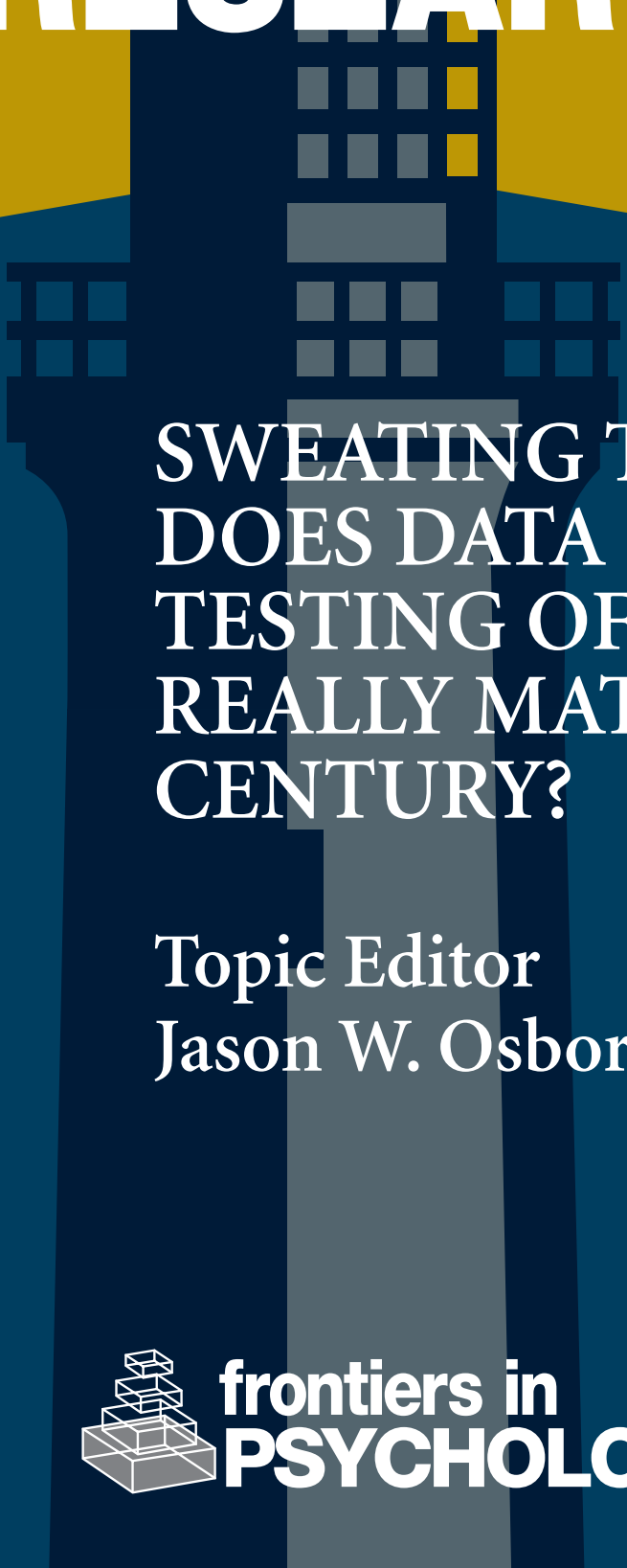
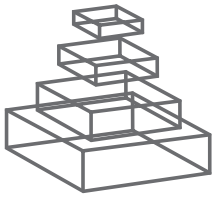
frontiers
RESEARCH TOPICS



SWEATING THE SMALL STUFF: DOES DATA CLEANING AND TESTING OF ASSUMPTIONS REALLY MATTER IN THE 21ST CENTURY?

Topic Editor
Jason W. Osborne

frontiers in
PSYCHOLOGY



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-155-0
DOI 10.3389/978-2-88919-155-0

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.
All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

SWEATING THE SMALL STUFF: DOES DATA CLEANING AND TESTING OF ASSUMPTIONS REALLY MATTER IN THE 21ST CENTURY?

Topic Editor:
Jason W. Osborne, University of Louisville**,** USA

Modern statistical software makes it easier than ever to do thorough data screening/cleaning and to test assumptions associated with the analyses researchers perform.

However, few authors (even in top-tier journals) seem to be reporting data cleaning/screening and testing assumptions associated with the statistical analyses being reported. Few popular textbooks seem to focus on these basics.

In the 21st Century, with our complex modern analyses, is data screening and cleaning still relevant? Do outliers or extreme scores matter any more? Does having normally distributed variables improve analyses? Are there new techniques for screening or cleaning data that researchers should be aware of?

Are most analyses robust to violations of most assumptions, to the point that researchers really don't need to pay attention to assumptions any more?

My goal for this special issue is examine this issue with fresh eyes and 21st century methods. I believe that we can demonstrate that these things do still matter, even when using "robust" methods or non-parametric techniques, and perhaps identify when they matter MOST or in what way they can most substantially affect the results of an analysis.

I believe we can encourage researchers to change their habits through evidence-based discussions revolving around these issues. It is possible we can even convince editors of important journals to include these aspects in their evaluation /review criteria, as many journals in the social sciences have done with effect size reporting in recent years.

I invite you to join me in demonstrating WHY paying attention to these mundane aspects of quantitative analysis can be beneficial to researchers.

Table of Contents

frontiers in
PSYCHOLOGY

EDITORIAL
published: 25 June 2013
doi: 10.3389/fpsyg.2013.00370

Is data cleaning and the testing of assumptions relevant in the 21st century?

Jason W. Osborne*

Educational and Counseling Psychology, University of Louisville, Louisville, Kentucky, USA
**Correspondence: jason.osborne@louisville.edu*

Edited by:
Axel Cleeremans, Université Libre de Bruxelles, Belgium

> You must understand fully what your assumptions say and what they imply. You must not claim that the "usual assumptions" are acceptable due to the robustness of your technique unless you really understand the implications and limits of this assertion in the context of your application. And you must absolutely never use any statistical method without realizing that you are implicitly making assumptions, and that the validity of your results can never be greater than that of the most questionable of these (Vardeman and Morris, 2003, p. 26).

Modern quantitative studies use sophisticated statistical analyses that rely upon numerous important assumptions to ensure the validity of the results and protection from mis-estimation of outcomes. Yet casual inspection of respected journals in various fields shows a marked absence of discussion of the mundane, basic staples of quantitative methodology such as data cleaning or testing of assumptions, leaving us in the troubling position of being surrounded by intriguing quantitative findings but not able to assess the quality or reliability of the knowledge base of our field.

Few of us become scientists in order to do harm to the literature. Indeed most of us seek to help people, improve the world in some way, to make a difference. However, all the effort in the world will not accomplish these goals in the absence of valid, reliable, generalizable results—which can only be had with clean (non-faulty) data and assumptions of analyses met.

WHERE DOES THIS IDEA OF DATA CLEANING AND TESTING ASSUMPTIONS COME FROM?

Researchers have discussed the importance of assumptions from the introduction of our early modern statistical tests (e.g., Pearson, 1901; Student, 1908; Pearson, 1931). Even the most recently-developed statistical tests are developed in a context of certain important assumptions about the data.

Mathematicians and statisticians developing the tests we take for granted today had to make certain explicit assumptions about the data in order to formulate the operations that occur "under the hood" when we perform statistical analyses. A common example is that the data (or errors) are normally distributed, or that all groups (errors) have roughly equal variance. Without these assumptions the formulae and conclusions are not valid.

Early in the 20th century these assumptions were the focus of vigorous debate and discussion. For example, since data rarely are perfectly normally distributed, how much of a deviation from normality is acceptable? Similarly, it is rare that two groups would have exactly identical variances, how close to equal is good enough to maintain the goodness of the results?

By the middle of the 20th century, researchers had assembled some evidence that some *minimal* violations of some assumptions had minimal effects on error rates under certain circumstances—in other words, if your variances are not exactly identical across all groups, but are relatively close, it is probably acceptable to interpret the results of that test despite this technical violation of assumptions. Box (1953) is credited with coining the term "robust" (Boneau, 1960) which usually indicates that violation of an assumption does not substantially influence the Type I error rate of the test[1]. Thus, many authors published studies showing that analyses such as simple one-factor ANOVA analyses are "robust" to non-normality of the populations (Pearson, 1931) and to variance inequality (Box, 1953) when group sizes are equal. This means that they concluded that modest (practical) violations of these assumptions would not increase the probability of Type I errors [although even Pearson (1931) notes that strong non-normality can bias results toward increased Type II errors].

These fundamental, important debates focused on minor (but practically insignificant) deviations from absolute normality or exactly equal variance, (i.e., if a skew of 0.01 or 0.05 would make results unreliable). Despite being relatively narrow in scope (e.g., primarily concerned with Type I error rates in the context of exactly equal sample sizes and relatively simple one-factor ANOVA analyses) these early studies appear to have given social scientists the impression that these basic assumptions are unimportant. These early studies do not mean, however, that *all* analyses are robust to *dramatic* violations of these assumptions, or attest to robustness without meeting the other conditions (e.g., exactly equal cell sizes).

These findings do not necessarily generalize to broad violations of any assumption under any condition, and leave open questions regarding Type II error rates and mis-estimation of effect sizes and confidence intervals. Unfortunately, the latter point seems to have been lost on many modern researchers. Recall that these early researchers on "robustness" were often applied statisticians working in places such as chemical and agricultural companies as well as research labs such as Bell Telephone Labs, not in the social sciences where data may be more likely to be messy. Thus, these authors are viewing "modest deviations" as exactly that- minor deviations from mathematical models of perfect normality and perfect equality of variance that are practically unimportant. Social scientists rarely see data that are as clean as that discussed in these robustness studies.

[1]Note that Type II error rates and mis-estimation of parameters is much less rarely discussed and investigated.

Further, important caveats came with conclusions around "robustness"—such as adequate sample sizes, equal group sizes, and relatively simple analyses such as one-factor ANOVA.

This mythology of robustness, however, appears to have taken root in the social sciences and may have been accepted as broad fact rather than narrowly, as intended. Through the latter half of the 20th century this term came to be used more often as researchers published narrowly-focused studies that appeared to reinforce the mythology of robustness, perhaps inadvertently indicating that robustness was the rule rather than the exception.

In one example of this type of research, studies reported that simple statistical procedures such as the Pearson Product-Moment Correlation and the One-Way ANOVA (e.g., Feir-Walsh and Toothaker, 1974; Havlicek and Peterson, 1977) were robust to even "substantial violations" of assumptions. It is perhaps not surprising that "robustness" appears to have become unquestioned canon among quantitative social scientists, despite the caveats to these latter assertions, and the important point that these assertions of robustness usually relates only to Type I error rates, yet other aspects of analyses (such as Type II error rates or the accuracy of the estimates of effects) might still be strongly influenced by violation of assumptions.

However, the finding that simple correlations might be robust to certain violations is not to say that similar but more complex procedures (e.g., multiple regression, path analysis, or structural equation modeling) are equally robust to these same violations. Similarly, should one-way ANOVA be robust to violations of assumptions[2], it is not clear that similar but more complex procedures (e.g., factorial ANOVA or ANCOVA) would be equally robust to these violations. Yet recent surveys of quantitative research in many sciences affirms that a relatively low percentage of authors in recent years report basic information such as having checked for extreme scores, normality of the data, or having tested assumptions of the statistical procedures being used (Keselman et al., 1998; Osborne, 2008; Osborne et al., 2012). It seems, then, that this "mythology of robustness" has led a substantial percentage of social science researchers to believe it unnecessary to check the goodness of their data and the assumptions that their tests are based on (or report having done so).

Recent surveys of top research journals in the social sciences[3] confirm that authors (and reviewers and editors) are disconcertingly casual about data cleaning and reporting of tests of assumptions. One prominent review of education and psychology research by Keselman et al. (1998) provided a thorough review of empirical social science during the 1990s. The authors reviewed studies from 17 prominent journals spanning different areas of education and psychology, focusing on empirical articles with ANOVA-type designs.

In looking at 61 studies utilizing univariate ANOVA between-subjects designs, the authors found that only 11.48% of authors reported anything related to assessing normality, almost uniformly assessing normality through descriptive rather than inferential methods. Further, only 8.20% reported assessing homogeneity of variance, and only 4.92% assessed both distributional assumptions and homogeneity of variance. While some earlier studies asserted ANOVA to be robust to violations of these assumptions (Feir-Walsh and Toothaker, 1974), more recent work contradicts this long-held belief, particularly where designs extend beyond simple One-Way ANOVA and where cell sizes are unbalanced (which seems fairly common in modern ANOVA analyses within the social sciences) (Wilcox, 1987; Lix et al., 1996).

In examining articles reporting multivariate analyses, Keselman et al. (1998) describe a more dire situation. None of the 79 studies utilizing multivariate ANOVA procedures reported examining relevant assumptions of variance homogeneity, and in only 6.33% of the articles was there any evidence of examining of distributional assumptions (such as normality).

Similarly, in their examination of 226 articles that utilized some type of repeated-measures analysis, only 15.50% made reference to some aspect of assumptions, but none appeared to report assessing sphericity, an important assumption in these designs that can lead to substantial inflation of error rates and mis-estimation of effects, when violated (Maxwell and Delaney, 1990, p. 474).

Finally, their assessment of articles utilizing covariance designs ($N = 45$) was equally disappointing—75.56% of the studies reviewed made no mention of any assumptions or sample distributions, and most (82.22%) failed to report any information about the assumption of homogeneity of regression slope, an assumption critical to the validity of ANCOVA designs.

Another survey of articles published in 1998 and 1999 volumes of well-respected Educational Psychology journals (Osborne, 2008) showed that indicators of high quality data cleaning in published articles were sorely lacking. Specifically, authors in these top educational psychology journals almost never reported testing any assumptions of the analyses used (only 8.30% reported having tested any assumption), only 26.0% reported reliability of data being analyzed, and none reported any significant data cleaning (e.g., examination of data for outliers, normality, analysis of missing data, random responding, etc.).

Finally, a recent survey of recent articles published in prominent APA journals 2009 volumes (Osborne et al., 2012) found improved, but uninspiring results (see Figure 1.1). For example, the percentage of authors reporting anything resembling minimal data cleaning ranged from 22 to 38% across journals. This represents a marked improvement from previous surveys, but still leaves a majority of authors failing to report any type of data cleaning or testing of assumptions, a troubling state of affairs. Similarly, between 10 and 32% reported checking for distributional assumptions, and 32–45% reported dealing with missing data in some way (although usually through methods considered sub-optimal). Clearly, even in the 21st century, the majority of authors in highly-respected scholarly journals fail to report information about these basic issues of quantitative methods.

When I wrote a whole book on data cleaning (Osborne, 2012), my goal was to debunk this mythology of robustness and *laissez-faire* that seems to have seeped into the zeitgeist of quantitative methods. The challenge handed to authors in this book was to

[2]To be clear, it is debatable as to whether these relatively simple procedures are as robust as previously asserted.

[3]Other reviewers in other sciences tend to find similar results, unfortunately.

go beyond the basics of data cleaning and testing assumptions—to show that assumptions and quality data are still relevant and important in the 21st century. They went above and beyond this challenge in many interesting—and unexpected ways. I hope that this is the beginning—or a continuation—of an important discussion that strikes at the very heart of our quantitative disciplines; namely, whether we can trust any of the results we read in journals, and whether we can apply (or generalize) those results beyond the limited scope of the original sample.

REFERENCES

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychol. Bull.* 57, 49–64. doi: 10.1037/h0041412

Box, G. (1953). Non-normality and tests on variances. *Biometrika* 40, 318.

Feir-Walsh, B., and Toothaker, L. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educ. Psychol. Meas.* 34, 789. doi: 10.1177/001316447403400406

Havlicek, L. L., and Peterson, N. L. (1977). Effect of the violation of assumptions upon significance levels of the Pearson r. *Psychol. Bull.* 84, 373–377. doi: 10.1037/0033-2909.84.2.373

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: an analysis of their ANOVA, MANOVA, and ANCOVA Analyses. *Rev. Edu. Res.* 68, 350–386. doi: 10.3102/00346543068003350

Lix, L., Keselman, J., and Keselman, H. (1996). Consequences of assumption violations revisited: a quantitative review of alternatives to the one-way analysis of variance "F" Test. *Rev. Educ. Res.* 66, 579–619.

Maxwell, S., and Delaney, H. (1990). *Designing Experiments and Analyzing Data: a Model Comparison Perspective*. Pacific Grove, CA: Brooks Cole Publishing Company.

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educ. Psychol.* 28, 1–10. doi: 10.1080/01443410701491718

Osborne, J. W. (2012). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Thousand Oaks, CA: Sage Publications.

Osborne, J. W., Kocher, B., and Tillman, D. (2012). "*Sweating the small stuff: do authors in APA journals clean data or test assumptions* (and should anyone care if they do)," in *Paper presented at the Annual meeting of the Eastern Education Research Association*, (Hilton Head, SC).

Pearson, E. (1931). The analysis of variance in cases of non-normal variation. *Biometrika* 23, 114.

Pearson, K. (1901). Mathematical contribution to the theory of evolution. VII: On the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 195, 1–47.

Student. (1908). The probable error of a mean. *Biometrika* 6, 1–25.

Vardeman, S., and Morris, M. (2003). Statistics and Ethics. *Am. Stat.* 57, 21–26. doi: 10.1198/0003130031072

Wilcox, R. (1987). New designs in analysis of variance. *Ann. Rev. Psychol.* 38, 29–60. doi: 10.1146/annurev.ps.38.020187.000333

Received: 16 April 2013; accepted: 06 June 2013; published online: 25 June 2013.

*Citation: Osborne JW (2013) Is data cleaning and the testing of assumptions relevant in the 21st century? Front. Psychol. **4**:370. doi: 10.3389/fpsyg.2013.00370*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 14 May 2012
doi: 10.3389/fpsyg.2012.00137

Are assumptions of well-known statistical techniques checked, and why (not)?

Rink Hoekstra[1,2]*, Henk A. L. Kiers[2] and Addie Johnson[2]

[1] GION –Institute for Educational Research, University of Groningen, Groningen, The Netherlands
[2] Department of Psychology, University of Groningen, Groningen, The Netherlands

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Jason W. Osborne, Old Dominion University, USA
Jelte M. Wicherts, University of Amsterdam, The Netherlands

****Correspondence:***
Rink Hoekstra, GION, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands
e-mail: r.hoekstra@rug.nl

A valid interpretation of most statistical techniques requires that one or more assumptions be met. In published articles, however, little information tends to be reported on whether the data satisfy the assumptions underlying the statistical techniques used. This could be due to self-selection: Only manuscripts with data fulfilling the assumptions are submitted. Another explanation could be that violations of assumptions are rarely checked for in the first place. We studied whether and how 30 researchers checked fictitious data for violations of assumptions in their own working environment. Participants were asked to analyze the data as they would their own data, for which often used and well-known techniques such as the *t*-procedure, ANOVA and regression (or non-parametric alternatives) were required. It was found that the assumptions of the techniques were rarely checked, and that if they were, it was regularly by means of a statistical test. Interviews afterward revealed a general lack of knowledge about assumptions, the robustness of the techniques with regards to the assumptions, and how (or whether) assumptions should be checked. These data suggest that checking for violations of assumptions is not a well-considered choice, and that the use of statistics can be described as opportunistic.

Keywords: assumptions, robustness, analyzing data, normality, homogeneity

INTRODUCTION

Most statistical techniques require that one or more assumptions be met, or, in the case that it has been proven that a technique is robust against a violation of an assumption, that the assumption is not violated too extremely. Applying the statistical techniques when assumptions are not met is a serious problem when analyzing data (Olsen, 2003; Choi, 2005). Violations of assumptions can seriously influence Type I and Type II errors, and can result in overestimation or underestimation of the inferential measures and effect sizes (Osborne and Waters, 2002). Keselman et al. (1998) argue that "The applied researcher who routinely adopts a traditional procedure without giving thought to its associated assumptions may unwittingly be filling the literature with nonreplicable results" (p. 351). Vardeman and Morris (2003) state "...absolutely never use any statistical method without realizing that you are implicitly making assumptions, and that the validity of your results can never be greater than that of the most questionable of these" (p. 26). According to the sixth edition of the APA Publication Manual, the methods researchers use "...must support their analytic burdens, including robustness to violations of the assumptions that underlie them..." [American Psychological Association (APA, 2009); p. 33]. The Manual does not explicitly state that researchers should check for possible violations of assumptions and report whether the assumptions were met, but it seems reasonable to assume that in the case that researchers do not check for violations of assumptions, they should be aware of the robustness of the technique.

Many articles have been written on the robustness of certain techniques with respect to violations of assumptions (e.g., Kohr and Games, 1974; Bradley, 1980; Sawilowsky and Blair, 1992; Wilcox and Keselman, 2003; Bathke, 2004), and many ways of checking to see if assumptions have been met (as well as solutions to overcoming problems associated with any violations) have been proposed (e.g., Keselman et al., 2008). Using a statistical test is one of the frequently mentioned methods of checking for violations of assumptions (for an overview of statistical methodology textbooks that directly or indirectly advocate this method, see e.g., Hayes and Cai, 2007). However, it has also been argued that it is not appropriate to check assumptions by means of tests (such as Levene's test) carried out before deciding on which statistical analysis technique to use because such tests compound the probability of making a Type I error (e.g., Schucany and Ng, 2006). Even if one desires to check whether or not an assumption is met, two problems stand in the way. First, assumptions are usually about the population, and in a sample the population is by definition not known. For example, it is usually not possible to determine the exact variance of the population in a sample-based study, and therefore it is also impossible to determine that two population variances are equal, as is required for the assumption of equal variances (also referred to as the assumption of homogeneity of variances) to be satisfied. Second, because assumptions are usually defined in a very strict way (e.g., all groups have equal variances in the population, or the variable is normally distributed in the population), the assumptions cannot reasonably be expected to be satisfied. Given these complications, researchers can usually only examine whether assumptions are not violated "too much" in their sample; for deciding on what is too much, information about

the robustness of the technique with regard to violations of the assumptions is necessary.

The assumptions of normality and of homogeneity of variances are required to be met for the *t*-test for independent group means, one of the most widely used statistical tests (Hayes and Cai, 2007), as well as for the frequently used techniques ANOVA and regression (Kashy et al., 2009). The assumption of normality is that the scores in the population in case of a *t*-test or ANOVA, and the population residuals in case of regression, be normally distributed. The assumption of homogeneity of variance requires equal population variances per group in case of a *t*-test or ANOVA, and equal population variances for every value of the independent variable for regression. Although researchers might be tempted to think that most statistical procedures are relatively robust against most violations, several studies have shown that this is often not the case, and that in the case of one-way ANOVA, unequal group sizes can have a negative impact on the technique's robustness (e.g., Havlicek and Peterson, 1977; Wilcox, 1987; Lix et al., 1996).

Many textbooks advise that the assumptions of normality and homogeneity of variance be checked graphically (Hazelton, 2003; Schucany and Ng, 2006), such as by making normal quantile plots for checking for normality. Another method, which is advised in many other textbooks (Hayes and Cai, 2007), is to use a so-called preliminary test to determine whether to continue with the intended technique or to use an alternative technique instead. Preliminary tests could, for example, be used to choose between a pooled *t*-test and a Welch *t*-test or between ANOVA and a non-parametric alternative. Following the argument that preliminary tests should not be used because, amongst others, they can inflate the probability of making a Type I error (e.g., Gans, 1981; Wilcox et al., 1986; Best and Rayner, 1987; Zimmerman, 2004, 2011; Schoder et al., 2006; Schucany and Ng, 2006; Rochon and Kieser, 2011), it has also been argued that in many cases unconditional techniques should be the techniques of choice (Hayes and Cai, 2007). For example, the Welch *t*-test, which does not require homogeneity of variance, would be seen *a priori* as preferable to the pooled variance *t*-test (Zimmerman, 1996; Hayes and Cai, 2007).

Although the conclusions one can draw when analyzing a data set with statistical techniques depend on whether the assumptions for that technique are met, and, if that is not the case, whether the technique is robust against violations of the assumption, no work, to our knowledge, describing whether researchers check for violations of assumptions in practice has been published. When possible violations of assumptions are checked, and why they sometimes are not, is a relevant question given the continuing prevalence of preliminary tests. For example, an inspection of the most recent 50 articles published in 2011 in *Psychological Science* that contained at least one *t*-test, ANOVA or regression analysis, revealed that in only three of these articles was the normality of the data or the homogeneity of variances discussed, leaving open the question of whether these assumptions were or were not checked in practice, and how. Keselman et al. (1998) showed in a review of 61 articles that used a between-subject univariate design that in only a small minority of articles anything about the assumptions of normality (11%) or homogeneity (8%) was mentioned, and that in only 5% of the articles something about both assumptions was mentioned. In the same article, Keselman et al. present results of another study in which 79 articles with a between-subject multivariate design were checked for references to assumptions, and again the assumptions were rarely mentioned. Osborne (2008) found similar results: In 96 articles published in high quality journals, checking of assumptions was reported in only 8% of the cases.

In the present paper a study is presented in which the behavior of researchers while analyzing data was observed, particularly with regard to the checking for violations of assumptions when analyzing data. It was hypothesized that the checking of assumptions might not routinely occur when analyzing data. There can, of course, be rational reasons for not checking assumptions. Researchers might, for example, have knowledge about the robustness of the technique they are using with respect to violations of assumptions, and therefore consider the checking of possible violations unnecessary. In addition to observing whether or not the data were checked for violations of assumptions, we therefore also administered a questionnaire to assess why data were not always checked for possible violations of assumptions. Specifically, we focused on four possible explanations for failing to check for violations of assumptions: (1) lack of knowledge of the assumption, (2) not knowing how to check whether the assumption has been violated, (3) not considering a possible violation of an assumption problematic (for example, because of the robustness of the technique), and (4) lack of knowledge of an alternative in the case that an assumption seems to be violated.

MATERIALS AND METHODS

PARTICIPANTS

Thirty Ph.D. students, 13 men and 17 women (mean age = 27, SD = 1.5), working at Psychology Departments (but not in the area of methodology or statistics) throughout The Netherlands, participated in the study. All had at least 2 years experience conducting research at the university level. Ph.D. students were selected because of their active involvement in the collection and analysis of data. Moreover, they were likely to have had their statistical education relatively recently, assuring a relatively up-to-date knowledge of statistics. They were required to have at least once applied a *t*-procedure, a linear regression analysis and an ANOVA, although not necessarily in their own research project. Ten participants were randomly selected from each of the Universities of Tilburg, Groningen, and Amsterdam, three cities in different regions of The Netherlands. In order to get 30 participants, 41 Ph.D. students were approached for the study, of whom 11 chose not to participate. Informed consent was obtained from all participants, and anonymity was ensured.

TASK

The task consisted of two parts: data analysis and questionnaire. For the *data analysis task* participants were asked to analyze six data sets, and write down their inferential conclusions for each data set. The *t*-test, ANOVA and regression (or unconditional alternatives to those techniques) were intended to be used, because they are relatively simple, frequently used, and because it was expected that most participants would be familiar with those techniques. Participants could take as much time as they wanted, and no limit was given to the length of the inferential conclusions. The second

part of the task consisted of filling in a *questionnaire* with questions about the participants' choices during the data analysis task, and about the participants' usual behavior with respect to assumption checking when analyzing data. All participants needed between 30 and 75 min to complete the data analysis task and between 35 and 65 min to complete the questionnaire. The questionnaire also included questions about participants' customs regarding visualization of data and inference, but these data are not presented here. All but three participants performed the two tasks at their own workplace. The remaining three used an otherwise unoccupied room in their department. During task performance, the first author was constantly present.

Data analysis task

The data for the six data sets that the participants were asked to analyze were offered in SPSS format, since every participant indicated using SPSS as their standard statistical package. Before starting to analyze the data, the participants were given a short description of a research question without an explicit hypothesis, but with a brief description of the variables in the SPSS file. The participants were asked to analyze the data sets and interpret the results as they do when they analyze and interpret their own data sets. Per data set, they were asked to write down an answer to the following question: "What do these results tell you about the situation in the population? Explain how you came to your conclusion". An example of such an instruction for which participants were expected to use linear regression analysis, translated from the Dutch, is shown in **Figure 1**. Participants were explicitly told that consultation of any statistical books or the internet was allowed, but only two participants availed themselves of this opportunity.

The short description of the six research questions was written in such a way as to suggest that two *t*-tests, two linear regression analyses and two ANOVAs should be carried out, without explicitly naming the analysis techniques. Of course, non-parametric or unconditional alternatives were considered appropriate, as well. The results of a pilot experiment indicated that the descriptions were indeed sufficient to guide people in using the desired technique: The five people tested in the pilot used the intended technique for each of the six data sets. All six research question descriptions, also in translated form, can be found in the Appendix to this study.

The six data sets differed with respect to the effect size, the significance of the outcomes, and to whether there was a "strong" violation of an assumption. Four of the six data sets contained significant effects, one of the two data sets for which a *t*-test was supposed to be used contained a clear violation of the assumption of normality, one of the two data sets for which ANOVA was supposed to be used contained a clear violation of the assumption of homogeneity of variance, and effect size was relatively large in three data sets and relatively small in the other data sets (see **Table 1** for an overview).

To get more information on which choices were made by the participants during task performance, and why these choices were made, participants were asked to "think aloud" during task performance. This was recorded on cassette. During task performance, the selections made within the SPSS program were noted by the first author, in order to be able to check whether there was any check for the assumptions relevant for the technique that was used. Furthermore, participants were asked to save the SPSS syntax files. For the analysis of task performance, the information from the notes made by the first author, the tape recordings and the syntax files were combined to examine behavior with respect to checking for violations of the assumptions of normality and homogeneity of variance. Of course, had participants chosen unconditional techniques for which one or both of these assumptions were not required to be met, their data for the scenario in question would not have been used, provided that a preliminary test was not carried out to decide whether to use the unconditional technique. However, in no cases were unconditional techniques used or even considered. The frequency of selecting preliminary tests was recorded separately.

Each data set was scored according to whether violations of the assumptions of normality and homogeneity of variance were checked for. A graphical assessment was counted as correctly

Table 1 | An overview of the properties of the six scenarios.

Scenario	Technique to be used	Effect size	*p*-Value	Violations of assumption
1	*t*-Test	Medium	0.04	Normality
2	*t*-Test	Very small	0.86	None
3	Regression analysis	Large	0.00	None
4	Regression analysis	Medium	0.01	None
5	ANOVA	Large	0.05	Homogeneity
6	ANOVA	Close to 0	0.58	None

A researcher is interested to what extent the weight of men can predict their self-esteem. She expects a linear relationship between weight and self-esteem. To study the relationship, she takes a random sample of 100 men, and administers a questionnaire to them to measure their self-esteem (on a scale from 0 to 50), and measures the participants' weight. In Column 1 of the SPSS file, the scores on the self-esteem questionnaire are given. The second column shows the weights of the men, measured in kilograms.

FIGURE 1 | An example of one of the research question descriptions. In this example, participants were supposed to answer this question by means of a regression analysis.

checking for the assumption, provided that the assessment was appropriate for the technique at hand. A correct check for the assumption of normality was recorded if, for the *t*-test and ANOVA, a graphical representation of the different groups was requested, except when the graph was used only to detect outliers. Merely looking at the numbers, without making a visual representation was considered insufficient. For regression analysis, making a plot of the residuals was considered to be a correct check of the assumption of normality. Deciding whether this was done explicitly was based on whether the participant made any reference to normality when thinking aloud. A second option was to make a QQ- or PP-plot of the residuals. Selecting the Kolmogorov–Smirnov test or the Shapiro–Wilk test within SPSS was considered checking for the assumption of normality using a preliminary test.

Three ways of checking for the assumption of homogeneity of variance for the *t*-test and ANOVA were considered adequate. The first was to make a graphical representation of the data in such a way that difference in variance between the groups was visible (e.g., boxplots or scatter plots, provided that they are given per group). A second way was to make an explicit reference to the variance of the groups. A final possibility was to compare standard deviations of the groups in the output, with or without making use of a rule of thumb to discriminate between violations and non-violations. For regression analysis, a scatter plot or a residual plot was considered necessary to check the assumption of homogeneity of variance. Although the assumption of homogeneity of variance assumes equality of the population variations, an explicit reference to the population was not required. The preliminary tests that were recorded included Levene's test, the *F*-ratio test, Bartlett's test, and the Brown–Forsythe test.

The frequency of using preliminary tests was reported separately from other ways of checking for assumptions. Although the use of preliminary tests is often considered an inappropriate method for checking assumptions, their use does show awareness of the existence of the assumption. Occurrences of checking for irrelevant assumptions, such as equal group sizes for the *t*-test, or normality of all scores for one variable (instead of checking for normality per group) for all three techniques were also counted, but scored as incorrectly checking for an assumption.

Questionnaire

The questionnaire addressed four explanations for why an assumption was not checked: (1) Unfamiliarity with the assumption, (2) Unfamiliarity with how to check the assumptions, (3) Violation of the assumption not being regarded problematic, and (4) Unfamiliarity with a remedy against a violation of the assumption. Each of these explanations was operationalized before the questionnaires were analyzed. The experimenter was present during questionnaire administration to stimulate the participants to answer more extensively, if necessary, or ask them to reformulate their answer when they seemed to have misread the question.

Unfamiliarity with the assumptions. Participants were asked to write down the assumptions they thought it was necessary to check for each of the three statistical techniques used in the study. Simply mentioning the assumption of normality or homogeneity of variance was scored as being familiar with the assumption, even if the participants did not specify what, exactly, was required to follow a normal distribution or which variances were supposed to be equal. Explaining the assumptions without explicitly mentioning them was also scored as being familiar with this assumption.

Unfamiliarity with how to check the assumptions. Participants were asked if they could think of a way to investigate whether there was a violation of each of the two assumptions (normality and homogeneity of variance) for *t*-tests, ANOVA and regression, respectively. Thus, the assumptions per technique were explicitly given, whether or not they had been correctly reported in answer to the previous question. For normality, specifying how to visualize the data in such a way that a possible violation was visible was categorized as a correct way of checking for assumption violations (for example: making a QQ-plot, or making a histogram), even when no further information was given about how to make such a visualization. Mentioning a measure of or a test for normality was also considered correct. For studying homogeneity of variance, rules of thumb or tests, such as Levene's test for testing equality of variances, were categorized as a correct way of checking this assumption, and the same holds for eyeballing visual representations from which variances could be deduced. Note that the criteria for a correct check are lenient, since they include preliminary tests that are usually considered inappropriate.

Violation of the assumption not being regarded problematic. For techniques for which it has been shown that they are robust against certain assumption violations, it can be argued that it makes sense *not* to check for these assumptions, because the outcome of this checking process would not influence the interpretation of the data anyway. To study this explanation, participants were asked per assumption and for the three techniques whether they considered a possible violation to be influential. Afterward, the answers that indicated that this influence was small or absent were scored as satisfying the criteria for this explanation.

Unfamiliarity with a remedy against a violation of an assumption. One could imagine that a possible violation of assumptions is not checked because no remedy for such violations is known. Participants were thus asked to note remedies for possible violations of normality and homogeneity of variance for each of the three statistical analysis techniques. Correct remedies were defined as transforming the data (it was not required that participants specify which transformation), using a different technique (e.g., a non-parametric technique when the assumption of normality has been violated) and increasing the sample size.

DATA ANALYSIS

All results are presented as percentages of the total number of participants or of the total number of analyzed data sets, depending on the specific research question. Confidence intervals (CIs) are given, but should be interpreted cautiously because the sample cannot be regarded as being completely random. The CIs for percentages were calculated by the so-called Score CIs (Wilson, 1927). All CIs are 95% CIs.

RESULTS

Of the six datasets that the 30 participants were required to analyze, in all but three instances the expected technique was chosen. In the remaining three instances, ANOVA was used to analyze data sets that were meant to be analyzed by means of a *t*-test. Since ANOVA is in this case completely equivalent to an independent-samples *t*-test, it can be concluded that an appropriate technique was chosen for all data sets. In none of these cases, an unconditional technique was chosen.

Violations of, or conformance with, the assumptions of normality and homogeneity of variance were correctly checked in 12% (95%CI = [8%, 18%]) and 23% (95%CI = [18%, 30%]), respectively, of the analyzed data sets. **Figure 2** shows for each of the three techniques how frequently possible violations of the assumptions of normality and homogeneity of variance occurred, and whether the checking was done correctly, or whether a preliminary test was used. Note that the assumption of normality was rarely checked for regression, and never correctly. In the few occasions that normality was checked the normality of the scores instead of the residuals was examined. Although this approach might be useful for studying the distribution of the scores, it is insufficient for determining whether the assumption of normality has been violated.

The percentages of participants giving each of the four reasons for not checking assumptions as measured by the questionnaire are given in **Figure 3**. A majority of the participants were unfamiliar with the assumptions. For each assumption, only a minority of participants mentioned at least one of the correct ways to check for a violation of the assumption. The majority of the participants failed to indicate that the alleged robustness of a technique against violations of the relevant assumption was a reason not to check these assumptions in the first place. Many participants did not know whether a violation of an assumption was important or not. Only in a minority of instances was an acceptable remedy for a violation of an assumption mentioned. No unacceptable remedies were mentioned. In general, participants indicated little knowledge of how to overcome a violation of one of the assumptions, and most participants reported never having looked for a remedy against a violation of statistical assumptions.

Participants had been told what the relevant assumptions were before they had to answer these questions. Therefore, the results for the last three explanations per assumption in **Figure 3** are reported for all participants, despite the fact that many participants reported being unfamiliar with the assumption. This implies that, especially for the assumption of normality and to a lesser extent for the assumption of equal variances, the results regarding the last three explanations should be interpreted with caution.

DISCUSSION

In order to examine people's understanding of the assumptions of statistical tests and their behavior with regard to checking these assumptions, 30 researchers were asked to analyze six data sets using the *t*-test, ANOVA, regression or a non-parametric

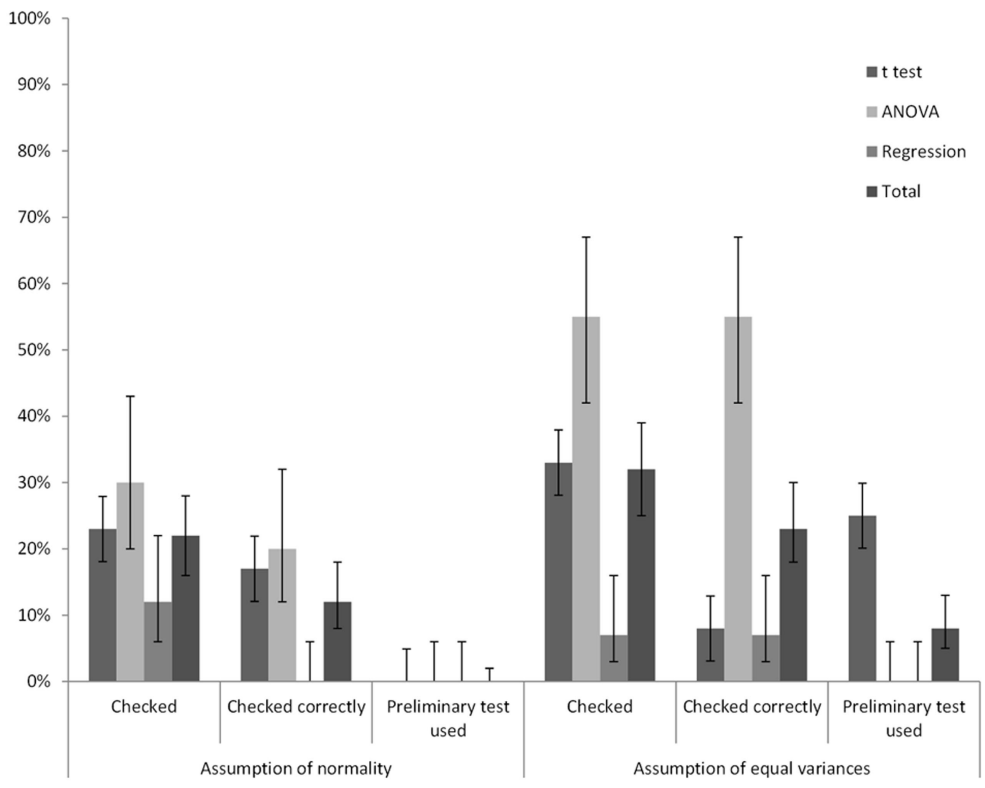


FIGURE 2 | The frequency of whether two assumptions were checked at all, whether they were checked correctly, and whether a preliminary test was used for three often used techniques in percentages of the total number of cases. Between brackets are 95% CIs for the percentages.

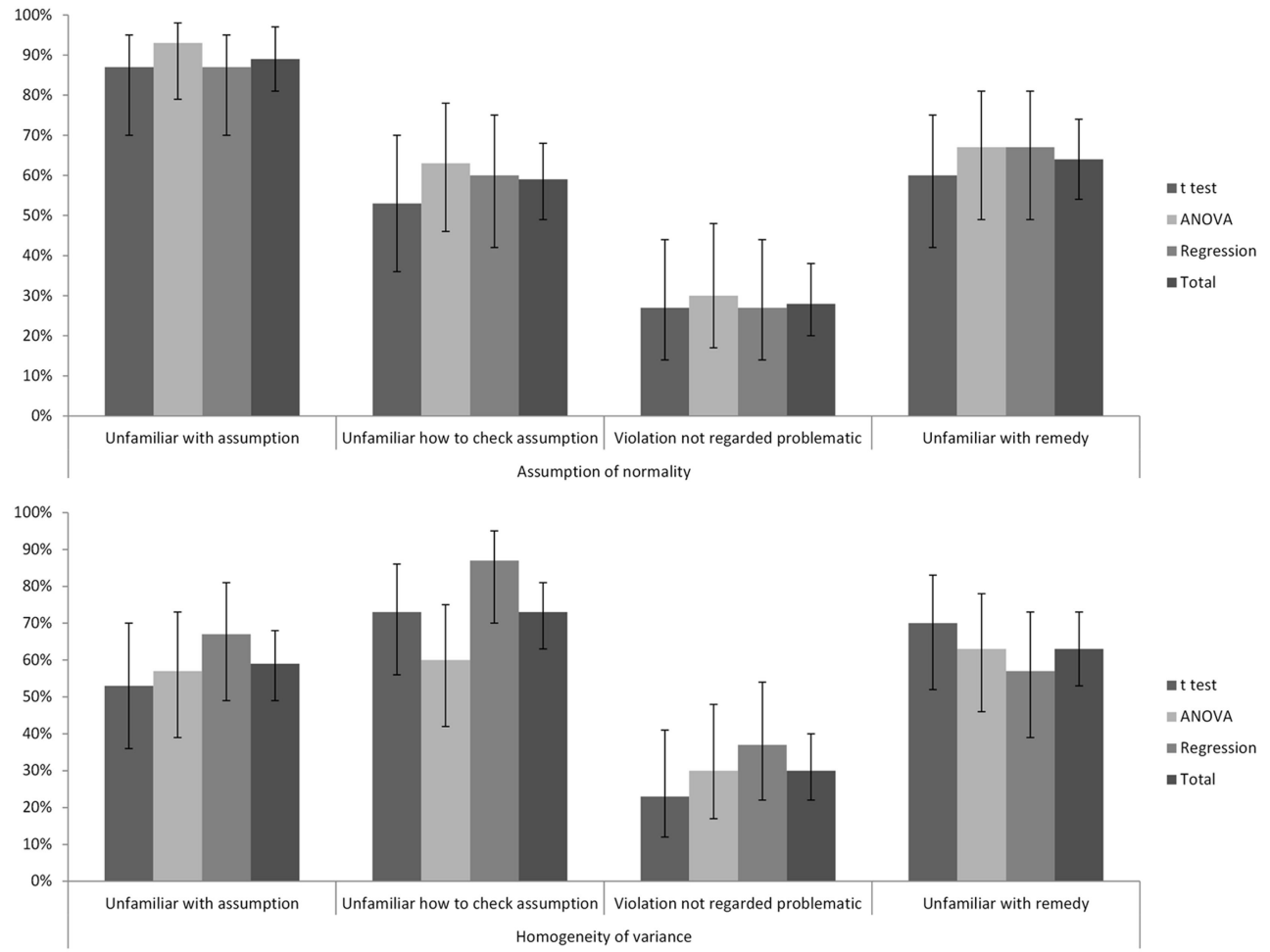


FIGURE 3 | Percentages of participants giving each of the explanations for not checking assumptions as a function of assumption and technique. Error bars indicate 95% CIs.

alternative, as appropriate. All participants carried out nominally appropriate analyses, but only in a minority of cases were the data examined for possible violations of assumptions of the chosen techniques. Preliminary test outcomes were rarely consulted, and only if given by default in the course of carrying out an analysis. The results of a questionnaire administered after the analyses were performed revealed that the failure to check for violations of assumptions could be attributed to the researchers' lack of knowledge about the assumptions, rather than to a deliberate decision not to check the data for violations of the assumptions.

Homogeneity of variance was checked for in roughly a third of all cases and the assumption of normality in less than a quarter of the data sets that were analyzed. Moreover, for the assumption of normality checks were often carried out incorrectly. An explanation for the finding that the assumption of homogeneity of variance was checked more often than the assumption of normality is the fact that a clear violation of this assumption can often be directly deduced from the standard deviations, whereas measures indicating normality are less common. Furthermore, many participants seemed familiar with a rule of thumb to check whether the assumption of homogeneity of variance for ANOVA is violated (e.g., largest standard deviation is larger than twice the smallest standard deviation), whereas such rules of thumb for checking possible violations of the assumption of normality were unknown to our participants. It was also found that Levene's test was often used as a preliminary test to choose between the pooled *t*-test and the Welch *t*-test, despite the fact that the use of preliminary tests is often discouraged (e.g., Wilcox et al., 1986; Zimmerman, 2004, 2011; Schucany and Ng, 2006). An obvious explanation for this could be that the outcomes of Levene's test are given as a default option for the *t* procedure in SPSS (this was the case in all versions that were used by the participants). The presence of Levene's test together with the corresponding *t*-tests may have led researchers to think that they should use this information. Support for this hypothesis is that preliminary tests were not carried out in any other cases.

It is possible that researchers have well-considered reasons for not checking for possible violations of assumption. For example, they may be aware of the robustness of a technique with respect to violations of a particular assumption, and quite reasonably chose

not to check to see if the assumption is violated. Our questionnaire, however, revealed that many researchers simply do not know which assumptions should be met for the *t*-test, ANOVA, and regression analysis. Only a minority of the researchers correctly named both assumptions, despite the fact that the statistical techniques themselves were well-known to the participants. Even when the assumptions were provided to the participants during the course of filling out the questionnaire, only a minority of the participants reported knowing a means of checking for violations, let alone which measures could be taken to remedy any possible violations or which tests could be used instead when violations could not be remedied.

A limitation of the present study is that, although researchers were asked to perform the tasks in their own working environment, the setting was nevertheless artificial, and for that reason the outcomes might have been biased. Researchers were obviously aware that they were being watched during this observation study, which may have changed their behavior. However, we expect that if they did indeed conduct the analyses differently than they would normally do, they likely attempted to perform better rather than worse than usual. A second limitation of the study is the relatively small number of participants. Despite this limited number and the resulting lower power, however, the effects are large, and the CIs show that the outcomes are unlikely to be due to chance alone. A third limitation is the possible presence of selection bias. The sample was not completely random because the selection of the universities involved could be considered a convenience sample. However, we have no reason to think that the sample is not representative of Ph.D. students at research universities. A fourth and last limitation is the fact that it is not clear what training each of the participants had on the topic of assumptions. However, all had their education in Psychology Departments in The Netherlands, where statistics is an important part of the basic curriculum. It is thus unlikely that they were not subjected to extensive discussion on the importance of meeting assumptions.

Our findings show that researchers are relatively unknowledgeable when it comes to when and how data should be checked for violations of assumptions of statistical tests. It is notable that the scientific community tolerates this lack of knowledge. One possible explanation for this state of affairs is that the scientific community as a whole does not consider it important to verify that the assumptions of statistical tests are met. Alternatively, other scientists may assume too readily that if nothing is said about assumptions in a manuscript, any crucial assumptions were met. Our results suggest that in many cases this might be a premature conclusion. It seems important to consider how statistical education can be improved to draw attention to the place of checking for assumptions in statistics and how to deal with possible violations (including deciding to use unconditional techniques). Requiring that authors describe how they checked for the violation of assumptions when the techniques applied are not robust to violations would, as Bakker and Wicherts (2011) have proposed, force researchers on both ends of the publishing process to show more awareness of this important issue.

REFERENCES

American Psychological Association. (2009). *Publication Manual of the American Psychological Association,* 6th Edn. Washington, DC: American Psychological Association

Bakker, M., and Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behav. Res. Methods* 43, 666–678.

Bathke, A. (2004). The ANOVA F test can still be used in some unbalanced designs with unequal variances and nonnormal data. *J. Stat. Plan. Inference* 126, 413–422.

Best, D. J., and Rayner, J. C. W. (1987). Welch's approximate solution for the Behrens–Fisher problem. *Technometrics* 29, 205–210.

Bradley, J. V. (1980). Nonrobustness in Z, t, and F tests at large sample sizes. *Bull. Psychon. Soc.* 16, 333–336.

Choi, P. T. (2005). Statistics for the reader: what to ask before believing the results. *Can. J. Anaesth.* 52, R1–R5.

Gans, D. J. (1981). Use of a preliminary test in comparing two sample means. *Commun. Stat. Simul. Comput.* 10, 163–174.

Havlicek, L. L., and Peterson, N. L. (1977). Effects of the violation of assumptions upon significance levels of the Pearson r. *Psychol. Bull.* 84, 373–377.

Hayes, A. F., and Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. *Br. J. Math. Stat. Psychol.* 60, 217–244.

Hazelton, M. L. (2003). A graphical tool for assessing normality, *Am. Stat.* 57, 285–288.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., and Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Pers. Soc. Psychol. Bull.* 35, 1131–1142.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., and Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychol. Methods* 13, 110–129.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., and Levin, J. R. (1998). Statistical practices of educational researchers: an analysis of their ANOVA, MANOVA and ANCOVA. *Rev. Educ. Res.* 68, 350.

Kohr, R. L., and Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *J. Exp. Educ.* 43, 61–69.

Lix, L. M., Keselman, J. C., and Keselman, H. J. (1996). Consequences of assumption violations revisited: a quantitative review of alternatives to the one-way analysis of variance F test. *Rev. Educ. Res.* 66, 579–620.

Olsen, C. H. (2003). Review of the use of statistics in infection and immunity. *Infect. Immun.* 71, 6689–6692.

Osborne, J. (2008). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educ. Psychol.* 28, 151–160.

Osborne, J. W., and Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Pract. Assess. Res. Evalu.* 8. Available at: http://www-psychology.concordia.ca/fac/kline/495/osborne.pdf

Rochon, J., and Kieser, M. (2011). A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample t-test. *Br. J. Math. Stat. Psychol.* 64, 410–426.

Sawilowsky, S. S., and Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychol. Bull.* 111, 352–360.

Schoder, V., Himmelmann, A., and Wilhelm, K. P. (2006). Preliminary testing for normality: some statistical aspects of a common concept. *Clin. Exp. Dermatol.* 31, 757–761.

Schucany, W. R., and Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample Student *t*. *Commun. Stat. Theory Methods* 35, 2275–2286.

Vardeman, S. B., and Morris, M. D. (2003). Statistics and ethics: some advice for young statisticians. *Am. Stat.* 57, 21.

Wilcox, R. R. (1987). New designs in analysis of variance. *Annu. Rev. Psychol.* 38, 29–60.

Wilcox, R. R., Charlin, V. L., and Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F∗ statistics. *Commun. Stat. Simul. Comput.* 15, 933–943.

Wilcox, R. R., and Keselman, H. J. (2003). Modern robust data analysis methods: measures of central tendency. *Psychol. Methods* 8, 254–274.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* 22, 209–212.

Zimmerman, D. W. (1996). Some properties of preliminary tests of equality of variances in the two-sample location problem. *J. Gen. Psychol.* 123, 217–231.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *Br. J. Math. Stat. Psychol.* 57, 173–181.

Zimmerman, D. W. (2011). A simple and effective decision rule for choosing a significance test to protect against non-normality. *Br. J. Math. Stat. Psychol.* 64, 388–409.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 November 2011; accepted: 20 April 2012; published online: 14 May 2012.

Citation: Hoekstra R, Kiers HAL and Johnson A (2012) Are assumptions of well-known statistical techniques checked, and why (not)? Front. Psychology ***3****:137. doi: 10.3389/fpsyg.2012.00137*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

APPENDIX

RESEARCH QUESTION DESCRIPTIONS

In this Appendix, the six research question descriptions are presented in translated form. Descriptions 1 and 2 were supposed to be answered by means of a *t*-test, Descriptions 3 and 4 by means of regression analysis, and Descriptions 5 and 6 by means of ANOVA.

1. A researcher is interested in the extent to which group A and group B differ in cognitive transcentivity. He has scores of 25 randomly selected participants from each of the two groups on a cognitive transcentivity test (with the range of possible scores from 0 to 25). In Column 1 of the SPSS file, the scores of the participants on the test are given, and in column 2 the group membership (group A or B) is given.
2. A researcher is interested in the extent to which group C and group D differ in cognitive transcentivity. He has scores of 25 randomly selected participants from each of the two groups on a cognitive transcentivity test (with the range of possible scores from 0 to 25). In Column 1 of the SPSS file, the scores of the participants on the test are given, and in column 2 the group membership (group C or D) is given.
3. A researcher is interested to what extent the weight of men can predict their self-esteem. She expects a linear relationship between weight and self-esteem. To study the relationship, she takes a random sample of 100 men, and administers a questionnaire to them to measure their self-esteem (on a scale from 0 to 50), and measures the participants' weight. In Column 1 of the SPSS file, the scores on the self-esteem questionnaire are given. The second column shows the weights of the men, measured in kilograms.
4. A researcher is interested to what extent the weight of women can predict their self-esteem. She expects a linear relationship between weight and self-esteem. To study the relationship, she takes a random sample of 100 women, and administers a questionnaire to them to measure their self-esteem (on a scale from 0 to 50), and measures the participants' weight. In Column 1 of the SPSS file, the scores on the self-esteem questionnaire are given. The second column shows the weights of the women, measured in kilograms.
5. A researcher is interested to what extent young people of three nationalities differ with respect to the time in which they can run the 100 meters. To study this, 20 persons between 20 and 30 years of age per nationality are randomly selected, and the times in which they run the 100 meters is measured. In Column 1 of the SPSS file, their times are given in seconds. The numbers "1, "2," and "3" in Column 2 represent the three different nationalities.
6. A researcher is interested to what extent young people of three *other* nationalities differ with respect to time in which they can run the 100 meters. To study this, 20 persons between 20 and 30 years of age per nationality are randomly selected, and the times in which they run the 100 meters is measured. In Column 1 of the SPSS file, their times are given in seconds. The numbers "1, "2," and "3" in Column 2 represent the three different nationalities.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 29 August 2012
doi: 10.3389/fpsyg.2012.00325

Statistical conclusion validity: some common threats and simple remedies

Miguel A. García-Pérez*

Facultad de Psicología, Departamento de Metodología, Universidad Complutense, Madrid, Spain

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Megan Welsh, University of Connecticut, USA
David Flora, York University, Canada

****Correspondence:***
Miguel A. García-Pérez, Facultad de Psicología, Departamento de Metodología, Campus de Somosaguas, Universidad Complutense, 28223 Madrid, Spain.
e-mail: miguel@psi.ucm.es

The ultimate goal of research is to produce dependable knowledge or to provide the evidence that may guide practical decisions. Statistical conclusion validity (SCV) holds when the conclusions of a research study are founded on an adequate analysis of the data, generally meaning that adequate statistical methods are used whose small-sample behavior is accurate, besides being logically capable of providing an answer to the research question. Compared to the three other traditional aspects of research validity (external validity, internal validity, and construct validity), interest in SCV has recently grown on evidence that inadequate data analyses are sometimes carried out which yield conclusions that a proper analysis of the data would not have supported. This paper discusses evidence of three common threats to SCV that arise from widespread recommendations or practices in data analysis, namely, the use of repeated testing and optional stopping without control of Type-I error rates, the recommendation to check the assumptions of statistical tests, and the use of regression whenever a bivariate relation or the equivalence between two variables is studied. For each of these threats, examples are presented and alternative practices that safeguard SCV are discussed. Educational and editorial changes that may improve the SCV of published research are also discussed.

Keywords: data analysis, validity of research, regression, stopping rules, preliminary tests

Psychologists are well aware of the traditional aspects of research validity introduced by Campbell and Stanley (1966) and further subdivided and discussed by Cook and Campbell (1979). Despite initial criticisms of the practically oriented and somewhat fuzzy distinctions among the various aspects (see Cook and Campbell, 1979, pp. 85–91; see also Shadish et al., 2002, pp. 462–484), the four facets of research validity have gained recognition and they are currently covered in many textbooks on research methods in psychology (e.g., Beins, 2009; Goodwin, 2010; Girden and Kabacoff, 2011). Methods and strategies aimed at securing research validity are also discussed in these and other sources. To simplify the description, *construct validity* is sought by using well-established definitions and measurement procedures for variables, *internal validity* is sought by ensuring that extraneous variables have been controlled and confounds have been eliminated, and *external validity* is sought by observing and measuring dependent variables under natural conditions or under an appropriate representation of them. The fourth aspect of research validity, which Cook and Campbell called *statistical conclusion validity* (SCV), is the subject of this paper.

Cook and Campbell, 1979, pp. 39–50) discussed that SCV pertains to the extent to which data from a research study can reasonably be regarded as revealing a link (or lack thereof) between independent and dependent variables *as far as statistical issues are concerned*. This particular facet was separated from other factors acting in the same direction (the three other facets of validity) and includes three aspects: (1) whether the study has enough statistical power to detect an effect if it exists, (2) whether there is a risk that the study will "reveal" an effect that does not actually exist, and (3) how can the magnitude of the effect be confidently estimated. They nevertheless considered the latter aspect as a mere step ahead once the first two aspects had been satisfactorily solved, and they summarized their position by stating that SCV "refers to inferences about whether it is reasonable to presume covariation given a specified α level and the obtained variances" (Cook and Campbell, 1979, p. 41). Given that mentioning "the obtained variances" was an indirect reference to statistical power and mentioning α was a direct reference to statistical significance, their position about SCV may have seemed to only entail consideration that the statistical decision can be incorrect as a result of Type-I and Type-II errors. Perhaps as a consequence of this literal interpretation, review papers studying SCV in published research have focused on power and significance (e.g., Ottenbacher, 1989; Ottenbacher and Maas, 1999), strategies aimed at increasing SCV have only considered these issues (e.g., Howard et al., 1983), and tutorials on the topic only or almost only mention these issues along with effect sizes (e.g., Orme, 1991; Austin et al., 1998; Rankupalli and Tandon, 2010). This emphasis on issues of significance and power may also be the reason that some sources refer to threats to SCV as "any factor that leads to a Type-I or a Type-II error" (e.g., Girden and Kabacoff, 2011, p. 6; see also Rankupalli and Tandon, 2010, Section 1.2), as if these errors had identifiable causes that could be prevented. It should be noted that SCV has also occasionally been purported to reflect the extent to which pre-experimental designs provide evidence for causation (Lee, 1985) or the extent to which

meta-analyses are based on representative results that make the conclusion generalizable (Elvik, 1998).

But Cook and Campbell's (1979, p. 80) aim was undoubtedly broader, as they stressed that SCV "is concerned with sources of random error *and with the appropriate use of statistics and statistical tests*" (italics added). Moreover, Type-I and Type-II errors are an essential and inescapable consequence of the statistical decision theory underlying significance testing and, as such, the potential occurrence of one or the other of these errors cannot be prevented. The actual occurrence of them for the data on hand cannot be assessed either. Type-I and Type-II errors will always be with us and, hence, SCV is only trivially linked to the fact that research will never unequivocally prove or reject any statistical null hypothesis or its originating research hypothesis. Cook and Campbell seemed to be well aware of this issue when they stressed that SCV refers to reasonable inferences given a specified significance level and a given power. In addition, Stevens (1950, p. 121) forcefully emphasized that "*it is a statistician's duty to be wrong* the stated number of times," implying that a researcher should accept the assumed risks of Type-I and Type-II errors, use statistical methods that guarantee the assumed error rates, and consider these as an essential part of the research process. From this position, these errors do not affect SCV unless their probability differs meaningfully from that which was assumed. And this is where an alternative perspective on SCV enters the stage, namely, whether the data were analyzed *properly* so as to extract conclusions that faithfully reflect what the data have to say about the research question. A negative answer raises concerns about SCV beyond the triviality of Type-I or Type-II errors. There are actually two types of threat to SCV from this perspective. One is when the data are subjected to thoroughly inadequate statistical analyses that do not match the characteristics of the design used to collect the data or that cannot logically give an answer to the research question. The other is when a proper statistical test is used but it is applied under conditions that alter the stated risk probabilities. In the former case, the conclusion will be wrong except by accident; in the latter, the conclusion will fail to be incorrect with the declared probabilities of Type-I and Type-II errors.

The position elaborated in the foregoing paragraph is well summarized in Milligan and McFillen's (1984, p. 439) statement that "under normal conditions (. . .) the researcher will not know when a null effect has been declared significant or when a valid effect has gone undetected (. . .) Unfortunately, the statistical conclusion validity, and the ultimate value of the research, rests on the explicit control of (Type-I and Type-II) error rates." This perspective on SCV is explicitly discussed in some textbooks on research methods (e.g., Beins, 2009, pp. 139–140; Goodwin, 2010, pp. 184–185) and some literature reviews have been published that reveal a sound failure of SCV in these respects.

For instance, Milligan and McFillen's (1984, p. 438) reviewed evidence that "the business research community has succeeded in publishing a great deal of incorrect and statistically inadequate research" and they dissected and discussed in detail four additional cases (among many others that reportedly could have been chosen) in which a breach of SCV resulted from gross mismatches between the research design and the statistical analysis. Similarly, García-Pérez (2005) reviewed alternative methods to compute confidence intervals for proportions and discussed three papers (among many others that reportedly could have been chosen) in which inadequate confidence intervals had been computed. More recently, Bakker and Wicherts (2011) conducted a thorough analysis of psychological papers and estimated that roughly 50% of published papers contain reporting errors, although they only checked whether the reported *p* value was correct and not whether the statistical test used was appropriate. A similar analysis carried out by Nieuwenhuis et al. (2011) revealed that 50% of the papers reporting the results of a comparison of two experimental effects in top neuroscience journals had used an incorrect statistical procedure. And Bland and Altman (2011) reported further data on the prevalence of incorrect statistical analyses of a similar nature.

An additional indicator of the use of inadequate statistical procedures arises from consideration of published papers whose title explicitly refers to a re-analysis of data reported in some other paper. A literature search for papers including in their title the terms "a re-analysis," "a reanalysis," "re-analyses," "reanalyses," or "alternative analysis" was conducted on May 3, 2012 in the Web of Science (WoS; http://thomsonreuters.com), which rendered 99 such papers with subject area "Psychology" published in 1990 or later. Although some of these were false positives, a sizeable number of them actually discussed the inadequacy of analyses carried out by the original authors and reported the results of proper alternative analyses that typically reversed the original conclusion. This type of outcome upon re-analyses of data are more frequent than the results of this quick and simple search suggest, because the information for identification is not always included in the title of the paper or is included in some other form: For a simple example, the search for the clause "a closer look" in the title rendered 131 papers, many of which also presented re-analyses of data that reversed the conclusion of the original study.

Poor design or poor sample size planning may, unbeknownst to the researcher, lead to unacceptable Type-II error rates, which will certainly affect SCV (as long as the null is not rejected; if it is, the probability of a Type-II error is irrelevant). Although insufficient power due to lack of proper planning has consequences on statistical tests, the thread of this paper de-emphasizes this aspect of SCV (which should perhaps more reasonably fit within an alternative category labeled *design validity*) and emphasizes the idea that SCV holds when statistical conclusions are incorrect with the stated probabilities of Type-I and Type-II errors (whether the latter was planned or simply computed). Whether or not the actual significance level used in the research or the power that it had is judged acceptable is another issue, which does not affect SCV: The statistical conclusion is valid within the stated (or computed) error probabilities. A breach of SCV occurs, then, when the data are not subjected to adequate statistical analyses or when control of Type-I or Type-II errors is lost.

It should be noted that a further component was included into consideration of SCV in Shadish et al.'s (2002) sequel to Cook and Campbell's (1979) book, namely, effect size. Effect size relates to what has been called a Type-III error (Crawford et al., 1998), that is, a statistically significant result that has no meaningful practical implication and that only arises from the use of a huge sample. This issue is left aside in the present paper because adequate consideration and reporting of effect sizes precludes Type-III errors,

although the recommendations of Wilkinson and The Task Force on Statistical Inference (1999) in this respect are not always followed. Consider, e.g., Lippa's (2007) study of the relation between sex drive and sexual attraction. Correlations generally lower than 0.3 in absolute value were declared strong as a result of p values below 0.001. With sample sizes sometimes nearing 50,000 paired observations, even correlations valued at 0.04 turned out significant in this study. More attention to effect sizes is certainly needed, both by researchers and by journal editors and reviewers.

The remainder of this paper analyzes three common practices that result in SCV breaches, also discussing simple replacements for them.

STOPPING RULES FOR DATA COLLECTION WITHOUT CONTROL OF TYPE-I ERROR RATES

The asymptotic theory that provides justification for null hypothesis significance testing (NHST) assumes what is known as *fixed sampling*, which means that the size n of the sample is not itself a random variable or, in other words, that the size of the sample has been decided in advance and the statistical test is performed once the entire sample of data has been collected. Numerous procedures have been devised to determine the size that a sample must have according to planned power (Ahn et al., 2001; Faul et al., 2007; Nisen and Schwertman, 2008; Jan and Shieh, 2011), the size of the effect sought to be detected (Morse, 1999), or the width of the confidence intervals of interest (Graybill, 1958; Boos and Hughes-Oliver, 2000; Shieh and Jan, 2012). For reviews, see Dell et al. (2002) and Maxwell et al. (2008). In many cases, a researcher simply strives to gather as large a sample as possible. Asymptotic theory supports NHST under fixed sampling assumptions, whether or not the size of the sample was planned.

In contrast to fixed sampling, *sequential sampling* implies that the number of observations is not fixed in advance but depends by some rule on the observations already collected (Wald, 1947; Anscombe, 1953; Wetherill, 1966). In practice, data are analyzed as they come in and data collection stops when the observations collected thus far satisfy some criterion. The use of sequential sampling faces two problems (Anscombe, 1953, p. 6): (i) devising a suitable stopping rule and (ii) finding a suitable test statistic and determining its sampling distribution. The mere statement of the second problem evidences that the sampling distribution of conventional test statistics for fixed sampling no longer holds under sequential sampling. These sampling distributions are relatively easy to derive in some cases, particularly in those involving negative binomial parameters (Anscombe, 1953; García-Pérez and Núñez-Antón, 2009). The choice between fixed and sequential sampling (sometimes portrayed as the "experimenter's intention"; see Wagenmakers, 2007) has important ramifications for NHST because the probability that the observed data are compatible (by any criterion) with a true null hypothesis generally differs greatly across sampling methods. This issue is usually bypassed by those who look at the data as a "sure fact" once collected, as if the sampling method used to collect the data did not make any difference or should not affect how the data are interpreted.

There are good reasons for using sequential sampling in psychological research. For instance, in clinical studies in which patients are recruited on the go, the experimenter may want to analyze data as they come in to be able to prevent the administration of a seemingly ineffective or even hurtful treatment to new patients. In studies involving a waiting-list control group, individuals in this group are generally transferred to an experimental group midway along the experiment. In studies with laboratory animals, the experimenter may want to stop testing animals before the planned number has been reached so that animals are not wasted when an effect (or the lack thereof) seems established. In these and analogous cases, the decision as to whether data will continue to be collected results from an analysis of the data collected thus far, typically using a statistical test that was devised for use in conditions of fixed sampling. In other cases, experimenters test their statistical hypothesis each time a new observation or block of observations is collected, and continue the experiment until they feel the data are conclusive one way or the other. Software has been developed that allows experimenters to find out how many more observations will be needed for a marginally non-significant result to become significant on the assumption that sample statistics will remain invariant when the extra data are collected (Morse, 1998).

The practice of repeated testing and optional stopping has been shown to affect in unpredictable ways the empirical Type-I error rate of statistical tests designed for use under fixed sampling (Anscombe, 1954; Armitage et al., 1969; McCarroll et al., 1992; Strube, 2006; Fitts, 2011a). The same holds when a decision is made to collect further data on evidence of a marginally (non) significant result (Shun et al., 2001; Chen et al., 2004). The inaccuracy of statistical tests in these conditions represents a breach of SCV, because the statistical conclusion thus fails to be incorrect with the assumed (and explicitly stated) probabilities of Type-I and Type-II errors. But there is an easy way around the inflation of Type-I error rates from within NHST, which solves the threat to SCV that repeated testing and optional stopping entail.

In what appears to be the first development of a sequential procedure with control of Type-I error rates in psychology, Frick (1998) proposed that repeated statistical testing be conducted under the so-called COAST (composite open adaptive sequential test) rule: If the test yields $p < 0.01$, stop collecting data and reject the null; if it yields $p > 0.36$, stop also and do not reject the null; otherwise, collect more data and re-test. The *low criterion* at 0.01 and the *high criterion* at 0.36 were selected through simulations so as to ensure a final Type-I error rate of 0.05 for paired-samples t tests. Use of the same low and high criteria rendered similar control of Type-I error rates for tests of the product-moment correlation, but they yielded slightly conservative tests of the interaction in 2×2 between-subjects ANOVAs. Frick also acknowledged that adjusting the low and high criteria might be needed in other cases, although he did not address them. This has nevertheless been done by others who have modified and extended Frick's approach (e.g., Botella et al., 2006; Ximenez and Revuelta, 2007; Fitts, 2010a,b, 2011b). The result is sequential procedures with stopping rules that guarantee accurate control of final Type-I error rates for the statistical tests that are more widely used in psychological research.

Yet, these methods do not seem to have ever been used in actual research, or at least their use has not been acknowledged. For instance, of the nine citations to Frick's (1998) paper listed in WoS as of May 3, 2012, only one is from a paper (published in 2011) in

which the COAST rule was reportedly used, although unintendedly. And not a single citation is to be found in WoS from papers reporting the use of the extensions and modifications of Botella et al. (2006) or Ximenez and Revuelta (2007). Perhaps researchers in psychology invariably use fixed sampling, but it is hard to believe that "data peeking" or "data monitoring" was never used, or that the results of such interim analyses never led researchers to collect some more data. Wagenmakers (2007, p. 785) regretted that "it is not clear what percentage of p values reported in experimental psychology have been contaminated by some form of optional stopping. There is simply no information in Results sections that allows one to assess the extent to which optional stopping has occurred." This incertitude was quickly resolved by John et al. (2012). They surveyed over 2000 psychologists with highly revealing results: Respondents affirmatively admitted to the practices of data peeking, data monitoring, or conditional stopping in rates that varied between 20 and 60%.

Besides John et al.'s (2012) proposal that authors disclose these details in full and Simmons et al.'s (2011) proposed list of requirements for authors and guidelines for reviewers, the solution to the problem is simple: Use strategies that control Type-I error rates upon repeated testing and optional stopping. These strategies have been widely used in biomedical research for decades (Bauer and Köhne, 1994; Mehta and Pocock, 2011). There is no reason that psychological research should ignore them and give up efficient research with control of Type-I error rates, particularly when these strategies have also been adapted and further developed for use under the most common designs in psychological research (Frick, 1998; Botella et al., 2006; Ximenez and Revuelta, 2007; Fitts, 2010a,b).

It should also be stressed that not all instances of repeated testing or optional stopping without control of Type-I error rates threaten SCV. A breach of SCV occurs only when the conclusion regarding the research question is based on the use of these practices. For an acceptable use, consider the study of Xu et al. (2011). They investigated order preferences in primates to find out whether primates preferred to receive the best item first rather than last. Their procedure involved several experiments and they declared that "three significant sessions (two-tailed binomial tests per session, $p < 0.05$) or 10 consecutive non-significant sessions were required from each monkey before moving to the next experiment. The three significant sessions were not necessarily consecutive (. . .) Ten consecutive non-significant sessions were taken to mean there was no preference by the monkey" (p. 2304). In this case, the use of repeated testing with optional stopping at a nominal 95% significance level for each individual test is part of the operational definition of an outcome variable used as a criterion to proceed to the next experiment. And, in any event, the overall probability of misclassifying a monkey according to this criterion is certainly fixed at a known value that can easily be worked out from the significance level declared for each individual binomial test. One may object to the value of the resultant risk of misclassification, but this does not raise concerns about SCV.

In sum, the use of repeated testing with optional stopping threatens SCV for lack of control of Type-I and Type-II error rates. A simple way around this is to refrain from these practices and adhere to the fixed sampling assumptions of statistical tests; otherwise, use the statistical methods that have been developed for use with repeated testing and optional stopping.

PRELIMINARY TESTS OF ASSUMPTIONS

To derive the sampling distribution of test statistics used in parametric NHST, some assumptions must be made about the probability distribution of the observations or about the parameters of these distributions. The assumptions of normality of distributions (in all tests), homogeneity of variances (in Student's two-sample t test for means or in ANOVAs involving between-subjects factors), sphericity (in repeated-measures ANOVAs), homoscedasticity (in regression analyses), or homogeneity of regression slopes (in ANCOVAs) are well known cases. The data on hand may or may not meet these assumptions and some parametric tests have been devised under alternative assumptions (e.g., Welch's test for two-sample means, or correction factors for the degrees of freedom of F statistics from ANOVAs). Most introductory statistics textbooks emphasize that the assumptions underlying statistical tests must be formally tested to guide the choice of a suitable test statistic for the null hypothesis of interest. Although this recommendation seems reasonable, serious consequences on SCV arise from following it.

Numerous studies conducted over the past decades have shown that the two-stage approach of testing assumptions first and subsequently testing the null hypothesis of interest has severe effects on Type-I and Type-II error rates. It may seem at first sight that this is simply the result of cascaded binary decisions each of which has its own Type-I and Type-II error probabilities; yet, this is the result of more complex interactions of Type-I and Type-II error rates that do not have fixed (empirical) probabilities across the cases that end up treated one way or the other according to the outcomes of the preliminary test: The resultant Type-I and Type-II error rates of the conditional test cannot be predicted from those of the preliminary and conditioned tests. A thorough analysis of what factors affect the Type-I and Type-II error rates of two-stage approaches is beyond the scope of this paper but readers should be aware that nothing suggests in principle that a two-stage approach might be adequate. The situations that have been more thoroughly studied include preliminary goodness-of-fit tests for normality before conducting a one-sample t test (Easterling and Anderson, 1978; Schucany and Ng, 2006; Rochon and Kieser, 2011), preliminary tests of equality of variances before conducting a two-sample t test for means (Gans, 1981; Moser and Stevens, 1992; Zimmerman, 1996, 2004; Hayes and Cai, 2007), preliminary tests of both equality of variances and normality preceding two-sample t tests for means (Rasch et al., 2011), or preliminary tests of homoscedasticity before regression analyses (Caudill, 1988; Ng and Wilcox, 2011). These and other studies provide evidence that strongly advises against conducting preliminary tests of assumptions. Almost all of these authors explicitly recommended against these practices and hoped for the misleading and misguided advice given in introductory textbooks to be removed. Wells and Hintze (2007, p. 501) concluded that "checking the assumptions using the same data that are to be analyzed, although attractive due to its empirical nature, is a fruitless endeavor because of its negative ramifications on the actual test of interest." The ramifications consist of substantial but

unknown alterations of Type-I and Type-II error rates and, hence, a breach of SCV.

Some authors suggest that the problem can be solved by replacing the formal test of assumptions with a decision based on a suitable graphical display of the data that helps researchers judge by eye whether the assumption is tenable. It should be emphasized that the problem still remains, because the decision on how to analyze the data is conditioned on the results of a preliminary analysis. The problem is not brought about by a formal preliminary test, but by the conditional approach to data analysis. The use of a non-formal preliminary test only prevents a precise investigation of the consequences on Type-I and Type-II error rates. But the "out of sight, out of mind" philosophy does not eliminate the problem.

It thus seems that a researcher must make a choice between two evils: either not testing assumptions (and, thus, threatening SCV as a result of the uncontrolled Type-I and Type-II error rates that arise from a potentially undue application of the statistical test) or testing them (and, then, also losing control of Type-I and Type-II error rates owing to the two-stage approach). Both approaches are inadequate, as applying non-robust statistical tests to data that do not satisfy the assumptions has generally as severe implications on SCV as testing preliminary assumptions in a two-stage approach. One of the solutions to the dilemma consists of switching to statistical procedures that have been designed for use under the two-stage approach. For instance, Albers et al. (2000) used second-order asymptotics to derive the size and power of a two-stage test for independent means preceded by a test of equality of variances. Unfortunately, derivations of this type are hard to carry out and, hence, they are not available for most of the cases of interest. A second solution consists of using classical test statistics that have been shown to be robust to violation of their assumptions. Indeed, dependable unconditional tests for means or for regression parameters have been identified (see Sullivan and D'Agostino, 1992; Lumley et al., 2002; Zimmerman, 2004, 2011; Hayes and Cai, 2007; Ng and Wilcox, 2011). And a third solution is switching to modern robust methods (see, e.g., Wilcox and Keselman, 2003; Keselman et al., 2004; Wilcox, 2006; Erceg-Hurn and Mirosevich, 2008; Fried and Dehling, 2011).

Avoidance of the two-stage approach in either of these ways will restore SCV while observing the important requirement that statistical methods should be used whose assumptions are not violated by the characteristics of the data.

REGRESSION AS A MEANS TO INVESTIGATE BIVARIATE RELATIONS OF ALL TYPES

Correlational methods define one of the branches of scientific psychology (Cronbach, 1957) and they are still widely used these days in some areas of psychology. Whether in regression analyses or in latent variable analyses (Bollen, 2002), vast amounts of data are subjected to these methods. Regression analyses rely on an assumption that is often overlooked in psychology, namely, that the predictor variables have fixed values and are measured without error. This assumption, whose validity can obviously be assessed without recourse to any preliminary statistical test, is listed in all statistics textbooks.

In some areas of psychology, predictors actually have this characteristic because they are physical variables defining the magnitude of stimuli, and any error with which these magnitudes are measured (or with which stimuli with the selected magnitudes are created) is negligible in practice. Among others, this is the case in psychophysical studies aimed at estimating *psychophysical functions* describing the form of the relation between physical magnitude and perceived magnitude (e.g., Green, 1982) or *psychometric functions* describing the form of the relation between physical magnitude and performance in a detection, discrimination, or identification task (Armstrong and Marks, 1997; Saberi and Petrosyan, 2004; García-Pérez et al., 2011). Regression or analogous methods are typically used to estimate the parameters of these relations, with stimulus magnitude as the independent variable and perceived magnitude (or performance) as the dependent variable. The use of regression in these cases is appropriate because the independent variable has fixed values measured without error (or with a negligible error). Another area in which the use of regression is permissible is in simulation studies on parameter recovery (García-Pérez et al., 2010), where the true parameters generating the data are free of measurement error by definition.

But very few other predictor variables used in psychology meet this requirement, as they are often test scores or performance measures that are typically affected by non-negligible and sometimes large measurement error. This is the case of the proportion of hits and the proportion of false alarms in psychophysical tasks, whose theoretical relation is linear under some signal detection models (DeCarlo, 1998) and, thus, suggests the use of simple linear regression to estimate its parameters. Simple linear regression is also sometimes used as a complement to statistical tests of equality of means in studies in which equivalence or agreement is assessed (e.g., Maylor and Rabbitt, 1993; Baddeley and Wilson, 2002), and in these cases equivalence implies that the slope should not differ significantly from unity and that the intercept should not differ significantly from zero. The use of simple linear regression is also widespread in priming studies after Greenwald et al. (1995; see also Draine and Greenwald, 1998), where the intercept (and sometimes the slope) of the linear regression of priming effect on detectability of the prime are routinely subjected to NHST.

In all the cases just discussed and in many others where the X variable in the regression of Y on X is measured with error, a study of the relation between X and Y through regression is inadequate and has serious consequences on SCV. The least of these problems is that there is no basis for assigning the roles of independent and dependent variable in the regression equation (as a non-directional relation exists between the variables, often without even a temporal precedence relation), but regression parameters will differ according to how these roles are assigned. In influential papers of which most researchers in psychology seem to be unaware, Wald (1940) and Mandansky (1959) distinguished regression relations from structural relations, the latter reflecting the case in which both variables are measured with error. Both authors illustrated the consequences of fitting a regression line when a structural relation is involved and derived suitable estimators and significance tests for the slope and intercept parameters of a structural relation. This topic was brought to the attention of psychologists by Isaac (1970) in a criticism of Treisman and Watts' (1966) use of simple linear regression to assess the equivalence of two alternative estimates of psychophysical sensitivity (d'

measures from signal detection theory analyses). The difference between regression and structural relations is briefly mentioned in passing in many elementary books on regression, the issue of fitting structural relations (sometimes referred to as *Deming's regression* or the *errors-in-variables regression model*) is addressed in detail in most intermediate and advance books on regression (e.g., Fuller, 1987; Draper and Smith, 1998) and hands-on tutorials have been published (e.g., Cheng and Van Ness, 1994; Dunn and Roberts, 1999; Dunn, 2007). But this type of analysis is not in the toolbox of the average researcher in psychology[1]. In contrast, recourse to this type analysis is quite common in the biomedical sciences.

Use of this commendable method may generalize when researchers realize that estimates of the slope β and the intercept α of a structural relation can be easily computed through

$$\hat{\beta} = \frac{S_y^2 - \lambda S_x^2 + \sqrt{\left(S_y^2 - \lambda S_x^2\right)^2 + 4\lambda S_{xy}^2}}{2S_{xy}}, \quad (1)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}, \quad (2)$$

where $\bar{X}$, $\bar{Y}$, S_x^2, S_y^2, and S_{xy} are the sample means, variances, and covariance of X and Y, and $\lambda = \sigma_{\varepsilon_y}^2/\sigma_{\varepsilon_x}^2$ is the ratio of the variances of measurement errors in Y and in X. When X and Y are the same variable measured at different times or under different conditions (as in Maylor and Rabbitt, 1993; Baddeley and Wilson, 2002), $\lambda = 1$ can safely be assumed (for an actual application, see Smith et al., 2004). In other cases, a rough estimate can be used, as the estimates of α and β have been shown to be robust except under extreme departures of the guesstimated λ from its true value (Ketellapper, 1983).

For illustration, consider Yeshurun et al. (2008) comparison of signal detection theory estimates of d' in each of the intervals of a two alternative forced-choice task, which they pronounced different as revealed by a regression analysis through the origin. Note that this is the context in which Isaac (1970) had illustrated the inappropriateness of regression. The data are shown in **Figure 1**, and Yeshurun et al. rejected equality of d'_1 and d'_2 because the regression slope through the origin (red line, whose slope is 0.908) differed significantly from unity: The 95% confidence interval for the slope ranged between 0.844 and 0.973. Using Eqs 1 and 2, the estimated structural relation is instead given by the blue line in **Figure 1**. The difference seems minor by eye, but the slope of the structural relation is 0.963, which is not significantly different from unity ($p = 0.738$, two-tailed; see Isaac, 1970, p. 215). This outcome, which reverses a conclusion raised upon inadequate data analyses, is representative of other cases in which the null hypothesis H_0: $\beta = 1$ was rejected. The reason is dual: (1) the slope of a structural relation is estimated with severe bias through regression (Riggs et al., 1978; Kalantar et al., 1995; Hawkins, 2002) and

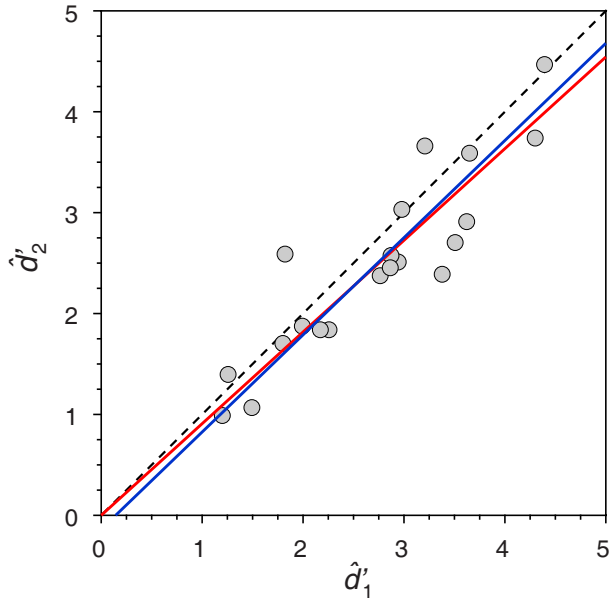


FIGURE 1 | Replot of data from Yeshurun et al. (2008, their Figure 8) with their fitted regression line through the origin (red line) and a fitted structural relation (blue line). The identity line is shown with dashed trace for comparison. For additional analyses bearing on the SCV of the original study, see García-Pérez and Alcalá-Quintana (2011).

(2) regression-based statistical tests of H_0: $\beta = 1$ render empirical Type-I error rates that are much higher than the nominal rate when both variables are measured with error (García-Pérez and Alcalá-Quintana, 2011).

In sum, SCV will improve if structural relations instead of regression equations were fitted when both variables are measured with error.

CONCLUSION

Type-I and Type-II errors are essential components of the statistical decision theory underlying NHST and, therefore, data can never be expected to answer a research question unequivocally. This paper has promoted a view of SCV that de-emphasizes consideration of these unavoidable errors and considers instead two alternative issues: (1) whether statistical tests are used that match the research design, goals of the study, and formal characteristics of the data and (2) whether they are applied in conditions under which the resultant Type-I and Type-II error rates match those that are declared as limiting the validity of the conclusion. Some examples of common threats to SCV in these respects have been discussed and simple and feasible solutions have been proposed. For reasons of space, another threat to SCV has not been covered in this paper, namely, the problems arising from multiple testing (i.e., in concurrent tests of more than one hypothesis). Multiple testing is commonplace in brain mapping studies and some implications on SCV have been discussed, e.g., by Bennett et al. (2009), Vul et al. (2009a,b), and Vecchiato et al. (2010).

All the discussion in this paper has assumed the frequentist approach to data analysis. In closing, and before commenting on

[1]SPSS includes a regression procedure called "two-stage least squares" which only implements the method described by Mandansky (1959) as "use of instrumental variables" to estimate the slope of the relation between X and Y. Use of this method requires extra variables with specific characteristics (variables which may simply not be available for the problem at hand) and differs meaningfully from the simpler and more generally applicable method to be discussed next

how SCV could be improved, a few words are worth about how Bayesian approaches fare on SCV.

THE BAYESIAN APPROACH

Advocates of Bayesian approaches to data analysis, hypothesis testing, and model selection (e.g., Jennison and Turnbull, 1990; Wagenmakers, 2007; Matthews, 2011) overemphasize the problems of the frequentist approach and praise the solutions offered by the Bayesian approach: Bayes factors (BFs) for hypothesis testing, credible intervals for interval estimation, Bayesian posterior probabilities, Bayesian information criterion (BIC) as a tool for model selection and, above all else, strict reliance on observed data and independence of the sampling plan (i.e., fixed vs. sequential sampling). There is unquestionable merit in these alternatives and a fair comparison with their frequentist counterparts requires a detailed analysis that is beyond the scope of this paper. Yet, I cannot resist the temptation of commenting on the presumed problems of the frequentist approach and also on the standing of the Bayesian approach with respect to SCV.

One of the preferred objections to *p* values is that they relate to data that were never collected and which, thus, should not affect the decision of what hypothesis the observed data support or fail to support. Intuitively appealing as it may seem, the argument is flawed because the referent for a *p* value is not other data sets that could have been observed in undone replications of the same experiment. Instead, the referent is the properties of the test statistic itself, which is guaranteed to have the declared sampling distribution when data are collected as assumed in the derivation of such distribution. Statistical tests are calibrated procedures with known properties, and this calibration is what makes their results interpretable. As is the case for any other calibrated procedure or measuring instrument, the validity of the outcome only rests on adherence to the usage specifications. And, of course, the test statistic and the resultant *p* value on application cannot be blamed for the consequences of a failure to collect data properly or to apply the appropriate statistical test.

Consider a two-sample *t* test for means. Those who need a referent may want to notice that the *p* value for the data from a given experiment relates to the uncountable times that such test has been applied to data from any experiment in any discipline. Calibration of the *t* test ensures that a proper use with a significance level of, say, 5% will reject a true null hypothesis on 5% of the occasions, no matter what the experimental hypothesis is, what the variables are, what the data are, what the experiment is about, who carries it out, or in what research field. What a *p* value indicates is how tenable it is that the *t* statistic will attain the observed value if the null were correct, with only a trivial link to the data observed in the experiment of concern. And this only places in a precise quantitative framework the logic that the man on the street uses to judge, for instance, that getting struck by lightning four times over the past 10 years is not something that could identically have happened to anybody else, or that the source of a politician's huge and untraceable earnings is not the result of allegedly winning top lottery prizes numerous times over the past couple of years. In any case, the advantage of the frequentist approach as regards SCV is that the probability of a Type-I or a Type-II error can be clearly and unequivocally stated, which is not to be mistaken for a statement that a *p* value is the probability of a Type-I error in the current case, or that it is a measure of the strength of evidence against the null that the current data provide. The most prevalent problems of *p* values are their potential for misuse and their widespread misinterpretation (Nickerson, 2000). But misuse or misinterpretation do not make NHST and *p* values uninterpretable or worthless.

Bayesian approaches are claimed to be free of these presumed problems, yielding a conclusion that is exclusively grounded on the data. In a naive account of Bayesian hypothesis testing, Malakoff (1999) attributes to biostatistician Steven Goodman the assertion that the Bayesian approach "says there is an X% probability that your hypothesis is true–not that there is some convoluted chance that if you assume the null hypothesis is true, you will get a similar or more extreme result if you repeated your experiment thousands of times." Besides being misleading and reflecting a poor understanding of the logic of calibrated NHST methods, what goes unmentioned in this and other accounts is that the Bayesian potential to find out the probability that the hypothesis is true will not materialize without two crucial extra pieces of information. One is the *a priori* probability of each of the competing hypotheses, which certainly does not come from the data. The other is the probability of the observed data under each of the competing hypothesis, which has the same origin as the frequentist *p* value and whose computation requires distributional assumptions that must necessarily take the sampling method into consideration.

In practice, Bayesian hypothesis testing generally computes BFs and the result might be stated as "the alternative hypothesis is *x* times more likely than the null," although the probability that this type of statement is wrong is essentially unknown. The researcher may be content with a conclusion of this type, but how much of these odds comes from the data and how much comes from the extra assumptions needed to compute a BF is undecipherable. In many cases research aims at gathering and analyzing data to make informed decisions such as whether application of a treatment should be discontinued, whether changes should be introduced in an educational program, whether daytime headlights should be enforced, or whether in-car use of cell phones should be forbidden. Like frequentist analyses, Bayesian approaches do not guarantee that the decisions will be correct. One may argue that stating how much more likely is one hypothesis over another bypasses the decision to reject or not reject any of them and, then, that Bayesian approaches to hypothesis testing are free of Type-I and Type-II errors. Although this is technically correct, the problem remains from the perspective of SCV: Statistics is only a small part of a research process whose ultimate goal is to reach a conclusion and make a decision, and researchers are in a better position to defend their claims if they can supplement them with a statement of the probability with which those claims are wrong.

Interestingly, analyses of decisions based on Bayesian approaches have revealed that they are no better than frequentist decisions as regards Type-I and Type-II errors and that parametric assumptions (i.e., the choice of prior and the assumed distribution of the observations) crucially determine the performance of Bayesian methods. For instance, Bayesian estimation is also subject to potentially large bias and lack of precision (Alcalá-Quintana and García-Pérez, 2004; García-Pérez and Alcalá-Quintana, 2007), the coverage probability of Bayesian credible intervals can be worse

than that of frequentist confidence intervals (Agresti and Min, 2005; Alcalá-Quintana and García-Pérez, 2005), and the Bayesian posterior probability in hypothesis testing can be arbitrarily large or small (Zaslavsky, 2010). On another front, use of BIC for model selection may discard a true model as often as 20% of the times, while a concurrent 0.05-size chi-square test rejects the true model between 3 and 7% of times, closely approximating its stated performance (García-Pérez and Alcalá-Quintana, 2012). In any case, the probabilities of Type-I and Type-II errors in practical decisions made from the results of Bayesian analyses will always be unknown and beyond control.

IMPROVING THE SCV OF RESEARCH

Most breaches of SCV arise from a poor understanding of statistical procedures and the resultant inadequate usage. These problems can be easily corrected, as illustrated in this paper, but the problems will not have arisen if researchers had had a better statistical training in the first place. There was a time in which one simply could not run statistical tests without a moderate understanding of NHST. But these days the application of statistical tests is only a mouse-click away and all that students regard as necessary is learning the rule by which p values pouring out of statistical software tell them whether the hypothesis is to be accepted or rejected, as the study of Hoekstra et al. (2012) seems to reveal.

One way to eradicate the problem is by improving statistical education at undergraduate and graduate levels, perhaps not just focusing on giving formal training on a number of methods but by providing students with the necessary foundations that will subsequently allow them to understand and apply methods for which they received no explicit formal training. In their analysis of statistical errors in published papers, Milligan and McFillen (1984, p. 461) concluded that "in doing projects, it is not unusual for applied researchers or students to use or apply a statistical procedure for which they have received no formal training. This is as inappropriate as a person conducting research in a given content area before reading the existing background literature on the topic. The individual simply is not prepared to conduct quality research. The attitude that statistical technology is secondary or less important to a person's formal training is shortsighted. Researchers are unlikely to master additional statistical concepts and techniques after leaving school. Thus, the statistical training in many programs must be strengthened. A single course in experimental design and a single course in multivariate analysis is probably insufficient for the typical student to master the course material. Someone who is trained only in theory and content will be ill-prepared to contribute to the advancement of the field or to critically evaluate the research of others." But statistical education does not seem to have changed much over the subsequent 25 years, as revealed by survey studies conducted by Aiken et al. (1990), Friedrich et al. (2000), Aiken et al. (2008), and Henson et al. (2010). Certainly some work remains to be done in this arena, and I can only second the proposals made in the papers just cited. But there is also the problem of the unhealthy over-reliance on narrow-breadth, clickable software for data analysis, which practically obliterates any efforts that are made to teach and promote alternatives (see the list of "Pragmatic Factors" discussed by Borsboom, 2006, pp. 431–434).

The last trench in the battle against breaches of SCV is occupied by journal editors and reviewers. Ideally, they also watch for problems in these respects. There is no known in-depth analysis of the review process in psychology journals (but see Nickerson, 2005) and some evidence reveals that the focus of the review process is not always on the quality or validity of the research (Sternberg, 2002; Nickerson, 2005). Simmons et al. (2011) and Wicherts et al. (2012) have discussed empirical evidence of inadequate research and review practices (some of which threaten SCV) and they have proposed detailed schemes through which feasible changes in editorial policies may help eradicate not only common threats to SCV but also other threats to research validity in general. I can only second proposals of this type. Reviewers and editors have the responsibility of filtering out (or requesting amendments to) research that does not meet the journal's standards, including SCV. The analyses of Milligan and McFillen (1984) and Nieuwenhuis et al. (2011) reveal a sizeable number of published papers with statistical errors. This indicates that some remains to be done in this arena too, and some journals have indeed started to take action (see Aickin, 2011).

ACKNOWLEDGMENTS

This research was supported by grant PSI2009-08800 (Ministerio de Ciencia e Innovación, Spain).

REFERENCES

Agresti, A., and Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 × 2 contingency tables. *Biometrics* 61, 515–523.

Ahn, C., Overall, J. E., and Tonidandel, S. (2001). Sample size and power calculations in repeated measurement analysis. *Comput. Methods Programs Biomed.* 64, 121–124.

Aickin, M. (2011). Test ban: policy of the Journal of Alternative and Complementary Medicine with regard to an increasingly common statistical error. *J. Altern. Complement. Med.* 17, 1093–1094.

Aiken, L. S., West, S. G., and Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *Am. Psychol.* 63, 32–50.

Aiken, L. S., West, S. G., Sechrest, L., and Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: a survey of PhD programs in North America. *Am. Psychol.* 45, 721–734.

Albers, W., Boon, P. C., and Kallenberg, W. C. M. (2000). The asymptotic behavior of tests for normal means based on a variance pre-test. *J. Stat. Plan. Inference* 88, 47–57.

Alcalá-Quintana, R., and García-Pérez, M. A. (2004). The role of parametric assumptions in adaptive Bayesian estimation. *Psychol. Methods* 9, 250–271.

Alcalá-Quintana, R., and García-Pérez, M. A. (2005). Stopping rules in Bayesian adaptive threshold estimation. *Spat. Vis.* 18, 347–374.

Anscombe, F. J. (1953). Sequential estimation. *J. R. Stat. Soc. Series B* 15, 1–29.

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics* 10, 89–100.

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. R. Stat. Soc. Ser. A* 132, 235–244.

Armstrong, L., and Marks, L. E. (1997). Differential effect of stimulus context on perceived length: implications for the horizontal–vertical illusion. *Percept. Psychophys.* 59, 1200–1213.

Austin, J. T., Boyle, K. A., and Lualhati, J. C. (1998). Statistical conclusion validity for organizational science researchers: a review. *Organ. Res. Methods* 1, 164–208.

Baddeley, A., and Wilson, B. A. (2002). Prose recall and amnesia: implications for the structure of working memory. *Neuropsychologia* 40, 1737–1743.

Bakker, M., and Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behav. Res. Methods* 43, 666–678.

Bauer, P., and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* 50, 1029–1041.

Beins, B. C. (2009). *Research Methods. A Tool for Life*, 2nd Edn. Boston, MA: Pearson Education.

Bennett, C. M., Wolford, G. L., and Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Soc. Cogn. Affect. Neurosci.* 4, 417–422.

Bland, J. M., and Altman, D. G. (2011). Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials* 12, 264.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annu. Rev. Psychol.* 53, 605–634.

Boos, D. D., and Hughes-Oliver, J. M. (2000). How large does n have to be for Z and t intervals? *Am. Stat.* 54, 121–128.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika* 71, 425–440.

Botella, J., Ximenez, C., Revuelta, J., and Suero, M. (2006). Optimization of sample size in controlled experiments: the CLAST rule. *Behav. Res. Methods Instrum. Comput.* 38, 65–76.

Campbell, D. T., and Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research.* Chicago, IL: Rand McNally.

Caudill, S. B. (1988). Type I errors after preliminary tests for heteroscedasticity. *Statistician* 37, 65–68.

Chen, Y. H. J., DeMets, D. L., and Lang, K. K. G. (2004). Increasing sample size when the unblinded interim result is promising. *Stat. Med.* 23, 1023–1038.

Cheng, C. L., and Van Ness, J. W. (1994). On estimating linear relationships when both variables are subject to errors. *J. R. Stat. Soc. Series B* 56, 167–183.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston, MA: Houghton Mifflin.

Crawford, E. D., Blumenstein, B., and Thompson, I. (1998). Type III statistical error. *Urology* 51, 675.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *Am. Psychol.* 12, 671–684.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychol. Methods* 3, 186–205.

Dell, R. B., Holleran, S., and Ramakrishnan, R. (2002). Sample size determination. *ILAR J.* 43, 207–213.

Draine, S. C., and Greenwald, A. G. (1998). Replicable unconscious semantic priming. *J. Exp. Psychol. Gen.* 127, 286–303.

Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis*, 3rd Edn. New York: Wiley.

Dunn, G. (2007). Regression models for method comparison data. *J. Biopharm. Stat.* 17, 739–756.

Dunn, G., and Roberts, C. (1999). Modelling method comparison data. *Stat. Methods Med. Res.* 8, 161–179.

Easterling, R. G., and Anderson, H. E. (1978). The effect of preliminary normality goodness of fit tests on subsequent inference. *J. Stat. Comput. Simul.* 8, 1–11.

Elvik, R. (1998). Evaluating the statistical conclusion validity of weighted mean results in meta-analysis by analysing funnel graph diagrams. *Accid. Anal. Prev.* 30, 255–266.

Erceg-Hurn, C. M., and Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *Am. Psychol.* 63, 591–601.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191.

Fitts, D. A. (2010a). Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. *Behav. Res. Methods* 42, 3–22.

Fitts, D. A. (2010b). The variable-criteria sequential stopping rule: generality to unequal sample sizes, unequal variances, or to large ANOVAs. *Behav. Res. Methods* 42, 918–929.

Fitts, D. A. (2011a). Ethics and animal numbers: Informal analyses, uncertain sample sizes, inefficient replications, and Type I errors. *J. Am. Assoc. Lab. Anim. Sci.* 50, 445–453.

Fitts, D. A. (2011b). Minimizing animal numbers: the variable-criteria sequential stopping rule. *Comp. Med.* 61, 206–218.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behav. Res. Methods Instrum. Comput.* 30, 690–697.

Fried, R., and Dehling, H. (2011). Robust nonparametric tests for the two-sample location problem. *Stat. Methods Appl.* 20, 409–422.

Friedrich, J., Buday, E., and Kerr, D. (2000). Statistical training in psychology: a national survey and commentary on undergraduate programs. *Teach. Psychol.* 27, 248–257.

Fuller, W. A. (1987). *Measurement Error Models.* New York: Wiley.

Gans, D. J. (1981). Use of a preliminary test in comparing two sample means. *Commun. Stat. Simul. Comput.* 10, 163–174.

García-Pérez, M. A. (2005). On the confidence interval for the binomial parameter. *Qual. Quant.* 39, 467–481.

García-Pérez, M. A., and Alcalá-Quintana, R. (2007). Bayesian adaptive estimation of arbitrary points on a psychometric function. *Br. J. Math. Stat. Psychol.* 60, 147–174.

García-Pérez, M. A., and Alcalá-Quintana, R. (2011). Testing equivalence with repeated measures: tests of the difference model of two-alternative forced-choice performance. *Span. J. Psychol.* 14, 1023–1049.

García-Pérez, M. A., and Alcalá-Quintana, R. (2012). On the discrepant results in synchrony judgment and temporal-order judgment tasks: a quantitative model. *Psychon. Bull. Rev.* (in press). doi:10.3758/s13423-012-0278-y

García-Pérez, M. A., Alcalá-Quintana, R., and García-Cueto, M. A. (2010). A comparison of anchor-item designs for the concurrent calibration of large banks of Likert-type items. *Appl. Psychol. Meas.* 34, 580–599.

García-Pérez, M. A., Alcalá-Quintana, R., Woods, R. L., and Peli, E. (2011). Psychometric functions for detection and discrimination with and without flankers. *Atten. Percept. Psychophys.* 73, 829–853.

García-Pérez, M. A., and Núñez-Antón, V. (2009). Statistical inference involving binomial and negative binomial parameters. *Span. J. Psychol.* 12, 288–307.

Girden, E. R., and Kabacoff, R. I. (2011). *Evaluating Research Articles. From Start to Finish*, 3rd Edn. Thousand Oaks, CA: Sage.

Goodwin, C. J. (2010). *Research in Psychology. Methods and Design*, 6th Edn. Hoboken, NJ: Wiley.

Graybill, F. A. (1958). Determining sample size for a specified width confidence interval. *Ann. Math. Stat.* 29, 282–287.

Green, B. G. (1982). The perception of distance and location for dual tactile figures. *Percept. Psychophys.* 31, 315–323.

Greenwald, A. G., Klinger, M. R., and Schuh, E. S. (1995). Activation by marginally perceptible ("subliminal") stimuli: dissociation of unconscious from conscious cognition. *J. Exp. Psychol. Gen.* 124, 22–42.

Hawkins, D. M. (2002). Diagnostics for conformity of paired quantitative measurements. *Stat. Med.* 21, 1913–1935.

Hayes, A. F., and Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. *Br. J. Math. Stat. Psychol.* 60, 217–244.

Henson, R. K., Hull, D. M., and Williams, C. S. (2010). Methodology in our education research culture: toward a stronger collective quantitative proficiency. *Educ. Res.* 39, 229–240.

Hoekstra, R., Kiers, H., and Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Front. Psychol.* 3:137. doi:10.3389/fpsyg.2012.00137

Howard, G. S., Obledo, F. H., Cole, D. A., and Maxwell, S. E. (1983). Linked raters' judgments: combating problems of statistical conclusion validity. *Appl. Psychol. Meas.* 7, 57–62.

Isaac, P. D. (1970). Linear regression, structural relations, and measurement error. *Psychol. Bull.* 74, 213–218.

Jan, S.-L., and Shieh, G. (2011). Optimal sample sizes for Welch's test under various allocation and cost considerations. *Behav. Res. Methods* 43, 1014–1022.

Jennison, C., and Turnbull, B. W. (1990). Statistical approaches to interim monitoring of clinical trials: a review and commentary. *Stat. Sci.* 5, 299–317.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532.

Kalantar, A. H., Gelb, R. I., and Alper, J. S. (1995). Biases in summary statistics of slopes and intercepts in linear regression with errors in both variables. *Talanta* 42, 597–603.

Keselman, H. J., Othman, A. R., Wilcox, R. R., and Fradette, K. (2004). The new and improved two-sample t test. *Psychol. Sci.* 15, 47–51.

Ketellapper, R. H. (1983). On estimating parameters in a simple linear errors-in-variables model. *Technometrics* 25, 43–47.

Lee, B. (1985). Statistical conclusion validity in ex post facto designs: practicality in evaluation. *Educ. Eval. Policy Anal.* 7, 35–45.

Lippa, R. A. (2007). The relation between sex drive and sexual attraction to men and women: a cross-national study of heterosexual, bisexual, and homosexual men and women. *Arch. Sex. Behav.* 36, 209–222.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annu. Rev. Public Health* 23, 151–169.

Malakoff, D. (1999). Bayes offers a "new" way to make sense of numbers. *Science* 286, 1460–1464.

Mandansky, A. (1959). The fitting of straight lines when both variables are subject to error. *J. Am. Stat. Assoc.* 54, 173–205.

Matthews, W. J. (2011). What might judgment and decision making research be like if we took a Bayesian approach to hypothesis testing? *Judgm. Decis. Mak.* 6, 843–856.

Maxwell, S. E., Kelley, K., and Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annu. Rev. Psychol.* 59, 537–563.

Maylor, E. A., and Rabbitt, P. M. A. (1993). Alcohol, reaction time and memory: a meta-analysis. *Br. J. Psychol.* 84, 301–317.

McCarroll, D., Crays, N., and Dunlap, W. P. (1992). Sequential ANOVAs and type I error rates. *Educ. Psychol. Meas.* 52, 387–393.

Mehta, C. R., and Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat. Med.* 30, 3267–3284.

Milligan, G. W., and McFillen, J. M. (1984). Statistical conclusion validity in experimental designs used in business research. *J. Bus. Res.* 12, 437–462.

Morse, D. T. (1998). MINSIZE: a computer program for obtaining minimum sample size as an indicator of effect size. *Educ. Psychol. Meas.* 58, 142–153.

Morse, D. T. (1999). MINSIZE2: a computer program for determining effect size and minimum sample size for statistical significance for univariate, multivariate, and nonparametric tests. *Educ. Psychol. Meas.* 59, 518–531.

Moser, B. K., and Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *Am. Stat.* 46, 19–21.

Ng, M., and Wilcox, R. R. (2011). A comparison of two-stage procedures for testing least-squares coefficients under heteroscedasticity. *Br. J. Math. Stat. Psychol.* 64, 244–258.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301.

Nickerson, R. S. (2005). What authors want from journal reviewers and editors. *Am. Psychol.* 60, 661–662.

Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* 14, 1105–1107.

Nisen, J. A., and Schwertman, N. C. (2008). A simple method of computing the sample size for chi-square test for the equality of multinomial distributions. *Comput. Stat. Data Anal.* 52, 4903–4908.

Orme, J. G. (1991). Statistical conclusion validity for single-system designs. *Soc. Serv. Rev.* 65, 468–491.

Ottenbacher, K. J. (1989). Statistical conclusion validity of early intervention research with handicapped children. *Except. Child.* 55, 534–540.

Ottenbacher, K. J., and Maas, F. (1999). How to detect effects: statistical power and evidence-based practice in occupational therapy research. *Am. J. Occup. Ther.* 53, 181–188.

Rankupalli, B., and Tandon, R. (2010). Practicing evidence-based psychiatry: 1. Applying a study's findings: the threats to validity approach. *Asian J. Psychiatr.* 3, 35–40.

Rasch, D., Kubinger, K. D., and Moder, K. (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Stat. Pap.* 52, 219–231.

Riggs, D. S., Guarnieri, J. A., and Addelman, S. (1978). Fitting straight lines when both variables are subject to error. *Life Sci.* 22, 1305–1360.

Rochon, J., and Kieser, M. (2011). A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample t-test. *Br. J. Math. Stat. Psychol.* 64, 410–426.

Saberi, K., and Petrosyan, A. (2004). A detection-theoretic model of echo inhibition. *Psychol. Rev.* 111, 52–66.

Schucany, W. R., and Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample Student t. *Commun. Stat. Theory Methods* 35, 2275–2286.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton Mifflin.

Shieh, G., and Jan, S.-L. (2012). Optimal sample sizes for precise interval estimation of Welch's procedure under various allocation and cost considerations. *Behav. Res. Methods* 44, 202–212.

Shun, Z. M., Yuan, W., Brady, W. E., and Hsu, H. (2001). Type I error in sample size re-estimations based on observed treatment difference. *Stat. Med.* 20, 497–513.

Simmons, J. P., Nelson, L. D., and Simoshohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366.

Smith, P. L., Wolfgang, B. F., and Sinclair, A. J. (2004). Mask-dependent attentional cuing effects in visual signal detection: the psychometric function for contrast. *Percept. Psychophys.* 66, 1056–1075.

Sternberg, R. J. (2002). On civility in reviewing. *APS Obs.* 15, 34.

Stevens, W. L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* 37, 117–129.

Strube, M. J. (2006). SNOOP: a program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behav. Res. Methods* 38, 24–27.

Sullivan, L. M., and D'Agostino, R. B. (1992). Robustness of the t test applied to data distorted from normality by floor effects. *J. Dent. Res.* 71, 1938–1943.

Treisman, M., and Watts, T. R. (1966). Relation between signal detectability theory and the traditional procedures for measuring sensory thresholds: estimating d' from results given by the method of constant stimuli. *Psychol. Bull.* 66, 438–454.

Vecchiato, G., Fallani, F. V., Astolfi, L., Toppi, J., Cincotti, F., Mattia, D., Salinari, S., and Babiloni, F. (2010). The issue of multiple univariate comparisons in the context of neuroelectric brain mapping: an application in a neuromarketing experiment. *J. Neurosci. Methods* 191, 283–289.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009a). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4, 274–290.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009b). Reply to comments on "Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition." *Perspect. Psychol. Sci.* 4, 319–324.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* 14, 779–804.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Ann. Math. Stat.* 11, 284–300.

Wald, A. (1947). *Sequential Analysis.* New York: Wiley.

Wells, C. S., and Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychol. Sch.* 44, 495–502.

Wetherill, G. B. (1966). *Sequential Methods in Statistics.* London: Chapman and Hall.

Wicherts, J. M., Kievit, R. A., Bakker, M., and Borsboom, D. (2012). Letting the daylight in: reviewing the reviewers and other ways to maximize transparency in science. *Front. Comput. Psychol.* 6:20. doi:10.3389/fncom.2012.00020

Wilcox, R. R. (2006). New methods for comparing groups: strategies for increasing the probability of detecting true differences. *Curr. Dir. Psychol. Sci.* 14, 272–275.

Wilcox, R. R., and Keselman, H. J. (2003). Modern robust data analysis methods: measures of central tendency. *Psychol. Methods* 8, 254–274.

Wilkinson, L., and The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *Am. Psychol.* 54, 594–604.

Ximenez, C., and Revuelta, J. (2007). Extending the CLAST sequential rule to one-way ANOVA under group sampling. *Behav. Res. Methods Instrum. Comput.* 39, 86–100.

Xu, E. R., Knight, E. J., and Kralik, J. D. (2011). Rhesus monkeys lack a consistent peak-end effect. *Q. J. Exp. Psychol.* 64, 2301–2315.

Yeshurun, Y., Carrasco, M., and Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: tests of the difference model. *Vision Res.* 48, 1837–1851.

Zaslavsky, B. G. (2010). Bayesian versus frequentist hypotheses testing in clinical trials with dichotomous and countable outcomes. *J. Biopharm. Stat.* 20, 985–997.

Zimmerman, D. W. (1996). Some properties of preliminary tests of equality of variances in the two-sample location problem. *J. Gen. Psychol.* 123, 217–231.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *Br. J. Math. Stat. Psychol.* 57, 173–181.

Zimmerman, D. W. (2011). A simple and effective decision rule for choosing a significance test to protect against non-normality. *Br. J. Math. Stat. Psychol.* 64, 388–409.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 May 2012; paper pending published: 29 May 2012; accepted: 14 August 2012; published online: 29 August 2012.

Citation: García-Pérez MA (2012) Statistical conclusion validity: some common threats and simple remedies. Front. Psychology ***3****:325. doi: 10.3389/fpsyg.2012.00325*
This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 15 February 2012
doi: 10.3389/fpsyg.2012.00034

Is coefficient alpha robust to non-normal data?

Yanyan Sheng*[1] *and* ***Zhaohui Sheng*[2]**

[1] *Department of Educational Psychology and Special Education, Southern Illinois University, Carbondale, IL, USA*
[2] *Department of Educational Leadership, Western Illinois University, Macomb, IL, USA*

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Pamela Kaliski, College Board, USA
James Stamey, Baylor University, USA

****Correspondence:***
Yanyan Sheng, Department of Educational Psychology and Special Education, Southern Illinois University, Wham 223, MC 4618, Carbondale, IL 62901-4618, USA.
e-mail: ysheng@siu.edu

Coefficient alpha has been a widely used measure by which internal consistency reliability is assessed. In addition to essential tau-equivalence and uncorrelated errors, normality has been noted as another important assumption for alpha. Earlier work on evaluating this assumption considered either exclusively non-normal error score distributions, or limited conditions. In view of this and the availability of advanced methods for generating univariate non-normal data, Monte Carlo simulations were conducted to show that non-normal distributions for true or error scores do create problems for using alpha to estimate the internal consistency reliability. The sample coefficient alpha is affected by leptokurtic true score distributions, or skewed and/or kurtotic error score distributions. Increased sample sizes, not test lengths, help improve the accuracy, bias, or precision of using it with non-normal data.

Keywords: coefficient alpha, true score distribution, error score distribution, non-normality, skew, kurtosis, Monte Carlo, power method polynomials

INTRODUCTION

Coefficient alpha (Guttman, 1945; Cronbach, 1951) has been one of the most commonly used measures today to assess internal consistency reliability despite criticisms of its use (e.g., Raykov, 1998; Green and Hershberger, 2000; Green and Yang, 2009; Sijtsma, 2009). The derivation of the coefficient is based on classical test theory (CTT; Lord and Novick, 1968), which posits that a person's observed score is a linear function of his/her unobserved true score (or underlying construct) and error score. In the theory, measures can be parallel (essential) tau-equivalent, or congeneric, depending on the assumptions on the units of measurement, degrees of precision, and/or error variances. When two tests are designed to measure the same latent construct, they are parallel if they measure it with identical units of measurement, the same precision, and the same amounts of error; tau-equivalent if they measure it with the same units, the same precision, but have possibly different error variance; essentially tau-equivalent if they assess it using the same units, but with possibly different precision and different amounts of error; or congeneric if they assess it with possibly different units of measurement, precision, and amounts of error (Lord and Novick, 1968; Graham, 2006). From parallel to congeneric, tests are requiring less strict assumptions and hence are becoming more general. Studies (Lord and Novick, 1968, pp. 87–91; see also Novick and Lewis, 1967, pp. 6–7) have shown formally that the population coefficient alpha equals internal consistency reliability for tests that are tau-equivalent or at least essential tau-equivalent. It underestimates the actual reliability for the more general congeneric test. Apart from essential tau-equivalence, coefficient alpha requires two additional assumptions: uncorrelated errors (Guttman, 1945; Novick and Lewis, 1967) and normality (e.g., Zumbo, 1999). Over the past decades, studies have well documented the effects of violations of essential tau-equivalence and uncorrelated errors (e.g., Zimmerman et al., 1993; Miller, 1995; Raykov, 1998; Green and Hershberger, 2000; Zumbo and Rupp, 2004; Graham, 2006; Green and Yang, 2009), which have been considered as two major assumptions for alpha. The normality assumption, however, has received little attention. This could be a concern in typical applications where the population coefficient is an unknown parameter and has to be estimated using the sample coefficient. When data are normally distributed, sample coefficient alpha has been shown to be an unbiased estimate of the population coefficient alpha (Kristof, 1963; van Zyl et al., 2000); however, less is known about situations when data are non-normal.

Over the past decades, the effect of departure from normality on the sample coefficient alpha has been evaluated by Bay (1973), Shultz (1993), and Zimmerman et al. (1993) using Monte Carlo simulations. They reached different conclusions on the effect of non-normal data. In particular, Bay (1973) concluded that a leptokurtic true score distribution could cause coefficient alpha to seriously underestimate internal consistency reliability. Zimmerman et al. (1993) and Shultz (1993), on the other hand, found that the sample coefficient alpha was fairly robust to departure from the normality assumption. The three studies differed in the design, in the factors manipulated and in the non-normal distributions considered, but each is limited in certain ways. For example, Zimmerman et al. (1993) and Shultz (1993) only evaluated the effect of non-normal error score distributions. Bay (1973), while looked at the effect of non-normal true score or error score distributions, only studied conditions of 30 subjects and 8 test items. Moreover, these studies have considered only two or three scenarios when it comes to non-normal distributions. Specifically, Bay (1973) employed uniform (symmetric platykurtic) and exponential (non-symmetric leptokurtic with positive skew) distributions for both true and error scores. Zimmerman et al. (1993) generated error scores from uniform, exponential, and mixed normal (symmetric leptokurtic) distributions, while Shultz (1993) generated them using exponential, mixed normal, and negative exponential

(non-symmetric leptokurtic with negative skew) distributions. Since the presence of skew and/or kurtosis determines whether and how a distribution departs from the normal pattern, it is desirable to consider distributions with varying levels of skew and kurtosis so that a set of guidelines can be provided. Generating univariate non-normal data with specified moments can be achieved via the use of power method polynomials (Fleishman, 1978), and its current developments (e.g., Headrick, 2010) make it possible to consider more combinations of skew and kurtosis.

Further, in the actual design of a reliability study, sample size determination is frequently an important and difficult aspect. The literature offers widely different recommendations, ranging from 15 to 20 (Fleiss, 1986), a minimum of 30 (Johanson and Brooks, 2010) to a minimum of 300 (Nunnally and Bernstein, 1994). Although Bay (1973) has used analytical derivations to suggest that coefficient alpha shall be robust against the violation of the normality assumption if sample size is large, or the number of items is large and the true score kurtosis is close to zero, it is never clear how many subjects and/or items are desirable in such situations.

In view of the above, the purpose of this study is to investigate the effect of non-normality (especially the presence of skew and/or kurtosis) on reliability estimation and how sample sizes and test lengths affect the estimation with non-normal data. It is believed that the results will not only shed insights on how non-normality affects coefficient alpha, but also provide a set of guidelines for researchers when specifying the numbers of subjects and items in a reliability study.

MATERIALS AND METHODS

This section starts with a brief review of the CTT model for coefficient alpha. Then the procedures for simulating observed scores used in the Monte Carlo study are described, followed by measures that were used to evaluate the performance of the sample alpha in each simulated situation.

PRELIMINARIES

Coefficient alpha is typically associated with true score theory (Guttman, 1945; Cronbach, 1951; Lord and Novick, 1968), where the test score for person i on item j, denoted as X_{ij}, is assumed to be a linear function of a true score (t_{ij}) and an error score (e_{ij}):

$$X_{ij} = t_{ij} + e_{ij}, \tag{1}$$

$i = 1, \ldots, n$ and $j = 1, \ldots, k$, where $E(e_{ij}) = 0$, $\rho_{te} = 0$, and $\rho_{e_{ij},e_{ij'}} = 0$. Here, e_{ij} denotes random error that reflects unpredictable trial-by-trial fluctuations. It has to be differentiated from systematic error that reflects situational or individual effects that may be specified. In the theory, items are usually assumed to be tau-equivalent, where true scores are restricted to be the same across items, or essentially tau-equivalent, where they are allowed to differ from item to item by a constant (υ_j). Under these conditions (1) becomes

$$X_{ij} = t_i + e_{ij} \tag{2}$$

for tau-equivalence, and

$$X_{ij} = t_i + \upsilon_j + e_{ij}, \tag{3}$$

where $\Sigma_j \upsilon_j = 0$, for essential tau-equivalence.

Summing across k items, we obtain a composite score (X_{i+}) and a scale error score (e_{i+}). The variance of the composite scores is then the summation of true score and scale error score variances:

$$\sigma^2_{X_+} = \sigma^2_t + \sigma^2_{e_+}. \tag{4}$$

The reliability coefficient, $\rho_{XX'}$, is defined as the proportion of composite score variance that is due to true score variance:

$$\rho_{XX'} = \frac{\sigma^2_t}{\sigma^2_{X_+}}. \tag{5}$$

Under (essential) tau-equivalence, that is, for models in (2) and (3), the population coefficient alpha, defined as

$$\alpha = \frac{k}{k-1}\frac{\sum\sum_{j \neq j'} \sigma_{X_j X_{j'}}}{\sigma^2_{X_+}},$$

or

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} \sigma^2_j}{\sigma^2_{X_+}}\right), \tag{6}$$

is equal to the reliability as defined in (5). As was noted, $\rho_{XX'}$ and α focus on the amount of random error and do not evaluate error that may be systematic.

Although the derivation of coefficient alpha based on Lord and Novick (1968) does not require distributional assumptions for t_i and e_{ij}, its estimation does (see Shultz, 1993; Zumbo, 1999), as the sample coefficient alpha estimated using sample variances s^2,

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} s^2_j}{s^2_{X_+}}\right), \tag{7}$$

is shown to be the maximum likelihood estimator of the population alpha assuming normal distributions (Kristof, 1963; van Zyl et al., 2000). Typically, we assume $t_i \sim N(\mu_t, \sigma^2_t)$ and $e_{ij} \sim N(0, \sigma^2_e)$, where σ^2_e has to be differentiated from the scale error score variance $\sigma^2_{e_+}$ defined in (4).

STUDY DESIGN

To evaluate the performance of the sample alpha as defined in (7) in situations where true score or error score distributions depart from normality, a Monte Carlo simulation study was carried out, where test scores of n persons ($n = 30, 50, 100, 1000$) for k items ($k = 5, 10, 30$) were generated assuming tau-equivalence and where the population reliability coefficient ($\rho_{XX'}$) was specified to be 0.3, 0.6, or 0.8 to correspond to unacceptable, acceptable, or very good reliability (Caplan et al., 1984, p. 306; DeVellis, 1991,

p. 85; Nunnally, 1967, p. 226). These are referred to as small, moderate, and high reliabilities in subsequent discussions. Specifically, true scores (t_i) and error scores (e_{ij}) were simulated from their respective distributions with $\sigma_e^2 = 1$, $\mu_t = 5$ and $\sigma_t^2 = \frac{\sigma_e^2 \rho_{XX'}}{(1-\rho_{XX'})k}$. The observed scores ($X_{ij}$) were subsequently obtained using Eq. (2).

In addition, true score or error score distributions were manipulated to be symmetric (so that skew, γ_1, is 0) or non-symmetric ($\gamma_1 > 0$) with kurtosis (γ_2) being 0, negative or positive. It is noted that only positively skewed distributions were considered in the study because due to the symmetric property, negative skew should have the same effect as positive skew. Generating non-normal distributions in this study involves the use of power method polynomials. Fleishman (1978) introduced this popular moment matching technique for generating univariate non-normal distributions. Headrick (2002, 2010) further extended from third-order to fifth-order polynomials to lower the skew and kurtosis boundary. As is pointed out by Headrick (2010, p. 26), for distributions with a mean of 0 and a variance of 1, the skew and kurtosis have to satisfy $\gamma_2 \geqslant \gamma_1^2 - 2$, and hence it is not plausible to consider all possible combinations of skew and kurtosis using power method polynomials. Given this, six distributions with the following combinations of skew and kurtosis were considered:

1. $\gamma_1 = 0, \gamma_2 = 0$ (normal distribution);
2. $\gamma_1 = 0, \gamma_2 = -1.385$ (symmetric platykurtic distribution);
3. $\gamma_1 = 0, \gamma_2 = 25$ (symmetric leptokurtic distribution);
4. $\gamma_1 = 0.96, \gamma_2 = 0.13$ (non-symmetric distribution);
5. $\gamma_1 = 0.48, \gamma_2 = -0.92$ (non-symmetric platykurtic distribution);
6. $\gamma_1 = 2.5, \gamma_2 = 25$ (non-symmetric leptokurtic distribution).

A normal distribution was included so that it could be used as a baseline against which the non-normal distributions could be compared. To actually generate univariate distributions using the fifth-order polynomial transformation, a random variate Z is first generated from a standard normal distribution, $Z \sim N(0,1)$. Then the following polynomial,

$$Y = c_0 + c_1 Z + c_2 Z^2 + c_3 Z^3 + c_4 Z^4 + c_5 Z^5 \tag{8}$$

is used to obtain Y. With appropriate coefficients ($c_0, \ldots, c_5$), Y would follow a distribution with a mean of 0, a variance of 1, and the desired levels of skew and kurtosis (see Headrick, 2002, for a detailed description of the procedure). A subsequent linear transformation would rescale the distribution to have a desired location or scale parameter. In this study, Y could be the true score (t_i) or the error score (e_{ij}). For the six distributions considered for t_i or e_{ij} herein, the corresponding coefficients are:

1. $c_0 = 0, c_1 = 1, c_2 = 0, c_3 = 0, c_4 = 0, c_5 = 0$;
2. $c_0 = 0, c_1 = 1.643377, c_2 = 0, c_3 = -0.319988, c_4 = 0, c_5 = 0.011344$;
3. $c_0 = 0, c_1 = 0.262543, c_2 = 0, c_3 = 0.201036, c_4 = 0, c_5 = 0.000162$;
4. $c_0 = -0.446924, c_1 = 1.242521, c_2 = 0.500764, c_3 = -0.184710, c_4 = -0.017947, c_5 = 0.003159$;
5. $c_0 = -0.276330, c_1 = 1.506715, c_2 = 0.311114, c_3 = -0.274078, c_4 = -0.011595, c_5 = 0.007683$;
6. $c_0 = -0.304852, c_1 = 0.381063, c_2 = 0.356941, c_3 = 0.132688, c_4 = -0.017363, c_5 = 0.003570$.

It is noted that the effect of the true score or error score distribution was investigated independently, holding the other constant by assuming it to be normal.

Hence, a total of 4 (sample sizes) × 3 (test lengths) × 3 (levels of population reliability) × 6 (distributions) × 2 (true or error score) = 432 conditions were considered in the simulation study. Each condition involved 100,000 replications, where coefficient alpha was estimated using Eq. (7) for simulated test scores (X_{ij}). The 100,000 estimates of α can be considered as random samples from the sampling distribution of $\hat{\alpha}$, and its summary statistics including the observed mean, SD, and 95% interval provide information about this distribution. In particular, the observed mean indicates whether the sample coefficient is biased. If it equals α, $\hat{\alpha}$ is unbiased; otherwise, it is biased either positively or negatively depending on whether it is larger or smaller than α. The SD of the sampling distribution is what we usually call the SE. It reflects the uncertainty in estimating α, with a smaller SE suggesting more precision and hence less uncertainty in the estimation. The SE is directly related to the 95% observed interval, as the larger it is, the more spread the distribution is and the wider the interval will be. With respect to the observed interval, it contains about 95% of $\hat{\alpha}$ around its center location from its empirical sampling distribution. If α falls inside the interval, $\hat{\alpha}$ is not significantly different from α even though it is not unbiased. On the other hand, if α falls outside of the interval, which means that 95% of the estimates differ from α, we can consider $\hat{\alpha}$ to be significantly different from α.

In addition to these summary statistics, the accuracy of the estimate was evaluated by the root mean square error (RMSE) and *bias*, which are defined as

$$\text{RMSE} = \sqrt{\frac{\sum (\hat{\alpha} - \alpha)^2}{100,000}}, \tag{9}$$

and

$$\text{bias} = \frac{\sum (\hat{\alpha} - \alpha)}{100,000}, \tag{10}$$

respectively. The larger the RMSE is, the less accurate the sample coefficient is in estimating the population coefficient. Similarly, the larger the absolute value of the *bias* is, the more bias the sample coefficient involves. As the equations suggest, RMSE is always positive, with values close to zero reflecting less error in estimating the actual reliability. On the other hand, *bias* can be negative or positive. A positive *bias* suggests that the sample coefficient tends to overestimate the reliability, and a negative *bias* suggests that it tends to underestimate the reliability. In effect, *bias* provides similar information as the observed mean of the sampling distribution of $\hat{\alpha}$.

RESULTS

The simulations were carried out using MATLAB (MathWorks, 2010), with the source code being provided in the Section "Appendix." Simulation results are summarized in **Tables 1–3** for conditions where true scores follow one of the six distributions specified in the previous section. Here, results from the five non-normal distributions were mainly compared with those from the normal distribution to determine if $\hat{\alpha}$ was affected by non-normality in true scores. Take the condition where a test of 5 items with the actual reliability being 0.3 was given to 30 persons as an example. A normal distribution resulted in an observed mean of 0.230 and a SE of 0.241 for the sampling distribution of $\hat{\alpha}$ (see **Table 1**). Compared with it, a symmetric platykurtic distribution, with an observed mean of 0.234 and a SE of 0.235, did not differ much.

Table 1 | Observed mean and SD of the sample alpha ($\hat{\alpha}$) for the simulated situations where the true score (t_i) distribution is normal or non-normal.

n	*k*	Mean ($\hat{\alpha}$)						SD ($\hat{\alpha}$)					
		dist1	**dist2**	**dist3**	**dist4**	**dist5**	**dist6**	**dist1**	**dist2**	**dist3**	**dist4**	**dist5**	**dist6**
$\rho_{XX'} = 0.3$													
30	5	0.230	0.234	0.198	0.230	0.231	0.201	0.241	0.235	0.290	0.242	0.237	0.288
	10	0.231	0.234	0.199	0.229	0.233	0.202	0.229	0.223	0.278	0.230	0.224	0.276
	30	0.231	0.233	0.199	0.230	0.233	0.200	0.221	0.215	0.269	0.222	0.216	0.270
50	5	0.252	0.253	0.233	0.253	0.254	0.232	0.176	0.172	0.214	0.177	0.172	0.214
	10	0.252	0.256	0.232	0.252	0.254	0.233	0.166	0.161	0.205	0.166	0.162	0.204
	30	0.254	0.254	0.231	0.252	0.254	0.233	0.160	0.156	0.202	0.160	0.157	0.199
100	5	0.269	0.269	0.258	0.268	0.269	0.258	0.118	0.116	0.148	0.119	0.117	0.148
	10	0.268	0.269	0.257	0.269	0.270	0.258	0.112	0.109	0.143	0.112	0.110	0.142
	30	0.269	0.270	0.257	0.268	0.269	0.256	0.108	0.105	0.141	0.108	0.106	0.140
1000	5	0.282	0.282	0.281	0.282	0.282	0.281	0.036	0.035	0.048	0.036	0.035	0.048
	10	0.282	0.282	0.281	0.282	0.282	0.281	0.034	0.033	0.046	0.034	0.033	0.046
	30	0.282	0.282	0.281	0.282	0.282	0.281	0.033	0.032	0.045	0.033	0.032	0.045
$\rho_{XX'} = 0.6$													
30	5	0.549	0.556	0.479	0.549	0.554	0.482	0.142	0.125	0.239	0.142	0.131	0.238
	10	0.551	0.557	0.481	0.549	0.554	0.480	0.133	0.117	0.232	0.136	0.122	0.232
	30	0.550	0.557	0.480	0.550	0.555	0.481	0.129	0.112	0.230	0.131	0.118	0.229
50	5	0.563	0.567	0.517	0.563	0.566	0.517	0.103	0.092	0.179	0.104	0.095	0.180
	10	0.563	0.567	0.516	0.563	0.566	0.517	0.097	0.086	0.176	0.098	0.089	0.174
	30	0.563	0.567	0.516	0.563	0.566	0.518	0.093	0.082	0.174	0.094	0.086	0.172
100	5	0.572	0.574	0.545	0.572	0.573	0.546	0.069	0.062	0.128	0.070	0.065	0.126
	10	0.572	0.574	0.545	0.572	0.573	0.547	0.066	0.057	0.126	0.066	0.060	0.124
	30	0.572	0.574	0.545	0.572	0.573	0.546	0.063	0.055	0.124	0.064	0.058	0.122
1000	5	0.580	0.580	0.576	0.580	0.580	0.577	0.021	0.019	0.043	0.021	0.020	0.042
	10	0.580	0.580	0.577	0.580	0.580	0.576	0.020	0.018	0.042	0.020	0.018	0.042
	30	0.580	0.580	0.576	0.580	0.580	0.576	0.019	0.017	0.042	0.019	0.018	0.041
$\rho_{XX'} = 0.8$													
30	5	0.771	0.778	0.701	0.770	0.776	0.703	0.072	0.056	0.171	0.075	0.062	0.172
	10	0.771	0.778	0.702	0.770	0.776	0.702	0.068	0.052	0.167	0.070	0.057	0.169
	30	0.771	0.778	0.701	0.771	0.776	0.702	0.066	0.049	0.166	0.068	0.055	0.167
50	5	0.778	0.782	0.733	0.778	0.780	0.733	0.052	0.041	0.125	0.053	0.045	0.125
	10	0.778	0.782	0.733	0.778	0.781	0.733	0.049	0.038	0.123	0.050	0.042	0.123
	30	0.778	0.782	0.732	0.778	0.781	0.733	0.048	0.036	0.122	0.049	0.040	0.122
100	5	0.782	0.784	0.757	0.782	0.784	0.757	0.035	0.028	0.086	0.036	0.031	0.085
	10	0.783	0.784	0.757	0.782	0.784	0.757	0.033	0.026	0.085	0.034	0.028	0.084
	30	0.783	0.784	0.757	0.782	0.784	0.757	0.032	0.024	0.084	0.033	0.027	0.084
1000	5	0.786	0.787	0.783	0.786	0.787	0.783	0.011	0.009	0.028	0.011	0.009	0.027
	10	0.786	0.787	0.783	0.787	0.787	0.783	0.010	0.008	0.028	0.010	0.009	0.027
	30	0.787	0.787	0.783	0.786	0.787	0.784	0.010	0.007	0.027	0.010	0.008	0.027

dist1, Normal distribution for t_i; dist2, distribution with negative kurtosis for t_i; dist3, distribution with positive kurtosis for t_i; dist4, skewed distribution for t_i; dist5, skewed distribution with negative kurtosis for t_i; dist6, skewed distribution with positive kurtosis for t_i.

Table 2 | Root mean square error and bias for estimating α for the simulated situations where the true score (t_i) distribution is normal or non-normal.

n	*k*	RMSE						bias					
		dist1	dist2	dist3	dist4	dist5	dist6	dist1	dist2	dist3	dist4	dist5	dist6
$\rho_{XX'} = 0.3$													
30	5	0.251	0.244	0.308	0.252	0.247	0.305	−0.070	−0.066	−0.102	−0.070	−0.069	−0.100
	10	0.240	0.233	0.296	0.241	0.234	0.292	−0.069	−0.067	−0.101	−0.071	−0.067	−0.098
	30	0.232	0.226	0.287	0.232	0.226	0.288	−0.069	−0.067	−0.101	−0.070	−0.067	−0.101
50	5	0.182	0.178	0.224	0.183	0.178	0.224	−0.048	−0.047	−0.067	−0.047	−0.046	−0.068
	10	0.173	0.167	0.216	0.173	0.169	0.215	−0.048	−0.044	−0.068	−0.048	−0.046	−0.067
	30	0.166	0.162	0.213	0.167	0.164	0.210	−0.046	−0.046	−0.069	−0.048	−0.046	−0.067
100	5	0.122	0.120	0.154	0.123	0.121	0.154	−0.031	−0.031	−0.042	−0.032	−0.031	−0.042
	10	0.116	0.114	0.149	0.116	0.114	0.148	−0.032	−0.031	−0.043	−0.031	−0.031	−0.042
	30	0.112	0.109	0.147	0.113	0.110	0.147	−0.031	−0.030	−0.043	−0.032	−0.031	−0.044
1000	5	0.040	0.040	0.052	0.041	0.040	0.051	−0.018	−0.018	−0.019	−0.018	−0.018	−0.019
	10	0.038	0.038	0.050	0.039	0.038	0.050	−0.018	−0.018	−0.019	−0.018	−0.018	−0.020
	30	0.038	0.037	0.049	0.038	0.037	0.049	−0.018	−0.018	−0.019	−0.018	−0.018	−0.019
$\rho_{XX'} = 0.6$													
30	5	0.151	0.132	0.268	0.151	0.139	0.266	−0.051	−0.044	−0.121	−0.051	−0.046	−0.118
	10	0.142	0.125	0.261	0.145	0.131	0.261	−0.050	−0.043	−0.120	−0.051	−0.046	−0.120
	30	0.139	0.120	0.260	0.140	0.126	0.258	−0.050	−0.043	−0.120	−0.051	−0.045	−0.119
50	5	0.109	0.097	0.198	0.110	0.101	0.198	−0.037	−0.033	−0.083	−0.037	−0.035	−0.083
	10	0.104	0.092	0.195	0.105	0.096	0.193	−0.037	−0.033	−0.084	−0.037	−0.034	−0.083
	30	0.100	0.088	0.193	0.102	0.092	0.191	−0.037	−0.033	−0.084	−0.037	−0.034	−0.083
100	5	0.075	0.067	0.139	0.076	0.070	0.137	−0.028	−0.026	−0.055	−0.028	−0.027	−0.054
	10	0.071	0.063	0.137	0.072	0.066	0.135	−0.028	−0.026	−0.056	−0.028	−0.027	−0.053
	30	0.069	0.061	0.135	0.070	0.064	0.133	−0.028	−0.026	−0.055	−0.028	−0.027	−0.054
1000	5	0.029	0.028	0.049	0.029	0.028	0.049	−0.020	−0.020	−0.024	−0.020	−0.020	−0.024
	10	0.028	0.027	0.048	0.029	0.027	0.048	−0.020	−0.020	−0.023	−0.020	−0.020	−0.024
	30	0.028	0.026	0.048	0.028	0.027	0.048	−0.020	−0.020	−0.024	−0.020	−0.020	−0.024
$\rho_{XX'} = 0.8$													
30	5	0.078	0.060	0.197	0.080	0.066	0.198	−0.030	−0.022	−0.099	−0.030	−0.024	−0.097
	10	0.074	0.056	0.194	0.076	0.062	0.196	−0.029	−0.023	−0.098	−0.030	−0.024	−0.099
	30	0.072	0.053	0.193	0.074	0.060	0.193	−0.029	−0.022	−0.099	−0.029	−0.024	−0.098
50	5	0.057	0.045	0.142	0.058	0.049	0.142	−0.022	−0.018	−0.067	−0.023	−0.020	−0.067
	10	0.054	0.042	0.140	0.055	0.046	0.140	−0.022	−0.018	−0.067	−0.022	−0.020	−0.067
	30	0.052	0.040	0.140	0.054	0.044	0.139	−0.022	−0.018	−0.068	−0.022	−0.019	−0.067
100	5	0.039	0.032	0.096	0.040	0.035	0.095	−0.018	−0.016	−0.043	−0.018	−0.016	−0.043
	10	0.038	0.030	0.095	0.038	0.033	0.094	−0.018	−0.016	−0.043	−0.018	−0.016	−0.043
	30	0.037	0.029	0.094	0.037	0.032	0.094	−0.018	−0.016	−0.043	−0.018	−0.016	−0.043
1000	5	0.017	0.016	0.033	0.017	0.016	0.032	−0.014	−0.013	−0.017	−0.014	−0.013	−0.017
	10	0.017	0.016	0.032	0.017	0.016	0.032	−0.014	−0.013	−0.017	−0.014	−0.013	−0.017
	30	0.017	0.015	0.032	0.017	0.016	0.032	−0.014	−0.013	−0.017	−0.014	−0.013	−0.017

dist1, Normal distribution for t_i; dist2, distribution with negative kurtosis for t_i; dist3, distribution with positive kurtosis for t_i; dist4, skewed distribution for t_i; dist5, skewed distribution with negative kurtosis for t_i; dist6, skewed distribution with positive kurtosis for t_i.

On the other hand, a symmetric leptokurtic distribution resulted in a much smaller mean (0.198) and a larger SE (0.290), indicating that the center location of the sampling distribution of $\hat{\alpha}$ was further away from the actual value (0.3) and more uncertainty was involved in estimating α. With respect to the accuracy of the estimate, **Table 2** shows that the normal distribution had a RMSE of 0.251 and a *bias* value of −0.070. The platykurtic distribution gave rise to smaller but very similar values: 0.244 for RMSE and −0.066 for *bias*, whereas the leptokurtic distribution had a relatively larger RMSE value (0.308) and a smaller *bias* value (−0.102), indicating that it involved more error and negative bias in estimating α. Hence, under this condition, positive kurtosis affected (the location and scale of) the sampling distribution of $\hat{\alpha}$ as well as the accuracy of using it to estimate α whereas negative kurtosis did not. Similar interpretations are used for the 95% interval shown in **Table 3**, except that one can also use the intervals to determine

Table 3 | Observed 95% interval of the sample alpha ($\hat{\alpha}$) for the simulated situations where the true score (t_i) distribution is normal or non-normal.

n	*k*	dist1		dist2		dist3		dist4		dist5		dist6	
		LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB
$\rho_{XX'} = 0.3$													
30	5	−0.351	0.580	−0.329	0.577	−0.490	0.635	−0.356	0.580	−0.342	0.576	−0.481	0.637
	10	−0.323	0.563	−0.305	0.556	−0.457	0.630	−0.328	0.561	−0.308	0.558	−0.450	0.624
	30	−0.303	0.550	−0.285	0.545	−0.435	0.618	−0.303	0.551	−0.286	0.547	−0.435	0.616
50	5	−0.155	0.528	−0.143	0.524	−0.252	0.587	−0.155	0.529	−0.147	0.524	−0.255	0.583
	10	−0.136	0.512	−0.115	0.508	−0.233	0.576	−0.134	0.514	−0.123	0.510	−0.229	0.573
	30	−0.116	0.505	−0.106	0.500	−0.219	0.571	−0.119	0.504	−0.109	0.501	−0.216	0.568
100	5	0.005	0.469	0.013	0.465	−0.062	0.522	0.004	0.469	0.010	0.466	−0.062	0.521
	10	0.020	0.457	0.027	0.454	−0.050	0.515	0.021	0.458	0.025	0.455	−0.046	0.512
	30	0.030	0.452	0.039	0.447	−0.044	0.512	0.028	0.451	0.035	0.447	−0.040	0.508
1000	5	0.208	0.350	0.211	0.349	0.186	0.374	0.208	0.350	0.210	0.348	0.186	0.373
	10	0.213	0.346	0.215	0.344	0.189	0.371	0.213	0.346	0.214	0.345	0.190	0.371
	30	0.215	0.343	0.217	0.342	0.192	0.369	0.215	0.344	0.217	0.343	0.192	0.369
$\rho_{XX'} = 0.6$													
30	5	0.212	0.754	0.258	0.742	−0.088	0.836	0.206	0.754	0.239	0.746	−0.086	0.834
	10	0.231	0.743	0.277	0.730	−0.067	0.833	0.219	0.744	0.261	0.734	−0.071	0.832
	30	0.239	0.737	0.289	0.723	−0.063	0.831	0.235	0.737	0.273	0.727	−0.059	0.828
50	5	0.325	0.723	0.357	0.713	0.111	0.809	0.322	0.724	0.348	0.718	0.105	0.807
	10	0.338	0.716	0.371	0.704	0.118	0.807	0.335	0.716	0.358	0.707	0.122	0.801
	30	0.349	0.711	0.377	0.697	0.127	0.806	0.343	0.710	0.370	0.702	0.130	0.801
100	5	0.417	0.689	0.439	0.680	0.270	0.770	0.416	0.689	0.430	0.683	0.274	0.768
	10	0.426	0.682	0.448	0.672	0.277	0.768	0.426	0.684	0.440	0.676	0.280	0.768
	30	0.432	0.678	0.452	0.668	0.283	0.767	0.430	0.679	0.446	0.672	0.286	0.764
1000	5	0.537	0.619	0.541	0.616	0.492	0.660	0.537	0.620	0.539	0.617	0.493	0.659
	10	0.539	0.617	0.544	0.613	0.494	0.660	0.539	0.617	0.543	0.615	0.495	0.658
	30	0.541	0.616	0.546	0.612	0.494	0.658	0.540	0.616	0.544	0.613	0.495	0.657
$\rho_{XX'} = 0.8$													
30	5	0.596	0.875	0.646	0.864	0.281	0.930	0.590	0.875	0.630	0.868	0.274	0.928
	10	0.607	0.869	0.655	0.857	0.292	0.929	0.598	0.869	0.641	0.861	0.283	0.926
	30	0.612	0.866	0.663	0.852	0.300	0.927	0.604	0.867	0.645	0.858	0.291	0.926
50	5	0.656	0.860	0.688	0.849	0.436	0.917	0.653	0.860	0.677	0.853	0.433	0.914
	10	0.664	0.855	0.696	0.844	0.444	0.916	0.660	0.856	0.686	0.848	0.439	0.913
	30	0.667	0.853	0.700	0.840	0.444	0.915	0.664	0.853	0.690	0.845	0.443	0.913
100	5	0.704	0.842	0.723	0.833	0.562	0.896	0.703	0.842	0.717	0.837	0.564	0.896
	10	0.708	0.838	0.728	0.829	0.567	0.896	0.706	0.840	0.722	0.833	0.568	0.895
	30	0.711	0.836	0.731	0.827	0.569	0.896	0.710	0.837	0.725	0.830	0.568	0.894
1000	5	0.764	0.807	0.769	0.803	0.726	0.837	0.764	0.807	0.768	0.804	0.728	0.836
	10	0.766	0.805	0.771	0.802	0.728	0.836	0.766	0.806	0.769	0.803	0.728	0.836
	30	0.766	0.805	0.772	0.801	0.728	0.836	0.766	0.805	0.770	0.802	0.729	0.836

dist1, Normal distribution for t_i; dist2, distribution with negative kurtosis for t_i; dist3, distribution with positive kurtosis for t_i; dist4, skewed distribution for t_i; dist5, skewed distribution with negative kurtosis for t_i; dist6, skewed distribution with positive kurtosis for t_i; LB, lower bound; UB, upper bound.

whether the sample coefficient was significantly different from α as described in the previous section.

Guided by these interpretations, one can make the following observations:

1. Among the five non-normal distributions considered for t_i, skewed or platykurtic distributions do not affect the mean or the SE for $\hat{\alpha}$(see **Table 1**). They do not affect the accuracy or bias in estimating α, either (see **Table 2**). On the other hand, symmetric or non-symmetric distributions with positive kurtosis tend to result in a much smaller average of $\hat{\alpha}$ with a larger SE (see **Table 1**), which in turn makes the 95% observed interval wider compared with the normal distribution (see **Table 3**). In addition, positive kurtosis tends to involve more bias in underestimating α with a reduced accuracy (see **Table 2**).

2. Sample size (n) and test length (k) play important roles for $\hat{\alpha}$ and its sampling distribution, as increased n or k tends to result in the mean of $\hat{\alpha}$ that is closer to the specified population reliability ($\rho_{XX'}$) with a smaller SE. We note that n has a larger and more apparent effect than k. Sample size further helps offset the effect of non-normality on the sampling distribution of $\hat{\alpha}$. In particular, when sample size gets large, e.g., $n = 1000$, departure from normal distributions (due to positive kurtosis) does not result in much different mean of $\hat{\alpha}$ although the SE is still slightly larger compared with normal distributions (see **Table 1**).
3. Increased n or k tends to increase the accuracy in estimating α while reducing bias. However, the effect of non-normality (due to positive kurtosis) on resulting in a larger estimating error and bias remains even with increased n and/or k (see **Table 2**). It is also noted that for all the conditions considered, $\hat{\alpha}$ has a consistently negative bias regardless of the shape of the distribution for t_i.
4. The 95% observed interval shown in **Table 3** agrees with the corresponding mean and SE shown in **Table 1**. It is noted that regardless of the population distribution for t_i, when n or k gets larger, $\hat{\alpha}$ has a smaller SE, and hence a narrower 95% interval, as the precision in estimating α increases. Given this, and that all intervals in the table, especially those for $n = 1000$, cover the specified population reliability ($\rho_{XX'}$), one should note that although departure from normality affects the accuracy, bias, and precision in estimating α, it does not result in systematically different $\hat{\alpha}$. In addition, when the actual reliability is small (i.e., $\rho_{XX'} = 0.3$), the use of large n is suggested, as when $n < 1000$, the 95% interval covers negative values of $\hat{\alpha}$. This is especially the case for the (symmetric or non-symmetric) distributions with positive kurtosis. For these distributions, at least 100 subjects are needed for $\hat{\alpha}$ to avoid relatively large estimation error when the actual reliability is moderate to large. For the other distributions, including the normal distribution, a minimum of 50 subjects is suggested for tests with a moderate reliability (i.e., $\rho_{XX'} = 0.6$), and 30 or more subjects are needed for tests with a high reliability (i.e., $\rho_{XX'} = 0.8$; see **Table 2**).

In addition, results for conditions where error scores depart from normal distributions are summarized in **Tables 4–6**. Given the design of the study, the results for the condition where e_{ij} followed a normal distribution are the same as those for the condition where the distribution for t_i was normal. For the purpose of comparisons, they are displayed in the tables again. Inspections of these tables result in the following findings, some of which are quite different from what are observed from **Tables 1–3**:

1. Symmetric platykurtic distributions or non-symmetric leptokurtic distributions consistently resulted in a larger mean but not a larger SE of $\hat{\alpha}$ than normal distributions (see **Table 4**). Some of the means, and especially those for non-symmetric leptokurtic distributions, are larger than the specified population reliability ($\rho_{XX'}$). This is consistent with the positive bias values in **Table 5**. On the other hand, symmetric leptokurtic, non-symmetric, or non-symmetric platykurtic distributions tend to have larger SE of $\hat{\alpha}$ than the normal distribution (see **Table 4**).
2. Sample size (n) and test length (k) have different effects on $\hat{\alpha}$ and its sampling distribution. Increased n consistently results in a larger mean of $\hat{\alpha}$ with a reduced SE. However, increased k may result in a reduced SE, but it has a negative effect on the mean in pushing it away from the specified population reliability ($\rho_{XX'}$), especially when $\rho_{XX'}$ is not large. In particular, with larger k, the mean of $\hat{\alpha}$ decreases to be much smaller for the non-normal distributions that are leptokurtic, non-symmetric, or non-symmetric platykurtic; but it increases to exceed $\rho_{XX'}$ for symmetric platykurtic or non-symmetric leptokurtic distributions. It is further observed that with increased n, the difference between non-normal and normal distributions of e_{ij} on the mean and SE of $\hat{\alpha}$ reduces. This is, however, not observed for increased k (see **Table 4**).
3. The RMSE and *bias* values presented in **Table 5** indicate that non-normal distributions for e_{ij}, especially leptokurtic, non-symmetric, or non-symmetric platykurtic distributions tend to involve larger error, if not bias, in estimating α. In addition, when k increases, RMSE or *bias* does not necessarily reduce. On the other hand, when n increases, RMSE decreases while *bias* increases. Hence, with larger sample sizes, there is more accuracy in estimating α, but bias is not necessarily reduced for symmetric platykurtic or non-symmetric leptokurtic distributions, as some of the negative bias values increase to become positive and non-negligible.
4. The effect of test length on the sample coefficient is more apparent in **Table 6**. From the 95% observed intervals for $\hat{\alpha}$, and particularly those obtained when the actual reliability is small to moderate (i.e., $\rho_{XX'} \leq 0.6$) with large sample sizes (i.e., $n = 1000$), one can see that when test length gets larger (e.g., $k = 30$), the intervals start to fail to cover the specified population reliability ($\rho_{XX'}$) regardless of the degree of the departure from the normality for e_{ij}. Given the fact that larger sample sizes result in less dispersion (i.e., smaller SE) in the sampling distribution of $\hat{\alpha}$ and hence a narrower 95% interval, and the fact that increased k pushes the mean of $\hat{\alpha}$ away from the specified reliability, this finding suggests that larger k amplifies the effect of non-normality of e_{ij} on $\hat{\alpha}$ in resulting in systematically biased estimates of α, and hence has to be avoided when the actual reliability is not large. With respect to sample sizes, similar patterns arise. That is, the use of large n is suggested when the actual reliability is small (i.e., $\rho_{XX'} = 0.3$), especially for tests with 30 items, whereas for tests with a high reliability (i.e., $\rho_{XX'} = 0.8$), a sample size of 30 may be sufficient. In addition, when the actual reliability is moderate, a minimum of 50 subjects is needed for $\hat{\alpha}$ to be fairly accurate for short tests ($k \leq 10$), and at least 100 are suggested for longer tests ($k = 30$; see **Table 5**).

Given the above results, we see that non-normal distributions for true or error scores do create problems for using coefficient alpha to estimate the internal consistency reliability. In particular, leptokurtic true score distributions that are either symmetric or skewed result in larger error and negative bias in estimating population α with less precision. This is similar to Bay's (1973) finding, and we see in this study that the problem remains even after increasing sample size to 1000 or test length to 30, although the effect is getting smaller. With respect to error score

Table 4 | Observed mean and SD of the sample alpha ($\hat{\alpha}$) for the simulated situations where the error score (e_{ij}) distribution is normal or non−normal.

n	*k*	Mean ($\hat{\alpha}$)						SD ($\hat{\alpha}$)					
		dist1	**dist2**	**dist3**	**dist4**	**dist5**	**dist6**	**dist1**	**dist2**	**dist3**	**dist4**	**dist5**	**dist6**
$\rho_{XX'} = 0.3$													
30	5	0.230	0.255	0.215	0.206	0.213	0.313	0.241	0.233	0.257	0.252	0.250	0.223
	10	0.231	0.295	0.158	0.174	0.185	0.367	0.229	0.207	0.256	0.248	0.248	0.195
	30	0.231	0.371	0.103	0.155	0.139	0.460	0.221	0.177	0.258	0.244	0.249	0.160
50	5	0.252	0.279	0.232	0.231	0.237	0.324	0.176	0.169	0.191	0.183	0.181	0.166
	10	0.252	0.316	0.180	0.198	0.213	0.380	0.166	0.150	0.187	0.180	0.178	0.143
	30	0.254	0.390	0.128	0.181	0.164	0.474	0.160	0.128	0.188	0.176	0.181	0.116
100	5	0.269	0.295	0.245	0.247	0.255	0.332	0.118	0.113	0.130	0.124	0.122	0.115
	10	0.268	0.331	0.196	0.216	0.229	0.390	0.112	0.101	0.128	0.121	0.120	0.098
	30	0.269	0.403	0.146	0.198	0.182	0.484	0.108	0.086	0.127	0.119	0.122	0.078
1000	5	0.282	0.308	0.254	0.261	0.269	0.338	0.036	0.035	0.040	0.038	0.037	0.036
	10	0.282	0.343	0.208	0.231	0.244	0.398	0.034	0.031	0.039	0.037	0.037	0.030
	30	0.282	0.414	0.161	0.213	0.197	0.493	0.033	0.026	0.039	0.036	0.037	0.024
$\rho_{XX'} = 0.6$													
30	5	0.549	0.550	0.565	0.552	0.551	0.571	0.142	0.140	0.163	0.141	0.140	0.159
	10	0.551	0.560	0.547	0.543	0.545	0.586	0.133	0.127	0.157	0.141	0.139	0.132
	30	0.550	0.615	0.452	0.482	0.500	0.669	0.129	0.102	0.180	0.160	0.156	0.093
50	5	0.563	0.564	0.574	0.565	0.564	0.579	0.103	0.100	0.121	0.103	0.101	0.118
	10	0.563	0.573	0.559	0.557	0.560	0.595	0.097	0.092	0.115	0.102	0.100	0.099
	30	0.563	0.625	0.472	0.499	0.518	0.676	0.093	0.074	0.131	0.115	0.112	0.068
100	5	0.572	0.573	0.579	0.574	0.574	0.584	0.069	0.068	0.084	0.069	0.068	0.082
	10	0.572	0.582	0.567	0.567	0.570	0.600	0.066	0.062	0.078	0.069	0.067	0.068
	30	0.572	0.633	0.484	0.511	0.530	0.681	0.063	0.050	0.088	0.078	0.075	0.046
1000	5	0.580	0.581	0.583	0.582	0.582	0.588	0.021	0.021	0.026	0.021	0.021	0.026
	10	0.580	0.589	0.574	0.576	0.578	0.605	0.020	0.019	0.024	0.021	0.020	0.021
	30	0.580	0.638	0.496	0.522	0.540	0.686	0.019	0.015	0.027	0.024	0.023	0.014
$\rho_{XX'} = 0.8$													
30	5	0.771	0.771	0.777	0.772	0.772	0.779	0.072	0.070	0.094	0.072	0.070	0.092
	10	0.771	0.771	0.776	0.773	0.772	0.777	0.068	0.067	0.081	0.068	0.067	0.080
	30	0.771	0.782	0.760	0.763	0.766	0.798	0.066	0.058	0.085	0.076	0.073	0.057
50	5	0.778	0.778	0.782	0.779	0.779	0.783	0.052	0.051	0.070	0.052	0.051	0.070
	10	0.778	0.778	0.781	0.779	0.779	0.784	0.049	0.048	0.060	0.049	0.048	0.059
	30	0.778	0.788	0.768	0.771	0.774	0.802	0.048	0.042	0.060	0.054	0.052	0.042
100	5	0.782	0.783	0.785	0.783	0.783	0.787	0.035	0.034	0.049	0.035	0.034	0.049
	10	0.783	0.783	0.785	0.784	0.784	0.787	0.033	0.033	0.041	0.033	0.033	0.041
	30	0.783	0.792	0.774	0.777	0.779	0.805	0.032	0.028	0.040	0.036	0.034	0.029
1000	5	0.786	0.787	0.788	0.787	0.787	0.789	0.011	0.010	0.015	0.011	0.011	0.015
	10	0.786	0.787	0.788	0.788	0.788	0.791	0.010	0.010	0.013	0.010	0.010	0.013
	30	0.787	0.795	0.779	0.781	0.784	0.807	0.010	0.009	0.012	0.011	0.010	0.009

dist1, Normal distribution for e_{ij}; dist2, distribution with negative kurtosis for e_{ij}; dist3, distribution with positive kurtosis for e_{ij}; dist4, skewed distribution for e_{ij}; dist5, skewed distribution with negative kurtosis for e_{ij}; dist6, skewed distribution with positive kurtosis for e_{ij}.

distributions, unlike conclusions from previous studies, departure from normality does create problems in the sample coefficient alpha and its sampling distribution. Specifically, leptokurtic, skewed, or non-symmetric platykurtic error score distributions tend to result in larger error and negative bias in estimating population α with less precision, whereas platykurtic or non-symmetric leptokurtic error score distributions tend to have increased positive bias when sample size, test length, and/or the actual reliability increases. In addition, different from conclusions made by Bay (1973) and Shultz (1993), an increase in test length does have an effect on the accuracy and bias in estimating reliability with the sample coefficient alpha when error scores are not normal, but it is in an undesirable manner. In particular, as is noted earlier, increased test length pushes the mean of $\hat{\alpha}$ away from the actual reliability, and hence causes the sample coefficient alpha to be significantly different from the population coefficient when the actual reliability is not high (e.g., $\rho_{XX'} \leq 0.6$) and the sample size is large (e.g., $n = 1000$). This could be due to the fact that e_{ij} is involved

Table 5 | Root mean square error and bias for estimating α for the simulated situations where the error score (e_{ij}) distribution is normal or non−normal.

n	*k*	RMSE						bias					
		dist1	dist2	dist3	dist4	dist5	dist6	dist1	dist2	dist3	dist4	dist5	dist6
$\rho_{XX'} = 0.3$													
30	5	0.251	0.237	0.271	0.269	0.264	0.224	−0.070	−0.045	−0.085	−0.094	−0.087	0.013
	10	0.240	0.208	0.293	0.278	0.273	0.206	−0.069	−0.005	−0.142	−0.126	−0.115	0.067
	30	0.232	0.191	0.325	0.284	0.297	0.226	−0.069	0.071	−0.197	−0.145	−0.162	0.160
50	5	0.182	0.170	0.202	0.195	0.192	0.167	−0.048	−0.022	−0.068	−0.069	−0.063	0.024
	10	0.173	0.151	0.222	0.207	0.198	0.164	−0.048	0.016	−0.120	−0.103	−0.088	0.080
	30	0.166	0.157	0.255	0.212	0.226	0.209	−0.046	0.090	−0.172	−0.119	−0.136	0.174
100	5	0.122	0.113	0.141	0.135	0.130	0.119	−0.031	−0.006	−0.055	−0.053	−0.045	0.032
	10	0.116	0.106	0.165	0.148	0.140	0.133	−0.032	0.031	−0.105	−0.084	−0.071	0.090
	30	0.112	0.134	0.200	0.157	0.170	0.200	−0.031	0.103	−0.154	−0.102	−0.118	0.184
1000	5	0.040	0.035	0.061	0.054	0.048	0.052	−0.018	0.008	−0.046	−0.039	−0.031	0.038
	10	0.038	0.053	0.100	0.079	0.067	0.102	−0.018	0.043	−0.092	−0.070	−0.056	0.098
	30	0.038	0.117	0.144	0.095	0.109	0.194	−0.018	0.114	−0.139	−0.087	−0.103	0.193
$\rho_{XX'} = 0.6$													
30	5	0.151	0.148	0.166	0.149	0.149	0.161	−0.051	−0.050	−0.035	−0.048	−0.050	−0.029
	10	0.142	0.133	0.165	0.152	0.149	0.133	−0.050	−0.040	−0.053	−0.057	−0.055	−0.014
	30	0.139	0.103	0.233	0.199	0.185	0.116	−0.050	0.015	−0.148	−0.118	−0.100	0.069
50	5	0.109	0.106	0.124	0.108	0.107	0.120	−0.037	−0.036	−0.026	−0.035	−0.036	−0.021
	10	0.104	0.096	0.123	0.111	0.107	0.099	−0.037	−0.027	−0.041	−0.043	−0.040	−0.005
	30	0.100	0.078	0.183	0.153	0.139	0.102	−0.037	0.025	−0.128	−0.101	−0.082	0.076
100	5	0.075	0.073	0.087	0.074	0.073	0.084	−0.028	−0.027	−0.022	−0.026	−0.026	−0.016
	10	0.071	0.065	0.085	0.076	0.074	0.068	−0.028	−0.018	−0.033	−0.033	−0.031	0.000
	30	0.069	0.060	0.146	0.118	0.103	0.093	−0.028	0.033	−0.116	−0.089	−0.070	0.081
1000	5	0.029	0.028	0.032	0.028	0.028	0.029	−0.020	−0.019	−0.017	−0.018	−0.018	−0.012
	10	0.028	0.022	0.036	0.032	0.030	0.022	−0.020	−0.011	−0.026	−0.024	−0.022	0.005
	30	0.028	0.041	0.108	0.082	0.064	0.087	−0.020	0.038	−0.104	−0.078	−0.060	0.086
$\rho_{XX'} = 0.8$													
30	5	0.078	0.076	0.097	0.077	0.076	0.095	−0.030	−0.029	−0.023	−0.028	−0.029	−0.021
	10	0.074	0.073	0.085	0.073	0.073	0.084	−0.029	−0.029	−0.024	−0.027	−0.028	−0.023
	30	0.072	0.060	0.094	0.085	0.080	0.057	−0.029	−0.018	−0.040	−0.037	−0.034	−0.003
50	5	0.057	0.055	0.073	0.057	0.055	0.072	−0.022	−0.022	−0.018	−0.021	−0.021	−0.017
	10	0.054	0.053	0.063	0.053	0.053	0.062	−0.022	−0.022	−0.019	−0.021	−0.021	−0.016
	30	0.052	0.044	0.068	0.061	0.058	0.042	−0.022	−0.012	−0.032	−0.029	−0.026	0.002
100	5	0.039	0.038	0.051	0.039	0.038	0.050	−0.018	−0.017	−0.015	−0.017	−0.017	−0.013
	10	0.038	0.037	0.044	0.037	0.037	0.043	−0.018	−0.017	−0.015	−0.016	−0.016	−0.013
	30	0.037	0.030	0.048	0.043	0.040	0.029	−0.018	−0.008	−0.026	−0.024	−0.021	0.005
1000	5	0.017	0.017	0.020	0.017	0.017	0.019	−0.014	−0.013	−0.012	−0.013	−0.013	−0.011
	10	0.017	0.016	0.017	0.016	0.016	0.016	−0.014	−0.013	−0.012	−0.012	−0.012	−0.010
	30	0.017	0.010	0.024	0.022	0.019	0.012	−0.014	−0.005	−0.021	−0.019	−0.016	0.007

dist1, Normal distribution for e_{ij}; dist2, distribution with negative kurtosis for e_{ij}; dist3, distribution with positive kurtosis for e_{ij}; dist4, skewed distribution for e_{ij}; dist5, skewed distribution with negative kurtosis for e_{ij}; dist6, skewed distribution with positive kurtosis for e_{ij}.

in each item, and hence an increase in the number of items would add up the effect of non-normality on the sample coefficient.

DISCUSSION

In practice, coefficient alpha is often used to estimate reliability with little consideration of the assumptions required for the sample coefficient to be accurate. As noted by Graham (2006, p. 942), students and researchers in education and psychology are often unaware of many assumptions for a statistical procedure, and this situation is much worse when it comes to measurement issues such as reliability. In actual applications, it is vital to not only evaluate the assumptions for coefficient alpha, but also understand them and the consequences of any violations.

Normality is not commonly considered as a major assumption for coefficient alpha and hence has not been well investigated. This study takes the advantage of recently developed techniques in generating univariate non-normal data to suggest that different from conclusions made by Bay (1973), Zimmerman et al. (1993),

Table 6 | Observed 95% interval of the sample alpha ($\hat{\alpha}$) for the simulated situations where the error score (e_{ij}) distribution is normal or non-normal.

n	*k*	dist1		dist2		dist3		dist4		dist5		dist6	
		LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB
$\rho_{XX'} = 0.3$													
30	5	−0.351	0.580	−0.305	0.592	−0.404	0.587	−0.403	0.572	−0.392	0.574	−0.220	0.638
	10	−0.323	0.563	−0.203	0.595	−0.459	0.529	−0.421	0.536	−0.416	0.543	−0.104	0.647
	30	−0.303	0.550	−0.058	0.629	−0.522	0.478	−0.437	0.508	−0.462	0.501	0.070	0.688
50	5	−0.155	0.528	−0.114	0.544	−0.212	0.531	−0.193	0.516	−0.179	0.519	−0.057	0.585
	10	−0.136	0.512	−0.031	0.551	−0.256	0.475	−0.223	0.481	−0.199	0.493	0.050	0.604
	30	−0.116	0.505	0.094	0.591	−0.307	0.423	−0.229	0.458	−0.257	0.447	0.203	0.653
100	5	0.005	0.469	0.042	0.486	−0.045	0.464	−0.031	0.456	−0.016	0.461	0.077	0.525
	10	0.020	0.457	0.107	0.501	−0.087	0.411	−0.052	0.421	−0.039	0.431	0.172	0.554
	30	0.030	0.452	0.211	0.549	−0.136	0.361	−0.067	0.399	−0.089	0.387	0.310	0.616
1000	5	0.208	0.350	0.237	0.372	0.172	0.330	0.184	0.332	0.193	0.338	0.264	0.405
	10	0.213	0.346	0.280	0.400	0.128	0.282	0.155	0.300	0.170	0.313	0.336	0.454
	30	0.215	0.343	0.360	0.463	0.082	0.234	0.139	0.280	0.122	0.267	0.444	0.537
$\rho_{XX'} = 0.6$													
30	5	0.212	0.754	0.212	0.752	0.171	0.794	0.212	0.756	0.210	0.752	0.186	0.796
	10	0.231	0.743	0.253	0.745	0.167	0.768	0.197	0.744	0.210	0.743	0.267	0.778
	30	0.239	0.737	0.370	0.766	0.015	0.711	0.099	0.713	0.121	0.721	0.444	0.803
50	5	0.325	0.723	0.331	0.722	0.289	0.758	0.326	0.725	0.329	0.722	0.302	0.762
	10	0.338	0.716	0.360	0.717	0.286	0.736	0.319	0.715	0.328	0.714	0.363	0.749
	30	0.349	0.711	0.455	0.743	0.167	0.675	0.229	0.680	0.256	0.691	0.520	0.783
100	5	0.417	0.689	0.422	0.688	0.389	0.716	0.420	0.691	0.422	0.689	0.398	0.721
	10	0.426	0.682	0.444	0.687	0.392	0.697	0.415	0.682	0.420	0.682	0.448	0.714
	30	0.432	0.678	0.521	0.718	0.287	0.632	0.338	0.643	0.363	0.656	0.579	0.759
1000	5	0.537	0.619	0.539	0.620	0.528	0.631	0.539	0.621	0.539	0.620	0.534	0.636
	10	0.539	0.617	0.551	0.624	0.524	0.619	0.533	0.614	0.537	0.616	0.562	0.644
	30	0.541	0.616	0.607	0.667	0.441	0.546	0.473	0.566	0.494	0.582	0.657	0.712
$\rho_{XX'} = 0.8$													
30	5	0.596	0.875	0.602	0.872	0.543	0.903	0.598	0.876	0.602	0.874	0.549	0.904
	10	0.607	0.869	0.611	0.868	0.575	0.889	0.608	0.870	0.611	0.869	0.581	0.890
	30	0.612	0.866	0.646	0.868	0.550	0.875	0.577	0.867	0.589	0.867	0.661	0.882
50	5	0.656	0.860	0.661	0.858	0.613	0.885	0.657	0.861	0.661	0.858	0.617	0.885
	10	0.664	0.855	0.667	0.854	0.637	0.872	0.665	0.856	0.667	0.855	0.644	0.873
	30	0.667	0.853	0.692	0.855	0.624	0.858	0.643	0.853	0.653	0.853	0.705	0.869
100	5	0.704	0.842	0.707	0.840	0.672	0.863	0.705	0.842	0.707	0.841	0.675	0.863
	10	0.708	0.838	0.711	0.838	0.692	0.853	0.711	0.840	0.711	0.839	0.696	0.854
	30	0.711	0.836	0.729	0.841	0.683	0.840	0.695	0.836	0.702	0.836	0.741	0.854
1000	5	0.764	0.807	0.766	0.806	0.756	0.816	0.766	0.807	0.766	0.807	0.757	0.817
	10	0.766	0.805	0.767	0.806	0.762	0.812	0.767	0.807	0.767	0.806	0.765	0.814
	30	0.766	0.805	0.778	0.812	0.754	0.802	0.759	0.802	0.762	0.803	0.789	0.824

dist1, Normal distribution for e_{ij}; dist2, distribution with negative kurtosis for e_{ij}; dist3, distribution with positive kurtosis for e_{ij}; dist4, skewed distribution for e_{ij}; dist5, skewed distribution with negative kurtosis for e_{ij}; dist6, skewed distribution with positive kurtosis for e_{ij}; LB, lower bound; UB, upper bound.

and Shultz (1993), coefficient alpha is not robust to the violation of the normal assumption (for either true or error scores). Non-normal data tend to result in additional error or bias in estimating internal consistency reliability. A larger error makes the sample coefficient less accurate, whereas more bias causes it to further under- or overestimate the actual reliability. We note that compared with normal data, leptokurtic true or error score distributions tend to result in additional negative bias, whereas platykurtic error score distributions tend to result in a positive bias. Neither case is desired in a reliability study, as the sample coefficient would paint an incorrect picture of the test's internal consistency by either estimating it with a larger value or a much smaller value and hence is not a valid indicator. For example, for a test with reliability being 0.6, one may calculate the sample alpha to be 0.4 because the true score distribution has a positive kurtosis, and conclude that the test is not reliable at all. On the other hand, one may have a test with actual reliability being 0.4. But because the error score distribution has a negative kurtosis, the sample

coefficient is calculated to be 0.7 and hence the test is concluded to be reliable. In either scenario, the conclusion on the test reliability is completely the opposite of the true situation, which may lead to an overlook of a reliable measure or an adoption of an unreliable instrument. Consequently, coefficient alpha is not suggested for estimating internal consistency reliability with non-normal data. Given this, it is important to make sure that in addition to satisfying the assumptions of (essential) tau-equivalence and uncorrelated errors, the sample data conform to normal distributions before one uses alpha in a reliability study.

Further, it is generally said that increased data sizes help approximate non-normal distributions to be normal. This is the case with sample sizes, not necessarily test lengths, in helping improve the accuracy, bias and/or precision of using the sample coefficient in reliability studies with non-normal data. Given the results of the study, we suggest that in order for the sample coefficient alpha to be fairly accurate and in a reasonable range, a minimum of 1000 subjects is needed for a small reliability, and a minimum of 100 is needed for a moderate reliability when the sample data depart from normality. It has to be noted that for the four sample size conditions considered in the study, the sample coefficient alpha consistently underestimates the population reliability even when normality is assumed (see **Table 2**). However, the degree of bias becomes negligible when sample size increases to 1000 or beyond.

In the study, we considered tests of 5, 10, or 30 items administered to 30, 50, 100, or 1000 persons with the actual reliability being 0.3, 0.6, or 0.8. These values were selected to reflect levels ranging from small to large in the sample size, test length, and population reliability considerations. When using the results, one should note that they pertain to these simulated conditions and may not generalize to other conditions. In addition, we evaluated the assumption of normality alone. That is, in the simulations, data were generated assuming the other assumptions, namely (essential) tau-equivalence and uncorrelated error terms, were satisfied. In practice, it is common for observed data to violate more than one assumption. Hence, it would also be interesting to see how non-normal data affect the sample coefficient when other violations are present. Further, this study looked at the sample coefficient alpha and its empirical sampling distribution without considering its sampling theory (e.g., Kristof, 1963; Feldt, 1965). One may focus on its theoretical SE (e.g., Bay, 1973; Barchard and Hakstian, 1997a,b; Duhachek and Iacobucci, 2004) and compare them with the empirical ones to evaluate the robustness of an interval estimation of the reliability for non-normal data.

REFERENCES

Barchard, K. A., and Hakstian, R. (1997a). The effects of sampling model on inference with coefficient alpha. *Educ. Psychol. Meas.* 57, 893–905.

Barchard, K. A., and Hakstian, R. (1997b). The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behav. Res.* 32, 169–191.

Bay, K. S. (1973). The effect of non-normality on the sampling distribution and standard error of reliability coefficient estimates under an analysis of variance model. *Br. J. Math. Stat. Psychol.* 26, 45–57.

Caplan, R. D., Naidu, R. K., and Tripathi, R. C. (1984). Coping and defense: constellations vs. components. *J. Health Soc. Behav.* 25, 303–320.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334.

DeVellis, R. F. (1991). *Scale Development.* Newbury Park, NJ: Sage Publications.

Duhachek, A., and Iacobucci, D. (2004). Alpha's standard error (ASE): an accurate and precise confidence interval estimate. *J. Appl. Psychol.* 89, 792–808.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika* 30, 357–370.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika* 43, 521–532.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiment.* New York: Wiley.

Graham, J. M. (2006). Congeneric and (essential) tau-equivalent estimates of score reliability. *Educ. Psychol. Meas.* 66, 930–944.

Green, S. B., and Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Struct. Equation Model.* 7, 251–270.

Green, S. B., and Yang, Y. (2009). Commentary on coefficient alpha: a cautionary tale. *Psychometrika* 74, 121–135.

Guttman, L. A. (1945). A basis for analyzing test-retest reliability. *Psychometrika* 10, 255–282.

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Comput. Stat. Data Anal.* 40, 685–711.

Headrick, T. C. (2010). *Statistical Simulation: Power Method Polynomials and Other Transformations.* Boca Raton, FL: Chapman & Hall.

Johanson, G. A., and Brooks, G. (2010). Initial scale development: sample size for pilot studies. *Educ. Psychol. Meas.* 70, 394–400.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika* 28, 221–238.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading: Addison-Wesley.

MathWorks. (2010). *MATLAB* (Version 7.11) [Computer software]. Natick, MA: MathWorks.

Miller, M. B. (1995). Coefficient alpha: a basic introduction from the perspectives of classical test theory and structural equation modeling. *Struct. Equation Model.* 2, 255–273.

Novick, M. R., and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika* 32, 1–13.

Nunnally, J. C. (1967). *Psychometric Theory.* New York: McGraw-Hill.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory*, 3rd Edn. New York: McGraw-Hill.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Appl. Psychol. Meas.* 22, 69–76.

Shultz, G. S. (1993). *A Monte Carlo study of the robustness of coefficient alpha.* Masters thesis, University of Ottawa, Ottawa.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74, 107–120.

van Zyl, J. M., Neudecker, H., and Nel, D. G. (2000). On the distribution of the maximum likelihood estimator for Cronbach's alpha. *Psychometrika* 65, 271–280.

Zimmerman, D. W., Zumbo, B. D., and Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educ. Psychol. Meas.* 53, 33–49.

Zumbo, B. D. (1999). *A Glance at Coefficient Alpha With an Eye Towards Robustness Studies: Some Mathematical Notes and a Simulation Model (Paper No. ESQBS-99-1).* Prince George, BC: Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia.

Zumbo, B. D., and Rupp, A. A. (2004). "Responsible modeling of measurement data for appropriate inferences: important advances in reliability and validity theory," in *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, ed. D. Kaplan (Thousand Oaks: Sage), 73–92.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 October 2011; paper pending published: 22 November 2011; accepted: 30 January 2012; published online: 15 February 2012.

*Citation: Sheng Y and Sheng Z (2012) Is coefficient alpha robust to non-normal data? Front. Psychology **3**:34. doi: 10.3389/fpsyg.2012.00034*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

APPENDIX

CODE IN MATLAB

```
function result=mcalpha(n,k,evar,rho,rep)
%
% mcalpha - obtain summary statistics for sample alphas
%
%  result=mcalpha(n,k,evar,rho,rep)
%
% returns the observed mean, standard deviation, and 95% interval (qtalpha)
% for sample alphas as well as the root mean square error (rmse) and bias for
% estimating the population alpha.
%
% The INPUT arguments:
%    n - sample size
%    k - test length
%    evar - error variance
%    rho - population reliability
%    rep - number of replications
%
       alphav=zeros(rep,1);
       tbcd=[0,1,0,0,0,0];
       ebcd=[0,1,0,0,0,0];
%
% note: tbcd and ebcd are vectors containing the six coefficients, c0,…,c5,
% used in equation (8) for true scores and error scores, respectively. Each
% of them can be set as:
% 1. [0,1,0,0,0,0] (normal)
% 2. [0,1.643377,0,-.319988,0,.011344] (platykurtic)
% 3. [0,0.262543,0,.201036,0,.000162] (leptokurtic)
% 4. [-0.446924 1.242521 0.500764 -0.184710 -0.017947,0.003159] (skewed)
% 5. [-.276330,1.506715,.311114,-.274078,-.011595,.007683] (skewed
% platykurtic)
% 6. [-.304852,.381063,.356941,.132688,-.017363,.003570] (skewed leptokurtic)
%
       for i=1:rep
           alphav(i)=alpha(n,k,evar,rho,tbcd,ebcd);
       end
       rmse=sqrt(mean((alphav-rho).^2));
       bias=mean(alphav-rho);
       qtalpha=quantile(alphav,[.025,.975]);
       result=[mean(alphav),std(alphav),qtalpha,rmse,bias];

function A=alpha(n,k,evar,rho,tbcd,ebcd)
%
% alpha - calculate sample alpha
%
%  alp=alpha(n,k,evar,rho,tbcd,ebcd)
%
% returns the sample alpha.
%
% The INPUT arguments:
%      n - sample size
%      k - test length
%      evar - error variance
%      rho - population reliability
```

```
%       rep - number of replications
%       tbcd - coefficients for generating normal/nonnormal true score
%              distributions using power method polynomials
%       ebcd - coefficients for generating normal/nonnormal error score
%              distributions using power method polynomials
%
        tvar=evar*rho/((1-rho)*k);
        t=rfsimu(tbcd,n,1,5,tvar);
        e=rfsimu(ebcd,n,k,0,evar);
        xn=t*ones(1,k)+e;
        x=round(xn);
        alp=k/(k-1)*(1-sum(var(x,1))/var(sum(x,2),1));

function X=rfsimu(bcd,n,k,mean,var)
%
% rfsimu - generate normal/nonnormal distributions using 5-th order power
%  method polynomials
%
%  X=rfsimu(bcd,n,k,mean,var)
%
% returns samples of size n by k drawn from a distribution with the desired
% moments.
%
% The INPUT arguments:
%       bcd - coefficients for generating normal/nonnormal distributions using
%             the 5-th order polynomials
%       k - test length
%       evar - error variance
%       rho - population reliability
%       rep - number of replications
%       tbcd - coefficients for generating normal/nonnormal true score
%              distributions using power method polynomials
%       ebcd - coefficients for generating normal/nonnormal error score
%              distributions using power method polynomials
%
        Z=randn(n,k);
        Y=bcd(1)+bcd(2)*Z+bcd(3)*Z.^2+bcd(4)*Z.^3+bcd(5)*Z.^4+bcd(6)*Z.^5;
        X=mean+sqrt(var)*Y;
```

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 12 April 2012
doi: 10.3389/fpsyg.2012.00102

The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data

***Kim Nimon*[1]*, *Linda Reichwein Zientek*[2] *and Robin K. Henson*[3]**

[1] *Department of Learning Technologies, University of North Texas, Denton, TX, USA*
[2] *Department of Mathematics and Statistics, Sam Houston State University, Huntsville, TX, USA*
[3] *Department of Educational Psychology, UNT, University of North Texas, Denton, TX, USA*

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Matthew D. Finkelman, Tufts University, USA
Martin Lages, University of Glasgow, UK

***Correspondence:**
Kim Nimon, Department of Learning Technologies, University of North Texas, 3940 North Elm Street, G150, Denton, TX 76207, USA.
e-mail: kim.nimon@unt.edu

The purpose of this article is to help researchers avoid common pitfalls associated with reliability including *incorrectly* assuming that (a) measurement error always attenuates observed score correlations, (b) different sources of measurement error originate from the same source, and (c) reliability is a function of instrumentation. To accomplish our purpose, we first describe what reliability is and why researchers should care about it with focus on its impact on effect sizes. Second, we review how reliability is assessed with comment on the consequences of cumulative measurement error. Third, we consider how researchers can use reliability generalization as a prescriptive method when designing their research studies to form hypotheses about whether or not reliability estimates will be acceptable given their sample and testing conditions. Finally, we discuss options that researchers may consider when faced with analyzing unreliable data.

Keywords: reliability, measurement error, correlated error

The vast majority of commonly used parametric statistical procedures assume data are measured without error (Yetkiner and Thompson, 2010). However, research indicates that there are at least three problems concerning application of the statistical assumption of reliable data. First and foremost, researchers frequently neglect to report reliability coefficients for their data (Vacha-Haase et al., 1999, 2002; Zientek et al., 2008). Presumably, these same researchers fail to consider if data are reliable and thus ignore the consequences of results based on data that are confounded with measurement error. Second, researchers often reference reliability coefficients from test manuals or prior research presuming that the same level of reliability applies to their data (Vacha-Haase et al., 2000). Such statements ignore admonitions from Henson (2001), Thompson (2003a), Wilkinson and APA Task Force on Statistical Inference (1999), and others stating that *reliability is a property inured to scores not tests*. Third, researchers that do consider the reliability of their data may attempt to correct for measurement error by applying Spearman's (1904) correction formula to sample data without considering how error in one variable relates to observed score components in another variable or the true score component of its own variable (cf. Onwuegbuzie et al., 2004; Lorenzo-Seva et al., 2010). These so-called *nuisance correlations*, however, can seriously influence the accuracy of the statistics that have been corrected by Spearman's formula (Wetcher-Hendricks, 2006; Zimmerman, 2007). In fact, as readers will see, the term *correction for attenuation* may be considered a misnomer as unreliable data do not always produce effects that are smaller than they would have been had data been measured with perfect reliability.

PURPOSE

The purpose of this article is to help researchers avoid common pitfalls associated with reliability including *incorrectly* assuming that (a) measurement error always attenuates observed score correlations, (b) different sources of measurement error originate from the same source, and (c) reliability is a function of instrumentation. To accomplish our purpose, the paper is organized as follows.

First, we describe what reliability is and why researchers should care about it. We focus on bivariate correlation (r) and discuss how reliability affects its magnitude. [Although the discussion is limited to r for brevity, the implications would likely extend to many other commonly used parametric statistical procedures (e.g., t-test, analysis of variance, canonical correlation) because many are "correlational in nature" (Zientek and Thompson, 2009, p. 344) and "yield variance-accounted-for effect sizes analogous to r^2" (Thompson, 2000, p. 263).] We present empirical evidence that demonstrates why measurement error does not always attenuate observed score correlations and why simple steps that attempt to correct for unreliable data may produce misleading results. Second, we review how reliability is commonly assessed. In addition to describing several techniques, we highlight the cumulative nature of different types of measurement error. Third, we consider how researchers can use reliability generalization (RG) as a prescriptive method when designing their research studies to form hypotheses about whether or not reliability estimates will be acceptable given their sample and testing conditions. In addition to reviewing RG theory and studies that demonstrate that reliability is a function of data and not instrumentation, we review barriers to conducting RG studies and propose a set of metrics to be included in research reports. It is our hope that editors will champion the inclusion of such data and thereby broaden what is known about the reliability of educational and psychological data published in research reports. Finally, we discuss options that researchers may consider when faced with analyzing unreliable data.

RELIABILITY: WHAT IS IT AND WHY DO WE CARE?

The predominant applied use of reliability is framed by classical test theory (CTT, Hogan et al., 2000) which conceptualizes observed scores into two independent additive components: (a) true scores and (b) error scores:

$$\text{Observed Score } (O_X) = \text{True Score } (T_X) + \text{Error Score } (E_X) \quad (1)$$

True scores reflect the construct of interest (e.g., depression, intelligence) while error scores reflect error in the measurement of the construct of interest (e.g., misunderstanding of items, chance responses due to guessing). Error scores are referred to as measurement error (Zimmerman and Williams, 1977) and stem from random and systematic occurrences that keep observed data from conveying the "truth" of a situation (Wetcher-Hendricks, 2006, p. 207). Systematic measurement errors are "those which consistently affect an individual's score because of some particular characteristic of the person or the test that has nothing to do with the construct being measured" (Crocker and Algina, 1986, p. 105). Random errors of measurement are those which "affect an individual's score because of purely chance happenings" (Crocker and Algina, 1986, p. 106).

The ratio between true score variance and observed score variance is referred to as reliability. In data measured with perfect reliability, the ratio between true score variance and observed score variance is 1 (Crocker and Algina, 1986). However, the nature of educational and psychological research means that most, if not all, variables are difficult to measure and yield reliabilities less than 1 (Osborne and Waters, 2002).

Researchers should care about reliability as the vast majority of parametric statistical procedures assume that sample data are measured without error (cf. Yetkiner and Thompson, 2010). Poor reliability even presents a problem for descriptive statistics such as the mean because part of the average score is actually error. It also causes problems for statistics that consider variable relationships because poor reliability impacts the magnitude of those results. Measurement error is even a problem in structural equation model (SEM) analyses, as poor reliability affects overall fit statistics (Yetkiner and Thompson, 2010). In this article, though, we focus our discussion on statistical analyses based on observed variable analyses because latent variable analyses are reported less frequently in educational and psychological research (cf. Kieffer et al., 2001; Zientek et al., 2008).

Contemporary literature suggests that unreliable data always attenuate observed score variable relationships (e.g., Muchinsky, 1996; Henson, 2001; Onwuegbuzie et al., 2004). Such literature stems from Spearman's (1904) correction formula that estimates a true score correlation ($r_{T_X T_Y}$) by dividing an observed score correlation ($r_{O_X O_Y}$) by the square root of the product of reliabilities ($r_{XX} r_{YY}$):

$$r_{T_X T_Y} = \frac{r_{O_X O_Y}}{\sqrt{r_{XX} r_{YY}}} \quad (2)$$

Spearman's formula suggests that the observed score correlation is solely a function of the true score correlation and the reliability of the measured variables such that the observed correlation between two variables can be no greater than the square root of the product of their reliabilities:

$$r_{O_X O_Y} = r_{T_X T_Y} \sqrt{r_{XX} r_{YY}} \quad (3)$$

Using derivatives of Eqs 2 and 3, Henson (2001) claimed, for example, that if one variable was measured with 70% reliability and another variable was measured with 60% reliability, the maximum possible observed score correlation would be 0.65 (i.e., $\sqrt{0.70 \times 0.60}$). He similarly indicated that the observed correlation between two variables will only reach its theoretical maximum of 1 when (a) the reliability of the variables are perfect and (b) the correlation between the true score equivalents is equal to 1. (Readers may also consult Trafimow and Rice, 2009 for an interesting application of this correction formula to behavioral task performance via potential performance theory).

The problem with the aforementioned claims is that they do not consider how error in one variable relates to observed score components in another variable or the true score component of its own variable. In fact, Eqs 2 and 3, despite being written for sample data, should only be applied to population data in the case when error does not correlate or share common variance (Zimmerman, 2007), as illustrated in **Figure 1**. However, in the case of correlated error in the population (see **Figure 2**) or in the case of sample

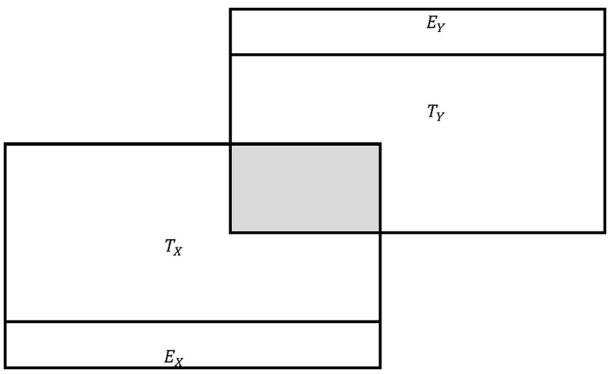


FIGURE 1 | Conceptual representation of observed score variance in the case of no correlated error. *T*, true scores; *E*, error scores. Note: shaded area denotes shared variance between T_X and T_Y.

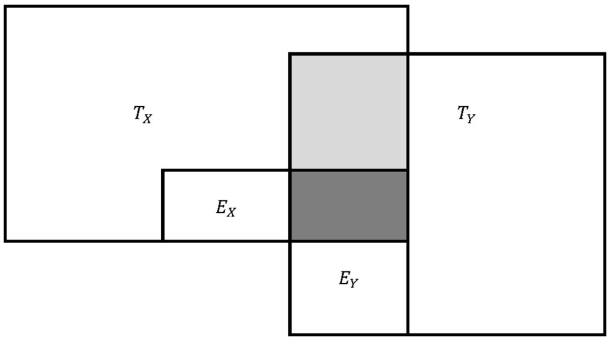


FIGURE 2 | Conceptual representation of observed score variance in the case of correlated error. T, true scores; *E*, error scores. Note: lighted shaded area denotes shared variance between T_X and T_Y. Dark shaded area denotes shared variance between E_X and E_Y.

data, the effect of error on observed score correlation is more complicated than Eqs 2 or 3 suggest. In fact, it is not uncommon for observed score correlations to be greater than the square root of the product of their reliabilities (e.g., see Dozois et al., 1998). In such cases, Spearman's correction formula (Eq. 2) will result in correlations greater than 1.00. While literature indicates that such correlations should be truncated to unity (e.g., Onwuegbuzie et al., 2004), a truncated correlation of 1.00 may be a less accurate estimate of the true score correlation than its observed score counterpart. As readers will see, observed score correlations may be *less* than or *greater* than their true score counterparts and therefore *less* or *more* accurate than correlations adjusted by Spearman's (1904) formula.

To understand why observed score correlations may not always be less than their true score counterparts, we present Charles's "correction for the full effect of measurement error" (Charles, 2005, p. 226). Although his formula cannot be used when "true scores and error scores are unknown," the formula clarifies the roles that reliability and error play in the formation of observed score correlation and identifies "the assumptions made in the derivation of the correction for attenuation" formula (Zimmerman, 2007, p. 923). Moreover, in the case when observed scores are available and true and error scores are hypothesized, the quantities in his formula can be given specific values and the full effect of measurement error on sample data can be observed.

Charles's (2005) formula extends Spearman's (1904) formula by taking into account correlations between error scores and between true scores and error scores that can occur in sample data:

$$r_{T_X T_Y} = \frac{r_{O_X O_Y}}{\sqrt{r_{XX} r_{YY}}} - \frac{r_{E_X E_Y}\sqrt{e_{XX}}\sqrt{e_{YY}}}{\sqrt{r_{XX} r_{YY}}} - \frac{r_{T_X E_Y}\sqrt{e_{YY}}}{\sqrt{r_{YY}}} - \frac{r_{T_Y E_X}\sqrt{e_{XX}}}{\sqrt{r_{XX}}} \quad (4)$$

Although not explicit, Charles's formula considers the correlations that exist between true scores and error scores of individual measures by defining error (e.g., e_{XX}, e_{YY}) as the ratio between error and observed score variance. Although error is traditionally represented as 1 – reliability (e.g., $1 - r_{XX}$), such representation is only appropriate for population data as the correlation between true scores and error scores for a given measure (e.g., $r_{T_X E_X}$) is assumed to be 0 in the population. Just as with $r_{E_X E_Y}$, $r_{T_X E_Y}$, $r_{T_Y E_X}$, correlations between true and error scores of individual measures ($r_{T_X E_X}$, $r_{T_Y E_Y}$) are not necessarily 0 in sample data. Positive correlations between true and error scores result in errors (e.g., e_{XX}) that are *less* than 1 – reliability (e.g., $1 - r_{xx}$), while negative correlations result in errors that are *greater* than 1 – reliability, as indicated in the following formula (Charles, 2005):

$$e_{XX} = \left(1 - r_{XX} - \frac{\mathrm{cov}_{T_X E_X}}{S^2_{O_X}}\right) \quad (5)$$

Through a series of simulation tests, Zimmerman (2007) demonstrated that an equivalent form of Eq. 4 accurately produces true score correlations for sample data and unlike Spearman's (1904) formula, always yields correlation coefficients between −1.00 and 1.00. From Eq. 4, one sees that Spearman's formula results in over corrected correlations when $r_{E_X E_Y}$, $r_{T_X E_Y}$, and $r_{T_Y E_X}$ are greater than 0, and under-corrected correlations when they are less than 0.

By taking Eq. 4 and solving for $r_{O_X O_Y}$, one also sees that the effect of unreliable data is more complicated than what is represented in Eq. 3:

$$r_{O_X O_Y} = r_{T_X T_Y}\sqrt{r_{XX} r_{YY}} + r_{E_X E_Y}\sqrt{e_{XX}}\sqrt{e_{YY}} + r_{T_X E_Y}\sqrt{e_{YY}}\sqrt{r_{XX}} + r_{T_Y E_X}\sqrt{e_{XX}}\sqrt{r_{YY}} \quad (6)$$

Equation 6 demonstrates why observed score correlations can be greater than the square root of the product of reliabilities and that the full effect of unreliable data on observed score correlation extends beyond true score correlation and includes the correlation between error scores, and correlations between true scores and error scores.

To illustrate the effect of unreliable data on observed score correlation, consider the case where $r_{T_X T_Y} = 0.50$, $r_{XX} = r_{YY} = 0.80$, $r_{E_X E_Y} = 0.50$, and $r_{T_X E_Y} = r_{T_Y E_X} = 0.10$. For the sake of parsimony, we assume $r_{T_X E_X} = r_{T_Y E_Y} = 0$ and therefore that $e_{XX} = 1 - r_{XX}$ and $e_{YY} = 1 - r_{YY}$. Based on Eq. 3, one would expect that the observed score correlation to be 0.40 ($0.50\sqrt{0.80 \times 0.80}$). However, as can be seen via the boxed points in **Figure 3**, the effect of correlated error, the correlation between T_X and E_Y, and the correlation between T_Y and E_X respectively increase the expected observed score correlation by 0.10, 0.04, and 0.04 resulting in an observed score correlation of 0.58, which is *greater* than the true score correlation of 0.50, and closer to the true score correlation of 0.50 than the Spearman (1904) correction resulting from Eq. 2 which equals 0.725 (i.e., $0.58/\sqrt{0.80 \times 0.80}$). This example shows that the attenuating effect of unreliable data (first term inEq. 6) is mitigated by the effect of correlated error (second term in Eq. 6) and the effects of correlations between true and error scores (third and forth terms in Eq. 6), assuming that the correlations are in the positive direction. Correlations in the negative direction serve to further attenuate the true score correlation beyond the first term in Eq. 6. This example further shows that observed score correlations are not always attenuated by measurement error and that in some cases an observed score correlation may provide an estimate that is closer to the true score correlation than a correlation that has been corrected by Spearman's formula.

As illustrated in **Figure 3**, the effect of correlated error and correlations between true and error scores tend to *increase* as reliability *decreases* and the magnitudes of $r_{T_X E_Y}$, $r_{T_Y E_X}$, and $r_{E_X E_Y}$ *increase*. The question, of course, is how big are these so-called "nuisance correlations" in real sample data? One can expect that, on average, repeated samples of scores would yield correlations of 0 for $r_{T_X E_Y}$ and $r_{T_Y E_X}$, as these correlations are assumed to be 0 in the population (Zimmerman, 2007). However, correlations between errors scores are not necessarily 0 in the population. Correlation between error scores can arise, for example, whenever tests are administered on the same occasion, consider the same construct, or are based on the same set of items (Zimmerman and Williams, 1977). In such cases, one can expect that, on average, repeated samples of error scores would approximate the level of correlated error in the population (Zimmerman, 2007). One can also expect that the variability of these correlations would increase

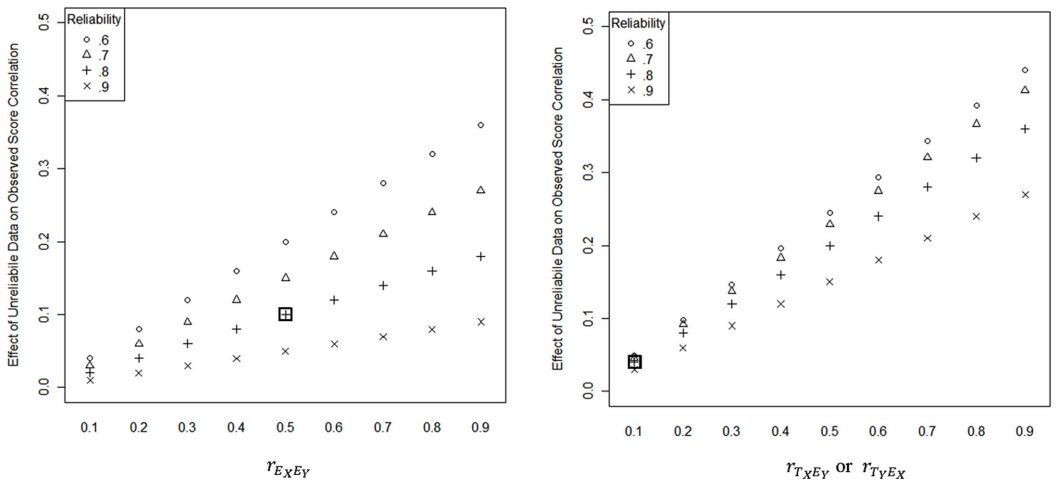


FIGURE 3 | Effect of unreliable data on observed score correlation as a function of reliability, correlated error ($r_{E_X E_Y}$), and correlations between true and error scores ($r_{T_X E_Y}$, $r_{T_Y E_X}$). Note: $r_{T_X E_X} = r_{T_Y E_Y} = 0$; $e_{XX} = e_{YY} = 1 -$ reliability. The boxed point on the left-hand panel indicates the effect of correlated error on an observed score correlation when $r_{T_X T_Y} = 0.50$, $r_{XX} = r_{YY} = 0.80$, and $r_{E_X E_Y} = 0.50$. The boxed point on the right-hand panel indicates the effect of a correlation between true and error scores (e.g., $r_{T_Y E_X}$) on an observed score correlation when $r_{T_X T_Y} = 0.50$, $r_{XX} = r_{YY} = 0.80$, and $r_{T_X E_Y} = r_{T_Y E_X} = 0.10$.

Table 1 | SAT data observed, error, and true scores.

	Writing scores			Reading scores		
State	**Observed (O_Y)**	**True (T_Y)**	**Error (E_Y)**	**Observed (O_Y)**	**True (T_Y)**	**Error (E_Y)**
Connecticut	513.0	512.0	1.0	509.0	509.8	−0.8
Delaware	476.0	485.0	−9.0	489.0	495.8	−6.8
Georgia	473.0	481.2	−8.2	485.0	491.4	−6.4
Maryland	491.0	496.4	−5.4	499.0	500.6	−1.6
Massachusetts	509.0	510.6	−1.6	513.0	513.2	−0.2
New Hampshire	511.0	510.4	0.6	523.0	521.0	2.0
New Jersey	497.0	495.8	1.2	495.0	495.4	−0.4
New York	476.0	480.4	−4.4	485.0	488.2	−3.2
North Carolina	474.0	481.2	−7.2	493.0	495.6	−2.6
Pennsylvania	479.0	482.2	−3.2	493.0	493.0	0.0
Rhode Island	489.0	491.4	−2.4	495.0	495.6	−0.6
South Carolina	464.0	473.8	−9.8	482.0	486.6	−4.6
Virginia	495.0	498.4	−3.4	512.0	511.4	0.6
M	488.2	492.2	−4.0	497.9	499.9	−1.9
SD	16.1	12.9	3.8	12.6	10.6	2.7

as the sample size (n) decreases. Indeed, Zimmerman (2007) found that the distributions for $r_{T_X E_Y}$, $r_{T_Y E_X}$, and $r_{E_X E_Y}$ yielded SDs of $\sim 1/\sqrt{n-1}$. The fact that there is sampling variance in these values makes "dealing with measurement error and sampling error pragmatically inseparable" (Charles, 2005, p. 226).

To empirically illustrate the full effect of unreliable data on observed score correlation, we build on the work of Wetcher-Hendricks (2006) and apply Eq. 6 to education and psychology examples. For education data, we analyzed SAT writing and reading scores (Benefield, 2011; College Board, 2011; Public Agenda, 2011). For psychology data, we analyzed scores from the Beck Depression Inventory-II (BDI-II; Beck, 1996) and Beck Anxiety Inventory (BAI; Beck, 1990). We close this section by contrasting CTT assumptions relating to reliability to population and sample data and summarizing how differences in those assumptions impact the full effect of reliability on observed score correlations.

EDUCATION EXAMPLE

We applied Eq. 6 to average SAT scores from 13 states associated with the original USA colonies (see **Table 1**). We selected these states as they were a cohesive group and were among the 17 states with the highest participation rates (Benefield, 2011; College

Board, 2011; Public Agenda, 2011). We used data reported for 2011 as observed data and the long-run average of SAT scores reported since the new form of the SAT was introduced as true scores, given the psychometric principle from Allen and Yen (1979) that long-run averages equal true score values (cf. Wetcher-Hendricks, 2006). To compute error scores, we subtracted true scores from observed scores. The components of Eq. 6 applied to the SAT data are presented in **Table 2** and yield the observed score correlation (0.90), as follows:

$$\begin{aligned} r_{O_X O_Y} &= 0.91\sqrt{0.71 \times 0.65} + 0.84\sqrt{0.05}\sqrt{0.06} \\ &\quad + 0.62\sqrt{0.06}\sqrt{0.71} + 0.68\sqrt{0.05}\sqrt{0.65} \\ 0.90 &= 0.62 + 0.04 + 0.12 + 0.12 \end{aligned} \tag{7}$$

While the reliability of the SAT data served to attenuate the true score correlation between reading and writing scores (cf. first term in Eq. 7), the correlations between (a) reading error scores and writing errors scores (cf. second term in Eq. 7), (b) reading error scores and writing true scores (cf. third term in Eq. 7), and (c) writing errors scores and reading true scores (cf. forth term in Eq. 7), served to mitigate the effect of that attenuation. Also note that the observed score correlation (0.90) is more in-line with the true score correlation (0.91) than what Spearman's (1904) correction formula yielded (i.e., $0.90/\sqrt{0.71 \times 0.65} = 1.33$). Given that Spearman's correction produced a value in excess of 1.00, it would be more accurate to report the observed score correlation, rather than follow conventional guidelines (e.g., Onwuegbuzie et al., 2004) and report 1.00.

Table 2 | Values for observed score correlation computation for SAT and beck data.

Component	SAT	Beck
$r(T_X, T_Y)$	0.91	0.69
$r(E_X, E_Y)$	0.84	0.76
$r(T_X, E_Y)$	0.62	0.47
$r(T_Y, E_X)$	0.68	−0.14
$r_{XX}(SD^2_{T_X}/SD^2_{O_X})$	0.71	0.79
$r_{YY}(SD^2_{T_Y}/SD^2_{O_Y})$	0.65	0.83
$e_{XX}(SD^2_{T_X}/SD^2_{O_X})$	0.05	0.21
$e_{YY}(SD^2_{T_Y}/SD^2_{O_Y})$	0.06	0.17

PSYCHOLOGY EXAMPLE

We applied Eq. 6 to average class BDI-II (Beck, 1996) and BAI (Beck, 1990) scores from Nimon and Henson, 2010; see **Table 3**). In their study, students responded to the BDI-II and BAI at two times within the same semester (i.e., time-1, time-2). Following Wetcher-Hendricks' (2006) example of using predicted scores as true scores, we used scores for time-1 as observed data and the predicted scores (regressing time-2 on time-1) as true scores. As in the education example, we subtracted true scores from observed scores to compute error scores. The components of Eq. 6 applied to Nimon and Henson's (2010) data are presented in **Table 2** and yield the observed score correlation of 0.81, as follows:

$$\begin{aligned} r_{O_X O_Y} &= 0.69\sqrt{0.79 \times 0.83} + 0.76\sqrt{0.21}\sqrt{0.17} \\ &\quad + 0.47\sqrt{0.17}\sqrt{0.79} - 0.14\sqrt{0.21}\sqrt{0.83} \\ 0.81 &= 0.56 + 0.14 + 0.17 - 0.06 \end{aligned} \tag{8}$$

In Nimon and Henson's (2010) data, the true score correlation (0.69) is *lower* than the observed score correlation (0.81). In this case, the attenuating effect of unreliability in scores was mitigated by other relationships involving error scores which, in the end, served to *increase* the observed correlation rather than *attenuate* it. As in the SAT data, Spearman's (1904) correction ($0.81/\sqrt{0.79 \times 0.83} = 1.00$) produced an over corrected correlation coefficient. The over-correction resulting from Spearman's correction was largely due to the formula not taking into account the correlation between the error scores and the correlation between the true anxiety score and the error depression score.

SUMMARY

Classical test theory can be used to prove that $\rho_{T_X E_X}$, $\rho_{T_Y E_Y}$, $\rho_{T_X E_Y}$, and $\rho_{T_Y E_X}$ all equal to 0 in a given population (Zimmerman, 2007). However, the tenets of CTT do not provide proof that $\rho_{E_X E_Y} = 0$. Furthermore, in the case of sample data, $r_{T_X E_X}$, $r_{T_Y E_Y}$, $r_{T_X E_Y}$, $r_{T_Y E_X}$, and $r_{E_X E_Y}$ are not necessarily zero.

Table 3 | Beck data observed, error, and true scores.

Class	Depression scores (BDI-II)			Anxiety scores (BAI)		
	Observed (O_Y)	True (T_Y)	Error (E_Y)	Observed (O_Y)	True (T_Y)	Error (E_Y)
1	6.67	7.76	−1.10	7.34	9.08	−1.74
2	10.26	10.45	−0.19	11.04	11.26	−0.22
3	5.92	4.75	1.17	9.69	8.69	1.00
4	7.21	7.40	−0.19	7.00	6.98	0.02
5	6.85	7.28	−0.43	7.31	6.40	0.91
6	6.78	6.67	0.12	9.27	8.84	0.43
7	6.13	6.82	−0.70	6.60	7.70	−1.09
8	11.54	10.23	1.32	12.62	11.93	0.69
M	7.67	7.67	0.00	8.86	8.86	0.00
SD	2.06	1.88	0.85	2.17	1.94	0.98

Because $r_{T_X E_X}, r_{T_Y E_Y}, r_{T_X E_Y}, r_{T_Y E_X}$, and $r_{E_X E_Y}$ may not be zero in any given sample, researchers cannot assume that poor reliability will *always* result in lower observed score correlations. As we have demonstrated, observed score correlations may be less than or greater than their true score counterparts and therefore less or more accurate than correlations adjusted by Spearman's (1904) formula.

Just as reliability affects the magnitude of observed score correlations, it follows that statistical significance tests are also impacted by measurement error. While error that causes observed score correlations to be greater than their true score counterparts increases the power of statistical significance tests, error that causes observed score correlations to be less than their true score counterparts decreases the power of statistical significance tests, with all else being constant. Consider the data from Nimon and Henson (2010) as an example. As computed by G*Power 3 (Faul et al., 2007), with all other parameters held constant, the power of the observed score correlation ($r_{O_X O_Y} = 0.81$, $1 - \beta = 0.90$) is greater than the power of true score correlation ($r_{T_X T_Y} = 0.69$, $1 - \beta = 0.62$). In this case, error in the data served to *decrease* the Type II error rate rather than *increase* it.

As we leave this section, it is important to note that the effect of reliability on observed score correlation *decreases* as reliability and sample size *increase*. Consider two research settings reviewed in Zimmerman (2007): In large *n* studies involving standardized tests, "many educational and psychological tests have generally accepted reliabilities of 0.90 or 0.95, and studies with 500 or 1,000 or more participants are not uncommon" (p. 937). In this research setting, the correction for and the effect of reliability on observed score correlation may be accurately represented by Eqs 2 and 3, respectively, as long as there is not substantial correlated error in the population. However, in studies involving a small number of participants and new instrumentation, reliability may be 0.70, 0.60, or lower. In this research setting, Eq. 2 may not accurately correct and Eq. 3 may not accurately represent the effect of measurement error on an observed score correlation. In general, if the correlation resulting from Eq. 2 is much greater than the observed score correlation, it is probably inaccurate as it does not consider the full effect of measurement error and error score correlations on the observed score correlation (cf. Eq. 4, Zimmerman, 2007).

RELIABILITY: HOW DO WE ASSESS?

Given that reliability affects the magnitude and statistical significance of sample statistics, it is important for researchers to assess the reliability of their data. The technique to assess reliability depends on the type of measurement error being considered. Under CTT, typical types of reliability assessed in educational and psychological research are test–retest, parallel-form, inter-rater, and internal consistency. After we present the aforementioned techniques to assess reliability, we conclude this section by countering a common myth regarding their collective nature.

TEST–RETEST

Reliability estimates that consider the consistency of scores across time are referred to as test–retest reliability estimates. Test–retest reliability is assessed by having a set of individuals take the same assessment at different points in time (e.g., week 1, week 2) and correlating the results between the two measurement occasions. For well-developed standardized achievement tests administered reasonably close together, test–retest reliability estimates tend to range between 0.70 and 0.90 (Popham, 2000).

PARALLEL-FORM

Reliability estimates that consider the consistency of scores across multiple forms are referred to as parallel-form reliability estimates. Parallel-form reliability is assessed by having a set of individuals take different forms of an instrument (e.g., short and long; Form A and Form B) and correlating the results. For well-developed standardized achievement tests, parallel-form reliability estimates tend to hover between 0.80 and 0.90 (Popham, 2000).

INTER-RATER

Reliability estimates that consider the consistency of scores across raters are referred to as inter-rater reliability estimates. Inter-rater reliability is assessed by having two (or more) raters assess the same set of individuals (or information) and analyzing the results. Inter-rater reliability may be found by computing consensus estimates, consistency estimates, or measurement estimates (Stemler, 2004):

Consensus

Consensus estimates of inter-rater reliability are based on the assumption that there should be exact agreement between raters. The most popular consensus estimate is simple percent-agreement, which is calculated by dividing the number of cases that received the same rating by the number of cases rated. In general, consensus estimates should be 70% or greater (Stemler, 2004). Cohen's kappa (κ; Cohen, 1960) is a derivation of simple percent-agreement, which attempts to correct for the amount of agreement that could be expected by chance:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{9}$$

where p_o is the observed agreement among raters and p_c is the hypothetical probability of chance agreement. Kappa values between 0.40 and 0.75 are considered moderate, and values between 0.75 and 1.00 are considered excellent (Fleiss, 1981).

Consistency

Consistency estimates of inter-rater reliability are based on the assumption that it is unnecessary for raters to yield the same responses as long as their responses are relatively consistent. Inter-rater reliability is typically assessed by correlating rater responses, where correlation coefficients of 0.70 or above are generally considered acceptable (Barrett, 2001).

Measurement

Measurement estimates of inter-rater reliability are based on the assumption that all rater information (including discrepant ratings) should be used in creating a scale score. Principal components analysis is a popular technique to compute the measurement estimate of inter-rater reliability (Harman, 1967). If the amount of shared variance in ratings that is accounted for by the first principal component is greater than 60%, it is assumed that raters are assessing a common construct (Stemler, 2004).

INTERNAL CONSISTENCY

Reliability estimates that consider item homogeneity, or the degree to which items on a test are internally consistent, are referred to as internal consistency reliability estimates. Measures of internal consistency are the most commonly reported form of reliability coefficient because they are readily available from a single administration of a test (Hogan et al., 2000; Henson, 2001). Internal consistency reliability is typically assessed by computing coefficient alpha (α; Cronbach, 1951):

$$\alpha = \frac{k}{(k-1)}\left[1 - \left(\frac{\sum \mathrm{SD}_i^2}{\mathrm{SD}_{\mathrm{Total}}^2}\right)\right] \quad (10)$$

where k refers to the number of items on the assessment device, i refers to item, and *Total* refers to the total scale score.

Note that the first part of the formula $[k/(k-1)]$ attempts to "correct" for potential bias in scales that have a small number of items. The rationale is that the more items in a scale, the less likely items will be biased. As k increases, the correction for bias becomes smaller. For two items, the correction is $2\,[2/(2-1)]$. For 10 items, the correction is 1.1, and for 100 items, the correction is only 1.01.

Due to the impact that internal consistency has on the interpretation of scale scores and variable relationships, researchers typically relate estimates of internal consistency to established benchmarks. Henson (2001) reviewed such benchmarks and cited 0.90 as a minimum internal consistency estimate for standardized test scores used for important educational decisions and 0.80 for scores used for general research purposes. Nunnally (1967) suggested minimum reliabilities of 0.60 or 0.50 for early stages of research, but this was increased to an exploratory standard of 0.70 in his second edition (1978, see also Nunnally and Bernstein, 1994). This change may have resulted in "many researchers citing Nunnally (1978) if they attained this loftier standard and citing the first edition if they did not!" (Henson, 2001, p. 181). In general, internal consistency estimates should be strong for most research purposes, although the exact magnitude of an acceptable coefficient alpha would depend on the purposes of the research.

For example, it is conceivable that coefficient alpha can be too high, which would occur when the items of measurement are highly redundant and measuring the same aspect of a construct. At the extreme of this case, all items would be perfectly correlated and thus alpha would be a perfect 1.00 (see Henson, 2001, for a demonstration). This would reflect poor measurement because of redundancy and, possibly, failure to reflect an appropriate breadth of items from the range of all possible items that could be used to measure the construct (cf. Hulin et al., 2001). Furthermore, a high coefficient alpha is sometimes misinterpreted as an indicator of unidimensionality. This is not the case, and in his summary thoughts on the history of his formula, Cronbach (2004) noted he had "cleared the air by getting rid of the assumption that the items of a test were unidimensional" (p. 397). It is certainly possible to find substantial alpha coefficients even when there are multiple (sometimes subtle) constructs represented in the data.

Conversely, low alphas may indeed reflect a failure to recognize multiple dimensions within a data set, particularly when those dimensions or factors are weakly correlated. In such cases, researchers should first explore the factor structure of their data prior to computation of alpha, and alpha generally should be computed at the subscale (e.g., factor) level rather than on a global test level when there are multiple constructs being assessed. The bottom line is that the interpretation of coefficient alpha when assessing constructs should consider (a) item representativeness and breadth and (b) desired overlap between items.

MULTIPLE SOURCES OF MEASUREMENT ERROR

It is important to note that the sources of measurement error described in this section are separate and cumulative (cf. Anastasi and Urbina, 1997). As noted by Henson (2001),

> Too many researchers believe that if they obtain $\alpha = 0.90$ for their scores, then the same 10% error would be found in a test–retest or inter-rater coefficient. Instead, assuming 10% error for internal consistency, stability, and inter-rater, then the overall measurement error would be 30%, not 10% because these estimates explain different sources of error (p. 182).

The point to be made here is that measurement error can originate from a variety of sources, which can lead to more cumulative measurement error than the researcher might suspect. Of course, this can impact observed relationships, effect sizes, and statistical power.

In order to get a better understanding of the sources of measurement error in scores, generalizability theory (G theory) can be employed which allows researchers to "(a) consider simultaneously multiple sources of measurement error, (b) consider measurement error interaction effects, and (c) estimate reliability coefficients for both "relative" and "absolute" decisions" (Thompson, 2003b, p. 43). As a full discussion of G theory is beyond the scope of this article, readers are directed to Shavelson and Webb (1991) for an accessible treatment. We continue with a discussion of how published reliability estimates can be used to inform research design and RG studies.

HOW DO WE PLAN? THE ROLE OF RG

As defined by Vacha-Haase (1998), RG is a method that helps characterize the reliability estimates for multiple administrations of a given instrument. Vacha-Haase further described RG as an extension of validity generalization (Schmidt and Hunter, 1977; Hunter and Schmidt, 1990) and stated that RG "characterizes (a) the typical reliability of scores for a given test across studies, (b) the amount of variability in reliability coefficients for given measures, and (c) the sources of variability in reliability coefficients across studies" (p. 6). RG assesses the variability in reliability estimates and helps identify how sample characteristics and sampling design impacts reliability estimates.

In a meta-analysis of 47 RG studies, Vacha-Haase and Thompson (2011) found that the average of the coefficient alpha means from these studies was 0.80 (SD = 0.09) with a range from 0.45 to 0.95. These results illustrate the extent to which reliability estimates can vary across studies, and in this case, across instruments. Because any given RG study quantifies the variation of reliability across studies for a given instrument, the results empirically demonstrate that the phrases "the reliability of the test" and "the test is not reliable" are inappropriate and that reliability is

a property inured to data, not instrumentation (Thompson and Vacha-Haase, 2000, p. 175).

Results from RG studies also provide empirical evidence that reliability estimates can vary according to sample characteristics. In their meta-analysis, Vacha-Haase and Thompson (2011) found that "the most commonly used predictor variables included gender (83.3% of the 47 RG studies), sample size (68.8%), age in years (54.2%), and ethnicity (52.1%)" (p. 162). Upon evaluating predictor variables across studies, they found number of items and the sample SD of scale scores to be noteworthy, as well as age and gender. However, as is true with all analyses Vacha-Haase and Thompson's review was contingent on the independent variables included in the models, as variable omission can impact results (cf. Pedhazur, 1997).

USING RG TO PLAN

While RG studies demonstrate the importance of assessing reliability estimates for the data in hand, they can also help researchers make educated decisions about the design of future studies. Researchers typically devote considerable energies toward study design because a poorly designed study is likely to produce results that are not useful or do not provide reliable answers to research questions. When data need to be collected from study participants, researchers must determine the most suitable instrument and should consult existing literature to understand the relationship between the reliability of the data to be measured and the population of interest. When available, an RG study can help guide researchers in the instrument selection process. By consulting reliability estimates from published reports, researchers can form hypotheses about whether or not reliability estimates will be acceptable given their sample and testing conditions.

To illustrate how RG studies can improve research design, we provide a hypothetical example. Presume that we want to conduct a study on a sample of fifth-grade students and plan to administer the Self-Description Questionnaire (SDQ; cf. Marsh, 1989). Because we want to conduct a study that will produce useful results, we endeavor to predict if scores from our sample are likely to produce acceptable levels of reliability estimates. Also, presume we are considering modifications such as shortening the 64-item instrument (because of limitations in the available time for administration) and changing the original five-point Likert type scale to a six-point Likert type scale (because of concern about response tendency with a middle option).

Results from Leach et al.'s (2006) RG study of the SDQ may help us decide if the SDQ might be an appropriate instrument to administer to our sample and if our proposed modifications might result in acceptable reliability estimates. For each domain of the SDQ, Leach et al. found that the reliability estimates tended to be within an acceptable range with general self-concept (GSC) scores yielding lower reliability estimates. However, even for GSC scores, the majority of the reliability estimates were within the acceptable range. Furthermore, Leach et al. found that "the most pervasive (predictor of reliability variation) seemed to be the role of the five-point Likert scale and use of the original version (unmodified) of the SDQ I" (p. 300).

The RG study suggests that SDQ I scores for our hypothetical example would likely yield acceptable levels of reliability presuming we did not modify the original instrument by shortening the instrument or changing the five-point Likert scale, and also assuming we employ a sample that is consistent with that for which the instrument was developed. These decisions help us with study design and mitigate our risk of producing results that might not be useful or yield biased effect sizes.

As illustrated, prior to administering an instrument, researchers should consult the existing literature to determine if an RG study has been conducted. RG studies have been published on a variety of measures and in a variety of journals. Researchers might first want to consult Vacha-Haase and Thompson (2011), as the authors provided references to 47 RG studies, including reports from Educational and Psychological Measurement, Journal of Nursing Measurement, Journal of Personality Assessment, Personality and Individual Differences, Personal Relationships, Journal of Marriage and Family, Assessment, Psychological Methods, Journal of Cross-Cultural Psychology, Journal of Managerial Issues, and International Journal of Clinical and Health Psychology. The fundamental message is that RG studies are published, and continue to be published, on a variety of measures in a variety of journals including journals focusing on measurement issues and substantive analyses.

BARRIERS TO CONDUCTING RG STUDIES

Researchers need to be cognizant of the barriers that impact RG results, as these barriers limit the generalization of results. Insufficient reporting of reliability estimates and sample characteristics are primary difficulties that impact the quality of RG results. When details about measurement and sampling designs are not provided, model misspecifications in RG studies may occur (Vacha-Haase and Thompson, 2011). As Dimitrov (2002) noted, misspecifications may "occur when relevant characteristics of the study samples are not coded as independent variables in RG analysis" (p. 794). When sampling variance is not included, the ability to conduct extensions to Vacha-Haase's (1998) RG method may also be impeded. For example, Rodriguez and Maeda (2006) noted that some RG studies make direct adjustments of alpha coefficients. However they noted problems with adjusting some but not all alphas in RG studies when researchers fail to publish sample variances.

Reliability estimates

Meticulous RG researchers have been discouraged to find that many of the studies they consult either (a) only report the reliabilities from previous studies (i.e., induct reliability coefficients) and not report reliabilities from their sample at hand or (b) do not report reliabilities at all (cf. Vacha-Haase et al., 2002). Vacha-Haase and Thompson (2011) found that "in an astounding 54.6% of the 12,994 primary reports authors did not even mention reliability!" and that "in 15.7% of the 12,994 primary reports, authors did mention score reliability but merely inducted previously reported values as if they applied to their data" (p. 161).

The *file drawer problem* of researchers not publishing results that were not statistically significant might be another factor that limits RG results. As discussed above, when reliability estimates are low, the ability to obtain noteworthy effect sizes can

be impacted. Rosenthal (1979) noted that the *file drawer problem* might, in the extreme case, result in journals that "are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g., $p > 0.05$) results" (p. 638). As noted by Rosenthal (1995), when conducting meta-analyses, a solution to the file drawer problem does not exist, "but reasonable boundaries can be established on the problem, and the degree of damage to any research conclusion that could be done by the file drawer problem can be estimated" (Rosenthal, 1995, p. 189). Because RG studies are meta-analyses of reliability estimates, RG studies are not immune to a biased sample of statistically significant studies and reliability estimates that never make it out of the file drawer might be lower than the estimates published in refereed journal publications.

Sample characteristics

Insufficient reporting of sample variance, sample characteristics, and sample design is another barrier impacting RG results. Insufficient reporting practices have been documented by several researchers. Miller et al. (2007) noted "given the archival nature of the analysis, however, selection of predictor variables was also limited to those that were reported in the reviewed articles" (p. 1057). Shields and Caruso (2004) noted limitations on coding variables and stated that "practical considerations and insufficient reporting practices in the literature restrict the number and type of predictor variables that can be coded" (p. 259). Within teacher education research, for example, Zientek et al. (2008) found that only 9% of the articles "included all of the elements necessary to possibly conduct a replication study" (p. 210) and that many studies fail to report both means and SD.

REPORTING RECOMMENDATIONS

In order to improve RG studies and remove barriers encountered with insufficient reporting practices, we propose a list of relevant information in **Figure 4** to be included in journal articles and research publications. Reporting this information will facilitate RG researchers' ability to conduct meaningful RG studies. Many of these items are necessary for study replication; hence they adhere to recommendations from the American Educational Research Association (AERA, 2006) and the American Psychological Association (APA, 2009b). We want to emphasize the importance of providing (a) the means and SD for each subscale, (b) the number of items for each subscale, and (c) the technique used to compute scale scores (e.g., sum, average).

To illustrate how these can be presented succinctly within a journal article format, we present a sample write-up in the Appendix. This narrative can serve as a guide for journal publications and research reports and follows the American Psychological Association (APA, 2009a) guidelines to help reduce bias when reporting sample characteristics. According to APA (2009a),

> Human samples should be fully described with respect to gender, age, and, when relevant to the study, race or ethnicity. Where appropriate, additional information should be presented (generation, linguistic background, socioeconomic status, national origin, sexual orientation, special interest group membership, etc.). (p. 4)

In addition to improving the ability to conduct RG studies, providing this information will allow future researchers to replicate studies and compare future findings with previous findings. To address journal space limitations and ease in readability, group means, SD, and reliability estimates may be disaggregated within

Score Characteristics Reliability estimate Type of reliability (e.g., alpha, test-retest) Mean Standard Deviation (*SD*) Scale Score (e.g., summated, average)	*Organizational Characteristics* Country or Geographic Region Organizational Type Organizational Size Other Organizational Characteristics
Scale Characteristics Language administered Format (e.g., paper, online) Reference number for scale (e.g., 0, 1) Points in scale (e.g., 5, 7) # of items Referent (e.g., self, other) Deviations (e.g., wording changes, items deleted)	*Sample Characteristics* Selection (e.g., random, purposeful) Sample Size Response Rate Age (Mean, SD) Gender (e.g., *n* Male) Ethnicity (*n* Latino/Hispanic/ *n* Not Latino/Hispanic) Race[a] (e.g., White, Black, Hispanic, Asian) Marital Status (e.g., *n* Married) Educational Level Participant Type Other Sample Characteristics
Educational Related Characteristics Education Level (e.g., *n* EC-16) Participant Type (e.g., *n* Regular, *n* Gifted, *n* Special) School Type (e.g., *n* Private, *n* Public, *n* Charter) Residential Area (e.g., *n* Suburban, *n* Rural, *n* Urban) College Major (e.g., *n* Education, *n* Psychology) Socio-economic Status (e.g., *n* Free/Reduced Lunch)	*Psychology Related Characteristics* Educational Level (e.g., *n* HS, *n* College) Participant Type (e.g., *n* Worker, *n* Manager) Org Type (e.g., *n* Private, *n* Public, *n* Non-profit) Firm Size (e.g., *n* Small, *n* Medium, *n* Large) Organizational/Job Tenure (Mean, SD) Industry (e.g., *n* Manufacturing, *n* Sales)

FIGURE 4 | Recommended data to report for each set of scores subjected to an inferential test. Note: Data should be reported for each set of scores analyzed across all measurement occasions (e.g., pre-test, post-test) and groups (e.g., gender, management level). [a]The Appendix adheres to the APA (2009a) recommendations for reporting race. Reporting of sample characteristics by race should follow APA (2009a) guidelines.

a table, as illustrated in the Appendix. Readers can also consult Pajares and Graham (1999) as a guide for presenting data.

WHAT DO WE DO IN THE PRESENCE OF UNRELIABLE DATA?

Despite the best-laid plans and research designs, researchers will at times still find data with poor reliability. In the real-world problem of conducting analyses on unreliable data, researchers are faced with many options which may include: (a) omitting variables from analyses, (b) deleting items from scale scores, (c) conducting "what if" reliability analyses, and (d) correcting effect sizes for reliability.

OMITTING VARIABLES FROM ANALYSES

Yetkiner and Thompson (2010) suggested that researchers omit variables (e.g., depression, anxiety) that exhibit poor reliability from their analyses. Alternatively, researchers may choose to conduct SEM analyses in the presence of poor reliability whereby latent variables are formed from item scores. The former become the units of analyses and yield statistics as if multiple-item scale scores had been measured without error. However, as noted by Yetkiner and Thompson, reliability is important even when SEM methods are used, as score reliability affects overall fit statistics.

DELETING ITEMS FROM SCALE SCORES

Rather than omitting an entire variable (e.g., depression, anxiety) from an analysis, a researcher may choose to omit one or more items (e.g., BDI-1, BAI-2) that are negatively impacting the reliability of the observed score. Dillon and Bearden (2001) suggested that researchers consider deleting items when scores from published instruments suffer from low reliability. Although "extensive revisions to prior scale dimensionality are questionable . . . one or a few items may well be deleted" in order to increase reliability (Dillon and Bearden, p. 69). Of course, the process of item deletion should be documented in the methods section of the article. In addition, we suggest that researchers report the reliability of the scale with and without the deleted items in order to add to the body of knowledge of the instrument and to facilitate the ability to conduct RG studies.

CONDUCTING "WHAT IF" RELIABILITY ANALYSES

Onwuegbuzie et al. (2004) proposed a "what if reliability" analysis for assessing the statistical significance of bivariate relationships. In their analysis, they suggested researchers use Spearman's (1904) correction formula and determine the "minimum sample size needed to obtain a statistically significant r based on observed reliability levels for x and y" (p. 236). They suggested, for example, that when $r_{O_X O_Y} = 0.30$, $r_{xx} = 0.80$, $r_{yy} = 0.80$, $r_{T_X T_Y}$, based on Spearman's formula, yields 0.38 ($0.30/\sqrt{(0.80 \times 0.80)}$) and "that this corrected correlation would be statistically significant with a sample size as small as 28" (p. 235).

Underlying the Onwuegbuzie et al. (2004) reliability analysis, presumably, is the assumption the error is uncorrelated in the population and sample. However, even in the case that such an assumption in tenable, the problem of "what if reliability" analysis is that the statistical significance of correlation coefficients that have been adjusted by Spearman's formula cannot be tested for statistical significance (Magnusson, 1967). As noted by Muchinsky (1996):

> Suppose an uncorrected validity coefficient of 0.29 is significantly different than zero at $p = 0.06$. Upon application of the correction for attenuation (Spearman's formula), the validity coefficient is elevated to 0.36. The inference cannot be drawn that the (corrected) validity coefficient is now significantly different from zero at $p < 0.05$ (p. 71).

As Spearman's formula does not fully account for the measurement error in an observed score correlation, correlations based on the formula have a different sampling distribution than correlations based on reliable data (Charles, 2005). Only in the case when the full effect of measurement error on a sample observed score correlation has been calculated (i.e., Eq. 4 or its equivalent) can inferences be drawn about the statistical significance of $r_{T_X T_Y}$.

CORRECTING EFFECT SIZES FOR RELIABILITY

In this article we presented empirical evidence that identified limitations associated with reporting correlations based on Spearman's (1904) correction. Based on our review of the theoretical and empirical literature concerning Spearman's correction, we offer researchers the following suggestions.

First, consider whether correlated errors exist in the population. If a research setting is consistent with correlated error (e.g., tests are administered on the same occasion, similar constructs, repeated measures), SEM analyses may be more appropriate to conduct where measurement error can be specifically modeled. However, as noted by Yetkiner and Thompson (2010), "score reliability estimates do affect our overall fit statistics, and so the quality of our measurement error estimates is important even in SEM" (p. 9).

Second, if Spearman's correction is greater than 1.00, do not truncate to unity. Rather consider the role that measurement and sampling error is playing in the corrected estimate. In some cases, the observed score correlation may be closer to the true score correlation than a corrected correlation that has been truncated to unity. Additionally, reporting the actual Spearman's correction provides more information than a value that has been truncated to unity.

Third, examine the difference between the observed score correlation and Spearman's correction. Several authors have suggested that a corrected correlation "very much higher than the original correlation" (i.e., 0.85 vs. 0.45) is "probably inaccurate" (Zimmerman, 2007, p. 938). A large difference between an observed correlation and corrected correlation "could be explained by correlated errors in the population, or alternatively because error are correlated with true scores or with each other in an anomalous sample" (Zimmerman, 2007, p. 938).

Fourth, if analyses based on Spearman's correction are reported, at a minimum also report results based on observed score correlations. Additionally, explicitly report the level of correlation error that is assumed to exist in the population.

CONCLUSION

In the present article, we sought to help researchers understand that (a) measurement error does not always attenuate observed score correlations in the presence of correlated errors, (b) different sources of measurement error are cumulative, and (c) reliability is a function of data, not instrumentation. We demonstrated that reliability impacts the magnitude and statistical significance tests

that consider variable relationships and identified techniques that applied researchers can use to fully understand the impact of measurement error on their data. We synthesized RG literature and proposed a reporting methodology that can improve the quality of future RG studies as well as substantive studies that they may inform.

In a perfect world, data would be perfectly reliable and researchers would not have worry to what degree their analyses were subject to *nuisance correlations* that exist in sample data. However, in the real-world, measurement error exists, can be systematic, and is unavoidably coupled with sampling error. As such, researchers must be aware of the full impact that measurement error has on their results and do all that they can *a priori* to select instruments that are likely to yield appropriate levels of reliability for their given sample. If considering these factors, we can better inform our collective research and move the fields of education and psychology forward by meticulously reporting the effects of reliability in our data.

REFERENCES

Allen, M. J., and Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educ. Res.* 35, 33–40.

American Psychological Association. (2009a). *Publication Manual of the American Psychological Association: Supplemental Material: Writing Clearly and Concisely*, Chap. 3. Available at: http://supp.apa.org/style/pubman-ch03.00.pdf

American Psychological Association. (2009b). *Publication Manual of the American Psychological Association*, 6th Edn. Washington, DC: American Psychological Association.

Anastasi, A., and Urbina, S. (1997). *Psychological Testing*, 7th Edn. Upper Saddle River, NJ: Prentice Hall.

Bandura, A. (2006). "Guide for constructing self-efficacy scales," in *Self-Efficacy Beliefs of Adolescents*, Vol. 5, eds F. Pajares and T. Urdan (Greenwich, CT: Information Age Publishing), 307–337.

Barrett, P. (2001). *Assessing the Reliability of Rating Data*. Available at: http://www.pbarrett.net/presentations/rater.pdf

Beck, A. T. (1990). *Beck Anxiety Inventory (BAI)*. San Antonio, TX: The Psychological Corporation.

Beck, A. T. (1996). *Beck Depression Inventory-II (BDI-II)*. San Antonio, TX: The Psychological Corporation.

Benefield, D. (2011). *SAT Scores by State, 2011*. Harrisburg, PA: Common Wealth Foundation. Available at: http://www.commonwealthfoundation.org/policyblog/detail/sat-scores-by-state-2011

Charles, E. P. (2005). The correction for attenuation due to measurement error: clarifying concepts, and creating confidence sets. *Psychol. Methods* 10, 206–226.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46.

College Board. (2011). *Archived SAT Data and Reports*. Available at: http://professionals.collegeboard.com/data-reports-research/sat/archived

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadsworth.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educ. Psychol. Meas.* 64, 391–418. (editorial assistance by Shavelson, R. J.).

Dillon, W. R., and Bearden, W. (2001). Writing the survey question to capture the concept. *J. Consum. Psychol.* 10, 67–69.

Dimitrov, D. M. (2002). Reliability: arguments for multiple perspectives and potential problems with generalizability across studies. *Educ. Psychol. Meas.* 62, 783–801.

Dozois, D. J. A., Dobson, K. S., and Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory –II. *Psychol. Assess.* 10, 83–89.

Faul, F., Erdfelder, E., Lang, A., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Edn. New York: John Wiley.

Harman, H. H. (1967). *Modern Factor Analysis*. Chicago: University of Chicago Press.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: a conceptual primer on coefficient alpha. *Meas. Eval. Couns. Dev.* 34, 177–189.

Hogan, T. P., Benjamin, A., and Brezinski, K. L. (2000). Reliability methods: a note on the frequency of use of various types. *Educ. Psychol. Meas.* 60, 523–531.

Hulin, C., Netemeyer, R., and Cudeck, R. (2001). Can a reliability coefficient be too high? *J. Consum. Psychol.* 10, 55–58.

Hunter, J. E., and Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage.

Kieffer, K. M., Reese, R. J., and Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: a methodological review. *J. Exp. Educ.* 69, 280–309.

Leach, L. F., Henson, R. K., Odom, L. R., and Cagle, L. S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educ. Psychol. Meas.* 66, 285–304.

Lorenzo-Seva, U., Ferrando, P. J., and Chico, E. (2010). Two SPSS programs for interpreting multiple regression results. *Behav. Res. Methods* 42, 29–35.

Magnusson, D. (1967). *Test Theory*. Reading, MA: Addison-Wesley.

Marat, D. (2005). Assessing mathematics self-efficacy of diverse students form secondary schools in Auckland: implications for academic achievement. *Issues Educ. Res.* 15, 37–68.

Marsh, H. W. (1989). Age and sex effects in multiple dimensions of self-concept: preadolescence to early adulthood. *J. Educ. Psychol.* 81, 417–430.

Miller, C. S., Shields, A. L., Campfield, D., Wallace, K. A., and Weiss, R. D. (2007). Substance use scales of the Minnesota Multiphasic Personality Inventory: an exploration of score reliability via meta-analysis. *Educ. Psychol. Meas.* 67, 1052–1065.

Muchinsky, P. M. (1996). The correction for attenuation. *Educ. Psychol. Meas.* 56, 63–75.

Nimon, K., and Henson, R. K. (2010). Validity of residualized variables after pretest covariance corrections: still the same variable? *Paper Presented at the Annual Meeting of the American Educational Research Association*, Denver, CO.

Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric Theory*, 2nd Edn. New York: McGraw-Hill.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory*, 3rd Edn. New York: McGraw-Hill.

Onwuegbuzie, A. J., Roberts, J. K., and Daniel, L. G. (2004). A proposed new "what if" reliability analysis for assessing the statistical significance of bivariate relationships. *Meas. Eval. Couns. Dev.* 37, 228–239.

Osborne, J., and Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Pract. Assess. Res. Eval.* 8. Available at: http://PAREonline.net/getvn.asp?v=8&n=2

Pajares, F., and Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemp. Educ. Psychol.* 24, 124–139.

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction*, 3rd Edn. Fort Worth, TX: Harcourt Brace.

Pintrich, P. R., Smith, D. A. F., Garcia, T., and McKeachie, W. J. (1991). *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Ann Arbor: University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning.

Popham, W. J. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders*, 3rd Edn. Needham, MA: Allyn & Bacon.

Public Agenda. (2011). *State-by-State SAT and ACT Scores*. Available at: http://www.publicagenda.org/charts/state-state-sat-and-act-scores

Rodriguez, M. C., and Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychol. Methods* 11, 306–322.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* 86, 638–641.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychol. Bull.* 118, 183–192.

Schmidt, F. L., and Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *J. Appl. Psychol.* 62, 529–540.

Shavelson, R., and Webb, N. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: Sage.

Shields, A. L., and Caruso, J. C. (2004). A reliability induction and reliability generalization study of the Cage Questionnaire. *Educ. Psychol. Meas.* 64, 254–270.

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101.

Stemler, S. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Pract. Assess. Res. Eval.* 9. Available at: http://PAREonline.net/getvn.asp?v=9&n=4

Thompson, B. (2000). "Ten commandments of structural equation modeling," in *Reading and Understanding More Multivariate Statistics*, eds L. Grimm and P. Yarnold (Washington, DC: American Psychological Association), 261–284.

Thompson, B. (2003a). "Guidelines for authors reporting score reliability estimates," in *Score Reliability: Contemporary Thinking on Reliability Issues*, ed. B. Thompson (Newbury Park, CA: Sage), 91–101.

Thompson, B. (2003b). "A brief introduction to generalizability theory," in *Score Reliability: Contemporary Thinking on Reliability Issues*, ed. B. Thompson (Newbury Park, CA: Sage), 43–58.

Thompson, B., and Vacha-Haase, T. (2000). Psychometrics is datametrics: the test is not reliable. *Educ. Psychol. Meas.* 60, 174–195.

Trafimow, D., and Rice, S. (2009). Potential performance theory (PPT): describing a methodology for analyzing task performance. *Behav. Res. Methods* 41, 359–371.

Vacha-Haase, T. (1998). Reliability generalization: exploring variance in measurement error affecting score reliability across studies. *Educ. Psychol. Meas.* 58, 6–20.

Vacha-Haase, T., Henson, R. K., and Caruso, J. C. (2002). Reliability generalization: moving toward improved understanding and use of score reliability. *Educ. Psychol. Meas.* 62, 562–569.

Vacha-Haase, T., Kogan, L, R., and Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: validity of score reliability inductions. *Educ. Psychol. Meas.* 60, 509–522.

Vacha-Haase, T., Ness, C., Nillson, J., and Reetz, D. (1999). Practices regarding reporting of reliability coefficients: a review of three journals. *J. Exp. Educ.* 67, 335–341.

Vacha-Haase, T., and Thompson, B. (2011). Score reliability: a retrospective look back at 12 years of reliability generalization studies. *Meas. Eval. Couns. Dev.* 44, 159–168.

Wetcher-Hendricks, D. (2006). Adjustments to the correction for attenuation. *Psychol. Methods* 11, 207–215.

Wilkinson, L., and APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanation. *Am. Psychol.* 54, 594–604.

Yetkiner, Z. E., and Thompson, B. (2010). Demonstration of how score reliability is integrated into SEM and how reliability affects all statistical analyses. *Mult. Linear Regress. Viewp.* 26, 1–12.

Zientek, L. R., Capraro, M. M., and Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: one look at the evidence cited in the AERA panel report. *Educ. Res.* 37, 208–216.

Zientek, L. R., and Thompson, B. (2009). Matrix summaries improve research reports: secondary analyses using published literature. *Educ. Res.* 38, 343–352.

Zimmerman, D. W. (2007). Correction for attenuation with biased reliability estimates and correlated errors in populations and samples. *Educ. Psychol. Meas.* 67, 920–939.

Zimmerman, D. W., and Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *J. Math. Psychol.* 16, 135–152.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 December 2011; paper pending published: 26 January 2012; accepted: 19 March 2012; published online: 12 April 2012.

*Citation: Nimon K, Zientek LR and Henson RK (2012) The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. Front. Psychology **3**:102. doi: 10.3389/fpsyg.2012.00102*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

APPENDIX

EXAMPLE WRITE-UP FOR SAMPLE, INSTRUMENT, AND RESULT SECTIONS

Sample

A convenience sample of 420 students (200 fifth graders, 220 sixth graders) were from a suburban public intermediate school in the southwest of the United States and included 190 Whites (100 males, 90 females; 135 regular education, 55 gifted education), 105 Blacks (55 males, 50 females; 83 regular education, 22 gifted education), 95 Hispanics (48 males, 47 females; 84 regular education, 11 gifted education), 18 Asians (9 males, 9 females; 13 regular education, 5 gifted education), and 12 Others (5 males, 7 females; 10 regular education, 2 gifted education). The school consisted of 45% of students in high-poverty as defined by number of students on free lunch. None of the students were in special education. Parental and/or student consent was obtained by 94% of the students, providing a high response rate.

INSTRUMENT

Marat (2005) included an instrument that contained several predictors of self-efficacy (see Pintrich et al., 1991; Bandura, 2006). In the present study, five constructs were included: Motivation Strategies (MS: 5 items); Cognitive Strategies (CS; 15 items); Resource Management Strategies (MS; 12 items); Self-Regulated Learning (SRL; 16 items); and Self-Assertiveness (SA; 6 items). No modifications were made to the items or the subscales but only five of the subscales from the original instrument listed above were administered. The English version of the instrument was administered via paper to students by the researchers during regular class time and utilized a five-point Likert scale anchored from 1 (not well) to 5 (very well). Composite scores were created for each construct by averaging the items for each subscale.

RESULTS

Coefficient alpha was calculated for the data in hand resulting in acceptable levels of reliability for MS (0.82, 0.84), CS (0.85, 0.84), RMS (0.91, 0.83), SRL (0.84, 86), and SA (0.87, 0.83), fall and spring, respectively (Thompson, 2003a).

GIFTED AND REGULAR STUDENTS

Table A1 provides the reliability coefficients, bivariate correlations, means, SD for each factor disaggregated by gifted and regular students.

Table A1 | Bivariate correlations, means, SD, and reliability coefficient disaggregated by gifted and regular education students.

	Gifted (n =95)								Regular (n =325)		
Factors	α	M	SD	1.	2.	3.	4.	5.	α	M	SD
1. MS											
2. CS											
3. RMS											
4. SRL											
5. SA											

Gifted Students Information

Provide bivariate correlations: Gifted students below the diagonal and Regular Education students above the diagonal.

Regular Students Information

α, Coefficient alpha; M, mean; SD, standard deviation. Bivariate correlations below the diagonal are for Gifted students and above the diagonal are for Regular education students. MS, motivation strategies; CS, cognitive strategies; RMS, resource management strategies; SRL, self-regulated learning; SA, self-assertiveness.

frontiers in
PSYCHOLOGY

PERSPECTIVE ARTICLE
published: 04 July 2012
doi: 10.3389/fpsyg.2012.00218

Replication unreliability in psychology: elusive phenomena or "elusive" statistical power?

Patrizio E. Tressoldi*

Dipartimento di Psicologia Generale, Università di Padova, Padova, Italy

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Fiona Fidler, University of Melbourne, Australia
Darrell Hull, University North Texas, USA
Donald Sharpe, University of Regina, Canada

****Correspondence:***
Patrizio E. Tressoldi, Dipartimento di Psicologia Generale, Università di Padova, Padova, Italy.
e-mail: patrizio.tressoldi@unipd.it

The focus of this paper is to analyze whether the unreliability of results related to certain controversial psychological phenomena may be a consequence of their low statistical power. Applying the Null Hypothesis Statistical Testing (NHST), still the widest used statistical approach, unreliability derives from the failure to refute the null hypothesis, in particular when exact or quasi-exact replications of experiments are carried out. Taking as example the results of meta-analyses related to four different controversial phenomena, subliminal semantic priming, incubation effect for problem solving, unconscious thought theory, and non-local perception, it was found that, except for semantic priming on categorization, the statistical power to detect the expected effect size (ES) of the typical study, is low or very low. The low power in most studies undermines the use of NHST to study phenomena with moderate or low ESs. We conclude by providing some suggestions on how to increase the statistical power or use different statistical approaches to help discriminate whether the results obtained may or may not be used to support or to refute the reality of a phenomenon with small ES.

Keywords: incubation effect, non-local perception, power, subliminal priming, unconscious thought theory

INTRODUCTION

ARE THERE ELUSIVE PHENOMENA OR IS THERE AN "ELUSIVE" POWER TO DETECT THEM?

When may a phenomenon be considered to be real or very probable, following the rules of current scientific methodology? Among the many requirements, there is a substantial consensus that replication is one of the more fundamental (Schmidt, 2009). In other words, a phenomenon may be considered real or very probable when it has been observed many times and preferably by different people or research groups. Whereas a failure to replicate is quite expected in the case of conceptual replication, or when the experimental procedure or materials entail relevant modifications, a failure in the case of an exact or quasi-exact replication, give rise to serious concerns about the reality of the phenomenon under investigation. This is the case in the four phenomena used as examples in this paper, namely, semantic subliminal priming, incubation effects on problem solving, unconscious thought, and non-local perception (NLP; e.g., Kennedy, 2001; Pratte and Rouder, 2009; Waroquier et al., 2009).

The focus of this paper is to demonstrate that for all phenomena with a moderate or small effect size (ES), approximately below 0.5 if we refer to standardized differences such as Cohen's *d*, the typical study shows a power level insufficient to detect the phenomenon under investigation.

Given that the majority of statistical analyses are based on the Null Hypothesis Statistical Testing (NHST) frequentist approach, their failure is determined by the rejection of the (nil) null hypothesis H_0, usually setting $\alpha < 0.05$. Even if this procedure is considered incorrect because the frequentist approach only supports H_0 rejection and not H_0 validity[1], it may be tolerated if there is proof of a high level of statistical power as recommended in the recent APA statistical recommendations [American Psychological Association, APA (2010)]: "Power: When applying inferential statistics, take seriously the statistical power considerations associated with the tests of hypotheses. Such considerations relate to the likelihood of correctly rejecting the tested hypotheses, given a particular alpha level, ES, and sample size. In that regard, routinely provide evidence that the study has sufficient power to detect effects of substantive interest. Be similarly careful in discussing the role played by sample size in cases in which not rejecting the null hypothesis is desirable (i.e., when one wishes to argue that there are no differences), when testing various assumptions underlying the statistical model adopted (e.g., normality, homogeneity of variance, homogeneity of regression), and in model fitting. pag. 30."

HOW MUCH POWER?

Statistical power depends on three classes of parameters: (1) the significance level (i.e., the Type I error probability) of the test, (2) the size(s) of the sample(s) used for the test, and (3) an ES parameter defining H_1 and thus indexing the degree of deviation from H_0 in the underlying population.

Power analysis should be used prospectively to calculate the minimum sample size required so that one can reasonably detect an effect of a given size. Power analysis can also be used to calculate

[1]The line of reasoning from "the null hypothesis is false" to "the theory is therefore true" involves the logical fallacy of affirming the consequent: "If the theory is true, the null hypothesis will prove to be false. The null hypothesis proved to be false; therefore, the theory must be true" (Nickerson, 2000).

the minimum ES that is likely to be detected in a study using a given sample size.

In most experimental designs, the accepted probability of making a Type I error is $\alpha = 0.05$ and the desired power is not less than 0.80. However, in order to define how to obtain such a level of power, it is necessary to know the ES of the phenomena being identified. It is intuitive that the smaller the phenomenon, the greater should be the means to detect it. This analogy is similar to the signal/noise relationship. The smaller the signal, the stronger must be the means to detect it in the noise. In psychological experiments, these means are the number of participants taking part in the study and the number of trials they are requested to perform.

Given that Power $= 1-\beta = ES^* \sqrt{N/SD^*}\ \alpha$, if we know the estimated ES of a phenomenon, after the definition of the desired power and the α level, the only free parameters is *N*, that is the number of participants or trials.

A review of 322 meta-analyses published before 1998, summarizing 25,000 studies referred to 474 social psychological effects, reports that the mean ES reported is $r = 0.21$ and the mode was less than $r = 0.10$ (Richard et al., 2003). For this observed mean ES, to obtain a statistical power for independent sample *t* tests $= >0.9$, the sample size for each group should be at least 90, a number rarely observed in the studies.

Setting aside the strong criticisms of the use of NHST (Cohen, 1994; Kline, 2004), a neglected aspect of this approach, is the control of how much statistical power is necessary to detect what the researcher aims to find[2].

This problem is not new and has already been raised by Sedlmeier and Gigerenzer (1989), Cohen (1992), Bezeau and Graves (2001), and Maxwell (2004) among others. However the widespread adherence to "The Null Ritual" as discussed by Gigerenzer et al. (2004), which consists in: (a) Set up a statistical null hypothesis of "no mean difference" or "zero correlation." (b) Don't specify the predictions of your research hypothesis or of any alternative substantive hypotheses; (c) Use 5% as a convention for rejecting the null; (d) If significant, accept your research hypothesis (e) Always perform this procedure, seems to prevent most researchers taking into account such fundamental statistical parameter. In Gigerenzer et al. (2004) check, covering the years 2000–2002, and encompassing with some 220 empirical articles, only nine researchers who computed the power of their tests were found.

Using the results of four recent meta-analyses related to "controversial" or "elusive" psychological phenomena, we illustrate the importance of using the available ESs to derive the appropriate number of participants to achieve a power $= >0.90$. Only if a replication fails with this level of power, it is legitimate to raise doubts about the reality of the phenomena under investigation.

CAN SUBLIMINALLY PRESENTED INFORMATION INFLUENCE BEHAVIOR?

This question is one of the most controversial questions in psychology and it remains an intriguing and strongly debated issue. Apart from the controversies relating to the controls that information (priming) was effectively masked and the identification of serious methodological flaws which caused great doubt as to the existence of subliminal processing, the debate is even hotter around the topic of the level of influence of this unconscious (subliminal) information. Hitherto, the debate as to whether subliminal priming reflects genuine semantic processing of the subliminal information or the formation of automatic S–R mappings remains unresolved.

The meta-analysis of Van den Bussche et al. (2009) tried to shed light on these questions by analyzing all the available literature between 1983 and December 2006. Their analysis was carried out separately from the two most studied protocols, subliminal priming for semantic categorization and subliminal priming for lexical decision and naming. If semantic subliminal priming facilitated the subsequent categorization of targets belonging to the same semantic category, it suggests that the primes were unconsciously categorized and processed semantically. The same effect is postulated if lexical decision and naming are faster or more accurate to semantically subliminal related prime–target pairs than to unrelated pairs. A naming task is similar to the lexical decision task except that the targets are all words, and participants are asked to name the targets aloud.

The synthesis of the main results is reported in **Table 1**.

DOES INCUBATION ENHANCE PROBLEM SOLVING?

The "incubation period" is the temporary shift away from an unsolved problem in order to allow a solution to emerge in the mind of the individual, seemingly with no additional effort, after he or she has put the problem aside for a period of time, having failed in initial attempts to solve it.

Among the questions under debate there is the problem of whether the nature of the discovery of the solution is really unconscious and if it is qualitatively different from that used to tackle problems that do not require such insight.

The meta-analysis of Sio and Ormerod (2009) tried to shed light on this and other related questions, analyzing all the available literature from 1964 to 2007. The main findings of their meta-analysis are reported in **Table 1**.

UNCONSCIOUS THOUGHT THEORY

The key assumption of Unconscious Thought Theory (UTT, Dijksterhuis et al., 2006) is that unconscious thought and conscious thought are characterized by different processes. That is, "unconscious thought" processes have a relatively large capacity – hence, they allow for an optimal decision strategy in which all attributes of chosen alternatives are weighted according to their importance. These unconscious processes require time, therefore the quality of decisions increases with the duration of unconscious thought. "Conscious thought" processes on the other hand, have a small capacity and therefore only allow for simplified decision making strategies. As summarized by Dijksterhuis et al., 2006, p. 105): "When a decision strategy warrants the careful and strict application of one specific rule, as in a lexicographic strategy, use conscious thought. When matters become more complicated and weighting is called for, as in the weighting strategy, use unconscious thought."

As expected, among the questions raised by this theory, a critical problem is whether really optimal decisions are obtained after

[2]Even if power estimate is important in meta-analyses (i.e., Valentine et al., 2009), in this paper we focus only on power estimates in single studies.

Table 1 | Descriptive statistics of the four meta-analyses, related to Unconscious Semantic Priming, Incubation effect, UTT, and NLP with the estimated power of a typical study, the mean and 95% CI power calculated from all studies included in the meta-analysis and the number of participants necessary to obtain a Power = 0.90.

Phenomena	Protocols	Source	*N* studies	Averaged *N* × study (range)	ES* (mean and 95% CI)	*p*	Estimated power of a typical study	Observed power (mean and 95% CI)	Estimated N to achieve power = 0.90
Semantic priming	Semantic categorization	Van den Bussche et al. (2009)	23	20 (6–80)	0.80 ± 0.2	n.a.	0.96	0.90 (±1.7)	
	Lexical and naming	Van den Bussche et al. (2009)	32	33 (9–132)	0.47 ± 0.11	n.a.	0.84	0.69 (±0.4.6)	40
Incubation effect		Sio and Ormerod (2009)	117	31 (7–278)	0.29 ± 0.09	n.a.	0.21	0.43 (±3.3)	100
Unconscious thought theory		Strick et al. (2011)	92	26 (14–55)	0.22 ± 0.08	1.2×10^{-8}	0.20§	0.19 (±0.72)	400§
Non-local perception	Remote Vision	Milton (1997)	78	34 (1–74)	0.16 ± 0.06	6×10^{-9}	0.55	n.a.	73
	Ganzfeld	Storm et al. (2010)	108	40 (7–128)	0.13 ± 0.04	8.3×10^{-11}	0.46	0.45 (±3.8)	110
	Forced-choice with normal state of consciousness	Storm et al. (in press)	72	128 (12–887)	0.011 ± 0.004	5.3×10^{-7}	0.07	0.065 (±0.81)	3450

**Random effect Cohen's d; §two groups comparison.*

a period of distraction from deliberate conscious mental activity for the same amount of time as would be the case had the decisions been taken deliberately.

The meta-analysis of Strick et al. (2011), aimed to give an answer to this and other related questions by analyzing all the available evidence up to May 2011. The main findings are reported in **Table 1**.

NON-LOCAL PERCEPTION

Non-local perception (NLP) is based on the hypothesis that the human mind may have quantum like properties, that is, that some of its functions, such as perceptual abilities, reasoning, etc., may be analyzed using quantum formalism.

The main non-local properties which are studied within the realm of quantum physics and which are supported by "extraordinary evidence" (see Genovese, 2005, 2010), are "entanglement" and "measurement interference." The first property, entanglement, allows two or more physical objects to behave as one even if they are separated in space and time. This "strange" property allows a form of immediate communication of the objects' characteristics over distances between or among the entangled objects, as has been observed in teleportation experiments (i.e., Bouwmeester et al., 1997). The possibility that quantum-like properties may be observed not only in physics but also even in biology and psychology has not only been studied theoretically (von Lucadou et al., 2007; Khrennikov, 2010; Walach and von Stillfried, 2011) but also experimentally (see Gutiérrez et al., 2010 for biology and Busemeyer et al., 2011, for psychology).

One of the main concerns about the studies related to different aspects of NLP, is their inconsistency in obtaining results satisfying the statistical cut off criteria to refute the null hypothesis that extra sensory perception does not exist, usually setting $\alpha < 0.05$.

This problem is recognized by some researchers involved in this field of research (Kennedy, 2001) and even more by all deniers of NLP evidence (e.g., Alcock, 2003).

The main findings of three meta-analysis related to three NLP different protocols, Remote Vision, that is NLP using a free-choice response (Milton, 1997), NLP in a Ganzfeld state using free-choice response (Storm et al., 2010) and NLP in a normal state of consciousness using a forced-choice response (Storm et al., in press), covering all evidence available from 1964 to 1992, 1974 to 2009, and 1987 to 2010 respectively, are reported in **Table 1**

POWER ESTIMATION

A synthesis of the descriptive statistics related to the four phenomena described above is presented in **Table 1** in decreasing order of magnitude of ESs. For each meta-analysis, the retrospective statistical power with $\alpha = 0.05$, achieved by a typical study using the mean of the number of participants of all studies included in the meta-analysis was estimated.

For all but one meta-analysis, it was also possible to calculate the mean and 95% CI *post hoc* power, using the number of participants of each study included in the meta-analyses and the estimated random ES, setting $\alpha = 0.05$.

Furthermore, for each of the four psychological phenomena, the number of participants necessary to obtain a statistical

power = 0.9 with $\alpha = 0.05$ given the observed random ESs, was estimated.

Statistical power was calculated using the software G*Power (Faul et al., 2007).

COMMENT

The results are quite clear: apart from the unconscious semantic priming for semantic categorization, where the number of participants in a typical experiment is sufficient to obtain a statistical power above 0.90, for all remaining phenomena, to achieve this level of power, it is necessary to increase the number of participants in a typical study, from a minimum of seven participants for the unconscious semantic priming for lexical decision and naming to around 3400 to investigate NLP using the forced-choice with normal state of consciousness protocol.

GENERAL DISCUSSION

The response to the question posed in the introduction, as to whether there are elusive phenomena or an elusive power to detect them, is quite clear. If there are clear estimates of ESs from the evidence of the phenomenon derived from a sufficient number of studies analyzed meta-analytically and their values are moderate or low, it is mandatory to increase the number of participants to achieve a statistical power of 0.90, with the inevitable consequence of investing more time and money into each study before interpreting the results as support for reality or unreality of a phenomenon.

Are there alternatives to this obligation? Yes, and we briefly illustrate some of these, also providing references for those interested in using them.

CONFIDENCE INTERVALS

In line with the statistical reform movement (i.e., Cumming, 2012), in the APA manual (American Psychological Association, APA, 2010), there are the following statistical recommendations "Alternatively, (to the use of NHST) use calculations based on a chosen target precision (confidence interval width) to determine sample sizes. Use the resulting confidence intervals to justify conclusions concerning ESs (e.g., that some effect is negligibly small) p. 30."

EQUIVALENCE TESTING

Equivalence tests are inferential statistics designed to provide evidence for a null hypothesis. Like effect tests, the nil–null is eschewed in equivalence testing. However unlike standard NHST, equivalence tests provide evidence that there is little difference or effect. A significant result in an equivalence test means that the hypothesis that the effects or differences are substantial can be rejected. Hence, equivalence tests are appropriate when researchers want to show little difference or effect (Levine et al., 2008).

EVALUATING INFORMATIVE HYPOTHESES

Evaluating specific expectations directly produces more useful results than sequentially testing traditional null hypotheses against catch-all rivals. Researchers are often interested in the evaluation of informative hypotheses and already know that the traditional null hypothesis is an unrealistic hypothesis. This presupposes that prior knowledge is often available; if this is not the case, testing the traditional null hypothesis is appropriate. In most applied studies, however, prior knowledge is indeed available in the form of specific expectations about the ordering of statistical parameters (Kuiper and Hoijtink, 2010; Van de Schoot et al., 2011).

BAYESIAN APPROACH

Another alternative is to abandon the frequentist approach and use a Bayesian one (Wagenmakers et al., 2011). With a Bayesian approach the problem of statistical power is substituted with parameter estimation and/or model comparison (Kruschke, 2011). In the first approach, assessing null values, the analyst simply sets up a range of candidate values, including the null value, and uses Bayesian inference to compute the relative credibility of all the candidate values. In the model comparison approach, the analyst sets up two competing models of what values are possible. One model posits that only the null value is possible whereas the alternative model posits that a broad range of other values is also possible. Bayesian inference is used to compute which model is more credible, given the data.

FINAL COMMENT

Is there a chance to abandon "The Null Ritual" in the near future and to think of science as cumulative knowledge? The answer is "yes" if we approach scientific discovery thinking meta-analytically (Cumming, 2012), that is, simply reporting observed (standardized) ES and the corresponding confidence intervals, both when NHST is refuted and when it is not refuted (Nickerson, 2000; American Psychological Association, APA, 2010) without drawing dichotomous decisions. The statistical approaches listed above are good tools to achieve this goal.

How many editors and reviewers are committed to pursuing it?

ACKNOWLEDGMENTS

Suggestions and comments by the reviewers were greatly appreciated for improving the clarity and quality of the paper. Proof Reading Service revised the English.

REFERENCES

Alcock, J. E. (2003). Give the null hypothesis a chance: reasons to remain doubtful about the existence of PSI. *J. Conscious. Stud.* 10, 29–50.

American Psychological Association. (2010). *Publication manual of the American Psychological Association*, 6th Edn, Washington, DC: American Psychological Association.

Bezeau, S., and Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *J. Clin. Exp. Neuropsychol.* 23, 399–406.

Bouwmeester, D., Pan, J. W., Mattle, K., Eibl, M., Weinfurter, H., and Zeilinger, A. (1997). Experimental quantum teleportation. *Nature* 390, 575–579.

Busemeyer, J. R., Pothos, E. M., Franco, R., and Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychol. Rev.* 118, 2, 193–218.

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 1,155–1,159.

Cohen, J. (1994). The earth is round (p < .085). *Am. Psychol.* 49, 997–1003.

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* New York: Routledge.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., and Van Baaren, R. B. (2006). On making the right choice: the deliberation-without attention effect. *Science* 311, 1005–1007.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191.

Genovese, M. (2005). Research on hidden variable theories, a review of recent progresses. *Phys. Rep.* 413, 319–396.

Genovese, M. (2010). Interpretations of quantum mechanics and measurement problem. *Adv. Sci. Lett.* 3, 249–258.

Gigerenzer, G., Krauss, S., and Vitouch, O. (2004). "The null ritual what you always wanted to know about significance testing but were afraid to ask," in *The Sage Handbook of Quantitative Methodology for the Social Sciences,* ed. D. Kaplan (Thousand Oaks, CA: Sage), 391–408.

Gutiérrez, R., Caetano, R., Woiczikowski, P. B., Kubar, T., Elstner, M., and Cuniberti, G. (2010). Structural fluctuations and quantum transport through DNA molecular wires, a combined molecular dynamics and model Hamiltonian approach. *New J. Phys.* 12, 023022.

Kennedy, J. E. (2001). Why is PSY so elusive? A review and proposed model. *J. Parapsychol.* 65, 219–246.

Khrennikov, A. Y. (2010). *Ubiquitous Quantum Structure from Psychology to Finance.* Berlin: Springer-Verlag.

Kline, R. B. (2004). *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research.* Washington, DC: APA.

Kruschke, J. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspect. Psychol. Sci.* 6, 299–312.

Kuiper, R. M., and Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychol. Methods* 15, 69–86.

Levine, T. R., Weber, R., Sun Park, H., and Hullett, C. R. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Hum. Commun. Res.* 34, 188–209.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol. Methods* 9, 147–163.

Milton, J. (1997). Meta-analysis of free-response ESP studies without altered states of consciousness. *J. Parapsychol.* 61, 279–319.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301.

Pratte, M. S., and Rouder, J. N. (2009). A task-difficulty artifact in subliminal priming. *Atten. Percept. Psychophys.* 71,1276–1283.

Richard, F. D., Bond, C. F., and Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Rev. Gen. Psychol.* 7, 331–363.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* 13, 90–100.

Sedlmeier, P., and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105, 309–316.

Sio, U. N., and Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychol. Bull.* 135, 94–120.

Storm, L., Tressoldi, P. E., and Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: assessing the noise reduction model in parapsychology. *Psychol. Bull.* 136, 471–485.

Storm, L., Tressoldi, P. E., and Di Risio, L. (in press). Meta-analysis of ESP studies, 1987–2010, assessing the success of the forced-choice design in parapsychology.

Strick, M., Dijksterhuis, A., Bos, M. W., Sjoerdsma, A., van Baaren, R. B., and Nordgren, L. F. (2011). A meta-analysis on unconscious thought effects. *Soc. Cogn.* 29, 738–762.

Valentine, J. C., Pigott, T. D., and Rothstein, H. R. (2009). How many studies to you need? A primer on statistical power for meta-analysis. *J. Educ. Behav. Stat.* 35, 215–247.

Van de Schoot, R., Hoijtink, H., and Jan-Willem, R. (2011). Moving beyond traditional null hypothesis testing: evaluating expectations directly. *Front. Psychol.* 2:24. doi:10.3389/fpsyg.2011.00024

Van den Bussche, E., den Noortgate, W., and Reynvoet, B. (2009). Mechanisms of masked priming: a meta-analysis. *Psychol. Bull.* 135, 452–477.

von Lucadou, W., Römer, H., and WAlach, H. (2007). Synchronistic phenomena as entanglement correlations in generalized quantum theory. *J. Conscious. Stud.* 14, 4, 50–74.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., and Van der Maas, H. (2011). Why psychologists must change the way they analyze their data: the case of psi. *J. Pers. Soc. Psychol.* 100, 426–432.

Walach, H., and von Stillfried, N. (2011). Generalised quantum theory. Basic idea and general intuition: a background story and overview. *Axiomathes* 21, 185–209.

Waroquier, L., Marchiori, D., Klein, O., and Cleeremans, A. (2009). Methodological pitfalls of the unconscious thought paradigm. *Judgm. Decis. Mak.* 4, 601–610.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 April 2012; accepted: 12 June 2012; published online: 04 July 2012.
*Citation: Tressoldi PE (2012) Replication unreliability in psychology: elusive phenomena or "elusive" statistical power? Front. Psychology **3**:218. doi: 10.3389/fpsyg.2012.00218*
This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 July 2012
doi: 10.3389/fpsyg.2012.00211

Distribution of variables by method of outlier detection

W. Holmes Finch*

Department of Educational Psychology, Ball State University, Muncie, IN, USA

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Matt Jans, The United States Census Bureau, USA
Avi Allalouf, National Institute for Testing and Evaluation, Israel

****Correspondence:***
W. Holmes Finch, Department of Educational Psychology, Ball State University, Muncie, IN 47304, USA.
e-mail: whfinch@bsu.edu

The presence of outliers can very problematic in data analysis, leading statisticians to develop a wide variety of methods for identifying them in both the univariate and multivariate contexts. In case of the latter, perhaps the most popular approach has been Mahalanobis distance, where large values suggest an observation that is unusual as compared to the center of the data. However, researchers have identified problems with the application of this metric such that its utility may be limited in some situations. As a consequence, other methods for detecting outlying observations have been developed and studied. However, a number of these approaches, while apparently robust and useful have not made their way into general practice in the social sciences. Thus, the goal of this study was to describe some of these methods and demonstrate them using a well known dataset from a popular multivariate textbook widely used in the social sciences. Results demonstrated that the methods do indeed result in datasets with very different distributional characteristics. These results are discussed in light of how they might be used by researchers and practitioners.

Keywords: Mahalanobis distance, minimum covariance determinant, minimum generalized variance, minimum volume ellipsoid, outliers, projection

INTRODUCTION

The presence of outliers is a ubiquitous and sometimes problematic aspect of data analysis. They can result from a variety of processes, including data recording and entry errors, obtaining samples from other than the target population, and sampling unusual individuals from the target population itself (Kruskal, 1988). Based on the standard definition of outliers, it is entirely possible that a dataset may not have any such cases, or it might have many. Given that they can arise from very different processes, outliers should not all be treated in the same manner. For example, those caused by data collection problems are likely to be removed from the sample prior to analysis, while those that are simply unusual members of the target population would be retained for data analysis. Finally, Kruskal noted that in some cases understanding the mechanism that caused outliers is the most important aspect of a given study. In other words, outliers can themselves provide useful information to researchers, and are not necessarily problematic in the sense of being bad data. The focus of this manuscript is not on the mechanism giving rise to outliers, but rather on methods for detecting them once they are present in the sample.

A number of authors have sought to precisely define what constitutes an outlier (e.g., Evans, 1999), and methods for detecting and dealing with them once detected remain an active area of research. It is well known that outliers can have a dramatic impact on the performance of common statistical analyses such as Pearson's correlation coefficient (Marascuilo and Serlin, 1988), univariate, and multivariate means comparisons (Kirk, 1995; Huberty and Olejnik, 2006), cluster analysis (Kaufman and Rousseeuw, 2005), multivariate means comparisons, and factor analysis (Brown, 2006), among others. For this reason researchers are strongly encouraged to investigate their data for the presence of outliers prior to conducting data analysis (Tabachnick and Fidell, 2007).

In the multivariate context, the most commonly recommended approach for outlier detection is the Mahalanobis Distance (D^2). While this approach can be an effective tool for such purpose, it also has weaknesses that might render it less than effective in many circumstances (Wilcox, 2005). The focus of this manuscript is on describing several alternative methods for multivariate outlier detection; i.e., observations that have unusual patterns on multiple variables as opposed to extreme scores on a single variable (univariate outliers). In addition, these approaches will be demonstrated, along with D^2, using a set of data taken from Tabachnick and Fidell (2007). The demonstration will utilize functions from the R software package for both outlier detection and data analysis after removal of the outliers. It should be noted that the focus of this manuscript is not on attempting to identify some optimal approach for dealing with outliers once they have been identified, which is an area of statistics itself replete with research, and which is well beyond the scope of this study. Suffice it to say that identification of outliers is only the first step in the process, and much thought must be given to how outliers will be handled. In the current study, they will be removed from the dataset in order to clearly demonstrate the differential impact of the various outlier detection methods on the data and subsequent analyses. However, it is not recommended that this approach to outliers be taken in every situation.

IMPACT OF OUTLIERS IN MULTIVARIATE ANALYSIS

Outliers can have a dramatic impact on the results of common multivariate statistical analyses. For example, they can distort correlation coefficients (Marascuilo and Serlin, 1988; Osborne and

Overbay, 2004), and create problems in regression analysis, even leading to the presence of collinearity among the set of predictor variables in multiple regression (Pedhazur, 1997). Distortions to the correlation may in turn lead to biased sample estimates, as outliers artificially impact the degree of linearity present between a pair of variables (Osborne and Overbay, 2004). In addition, methods based on the correlation coefficient such as factor analysis and structural equation modeling are also negatively impacted by the presence of outliers in data (Brown, 2006). Cluster analysis is particularly sensitive to outliers with a distortion of cluster results when outliers are the center or starting point of the analysis (Kaufman and Rousseeuw, 2005). Outliers can also themselves form a cluster, which is not truly representative of the broader array of values in the population. Outliers have also been shown to detrimentally impact testing for mean differences using ANOVA through biasing group means where they are present (Osborne and Overbay, 2004).

While outliers can be problematic from a statistical perspective, it is not always advisable to remove them from the data. When these observations are members of the target population, their presence in the dataset can be quite informative regarding the nature of the population (e.g., Mourão-Miranda et al., 2011). To remove outliers from the sample in this case would lead to loss of information about the population at large. In such situations, outlier detection would be helpful in terms of identifying members of the target population who are unusual when compared to the rest, but these individuals should not be removed from the sample (Zijlstra et al., 2011).

METHODS OF MULTIVARIATE OUTLIER DETECTION

Given the negative impact that outliers can have on multivariate statistical methods, their accurate detection is an important matter to consider prior to data analysis (Tabachnick and Fidell, 2007; Stevens, 2009). In popular multivariate statistics texts, the reader is recommended to use D^2 for multivariate outlier detection, although as is described below, there are several alternatives for multivariate outlier detection that may prove to be more effective than this standard approach. Prior to discussing these methods however, it is important to briefly discuss general qualities that make for an effective outlier detection method. Readers interested in a more detailed treatment are referred to two excellent texts by Wilcox (2005, 2010).

When thinking about the impact of outliers, perhaps the key consideration is the breakdown point of the statistical analysis in question. The breakdown point can be thought of as the minimum proportion of a sample that can consist of outliers after which point they will have a notable impact on the statistic of interest. In other words, if a statistic has a breakdown point of 0.1, then 10% of the sample could consist of outliers without markedly impacting the statistic. However, if the next observation beyond this 10% was also an outlier, the statistic in question would then be impacted by its presence (Maronna et al., 2006). Comparatively, a statistic with a breakdown point of 0.3 would be relatively more impervious to outliers, as it would not be impacted until more than 30% of the sample was made up of outliers. Of course, it should be remembered that the degree of this impact is dependent on the magnitude of the outlying observation, such that more extreme outliers would have a greater impact on the statistic than would a less extreme value. A high breakdown point is generally considered to be a positive attribute.

While the breakdown point is typically thought of as a characteristic of a statistic, it can also be a characteristic of a statistic in conjunction with a particular method of outlier detection. Thus, if a researcher calculates the sample mean after removing outliers using a method such as D^2, the breakdown point of the combination of mean and outlier detection method will be different than that of the mean by itself. Finally, although having a high breakdown point is generally desirable, it is also true that statistics with higher breakdown points (e.g., the median, the trimmed mean) are often less accurate in estimating population parameters when the data are drawn from a multivariate normal distribution (Genton and Lucas, 2003).

Another important property for a statistical measure of location (e.g., mean) is that it exhibit both location and scale equivariance (Wilcox, 2005). Location equivariance means that if a constant is added to each observation in the data set, the measure of location will be increased by that constant value. Scale equivariance occurs when multiplication of each observation in the data set by a constant leads to a change in the measure of location by the same constant. In other words, the scale of measurement should not influence relative comparisons of individuals within the sample or relative comparisons of group measures of location such as the mean. In the context of multivariate data, these properties for measures of location are referred to as affine equivariance. Affine equivariance extends the notion of equivariance beyond changes in location and scale to measures of multivariate dispersion. Covariance matrices are affine equivariant, for example, though they are not particularly robust to the presence of outliers (Wilcox, 2005). A viable approach to dealing with multivariate outliers must maintain affine equivariance.

Following is a description of several approaches for outlier detection. For the most part, these descriptions are presented conceptually, including technical details only when they are vital to understanding how the methods work. References are provided for the reader who is interested in learning more about the technical aspects of these approaches. In addition to these descriptions, **Table 1** also includes a summary of each method, including fundamental equations, pertinent references, and strengths and weaknesses.

MAHALANOBIS DISTANCE

The most commonly recommended approach for multivariate outlier detection is D^2, which is based on a measure of multivariate distance first introduced by Mahalanobis (1936), and which has been used in a wide variety of contexts. D^2 has been suggested as the default outlier detection method by a number of authors in popular textbooks used by researchers in a variety of fields (e.g., Johnson and Wichern, 2002; Tabachnick and Fidell, 2007). In practice, a researcher would first calculate the value of D^2 for each subject in the sample, as follows:

$$D_i^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \quad (1)$$

Table 1 | Summary of outlier detection methods.

Method	Equation	Reference	Strengths	Weaknesses
D_i^2	$(x_i - \bar{x})' S^{-1} (x_i - \bar{x})$	Mahalanobis (1936)	Intuitively easy to understand; easy to calculate; familiar to other researchers	Sensitive to outliers; assumes data are continuous
MVE	Identify subset of data contained within the ellipsoid that has minimized volume	Rousseeuw and Leroy (1987)	Yields mean with maximum possible breakdown point	May remove as much as 50% of sample
MCD	Identify subset of data that minimizes the determinant of the covariance matrix	Rousseeuw and van Driessen (1999)	Yields mean with maximum possible breakdown point	May remove as much as 50% of sample
MGV	Calculate MAD version of D^2 as $\sum_{j=1}^{n} \sqrt{\sum_{l=1}^{p} \left(\frac{x_{jl} - x_{il}}{\text{MAD}_l}\right)^2}$ to identify most central points; calculate variance of this central set as additional observations are added one by one; examine this generalized variance and retain those with values less than the adjusted median $M_G + \sqrt{\chi^2_{0.975.p}}(q_3 - q_1)$	Wilcox (2005)	Typically removes fewer observations than either MVE or MCD	Generally does not have as high a breakdown point as MVE or MCD
P1	Identify the multivariate center of data using MCD or MVE and then determine its relative distance from this center (depth); use the MGV criteria based on this depth to identify outliers	Donoho and Gasko (1992)	Approximates an affine equivariant outlier detection method; may not exclude as many cases as MVE or MCD	Will not typically lead to a mean with the maximum possible breakdown point
P2	Identify all possible lines between all pairs of observations in order to determine depth of each point	Donoho and Gasko (1992)	Some evidence that this method is more accurate than P1 in terms of identifying outliers	Extensive computational time, particularly for large datasets
P3	Same approach as P1 except that the criteria for identifying outliers is $M_D + \sqrt{\chi^2_{0.975.p}}(\frac{\text{MAD}_i}{0.6745})$	Donoho and Gasko (1992)	May yield a mean with a higher breakdown point than other projection methods	Will likely lead to exclusion of more observations as outliers than will other projection approaches

where

x_i = Vector of scores on the set of p variables for subject i

$\bar{x}$ = Vector of sample means on the set of p variables

S = Covariance matrix for the p variables

A number of recommendations exist in the literature for identifying when this value is large; i.e., when an observation might be an outlier. The approach used here will be to compare D_i^2 to the χ^2 distribution with p degrees of freedom and declare an observation to be an outlier if its value exceeds the quantile for some inverse probability; i.e., $\chi^2_p(0.005)$ (Mahalanobis).

D^2 is easy to compute using existing software and allows for direct hypothesis testing regarding outlier status (Wilcox, 2005). Despite these advantages, D^2 is sensitive to outliers because it is based on the sample covariance matrix, **S**, which is itself sensitive to outliers (Wilcox, 2005). In addition, D^2 assumes that the data are continuous and not categorical so that when data are ordinal, for example, it may be inappropriate for outlier detection (Zijlstra et al., 2007). Given these problems, researchers have developed alternatives to multivariate outlier detection that are more robust and more flexible than D^2.

MINIMUM VOLUME ELLIPSOID

One of the earliest of alternative approach to outlier detection was the Minimum Volume Ellipsoid (MVE), developed by Rousseeuw and Leroy (1987). In concept, the goal behind this method is to identify a subsample of observations of size h (where $h < n$) that creates the smallest volume ellipsoid of data points, based on the values of the variables. By definition, this ellipsoid should be free of outliers, and estimates of central tendency and dispersion would be obtained using just this subset of observations. The MVE approach to dealing with outliers can, in practice, be all but intractable to carry out as the number of possible ellipsoids to investigate will typically be quite large. Therefore, an alternative approach is to take a large number of random samples of size h with replacement, where

$$h = \frac{n}{2} + 1, \tag{2}$$

and calculate the volume of the ellipsoids created by each. The final sample to be used in further analyses is that which yields the smallest ellipsoid. An example of such an ellipsoid based on MVE can be seen in **Figure 1**. The circles represent observations that have been retained, while those marked with a star represent outliers that will be removed from the sample for future analyses.

MINIMUM COVARIANCE DETERMINANT

The minimum covariance determinant (MCD) approach to outlier detection is similar to the MVE in that it searches for a portion of the data that eliminates the presence and impact of outliers. However, whereas MVE seeks to do this by minimizing the volume of an ellipsoid created by the retained points, MCD does it by

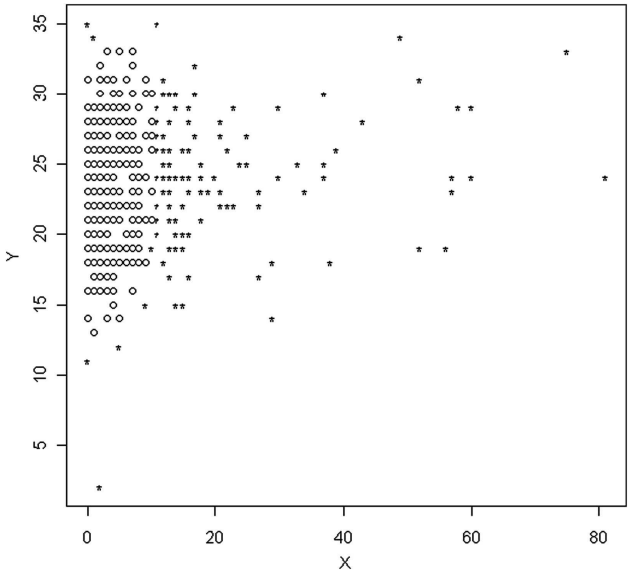


FIGURE 1 | Scatterplot of observations identified as outliers based on the MVE method.

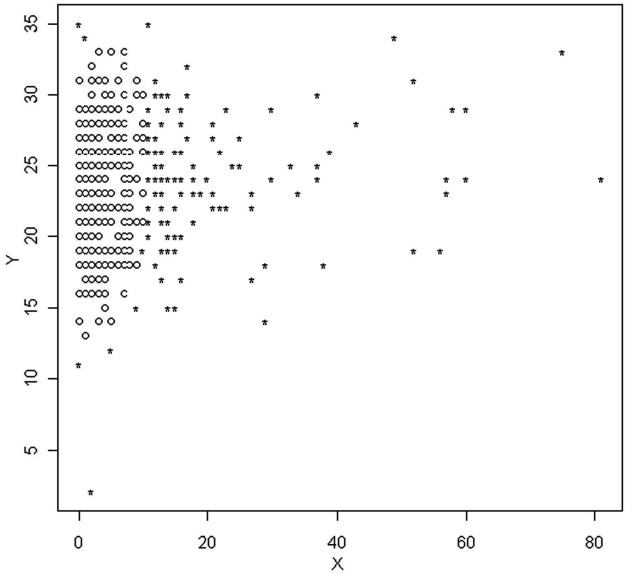


FIGURE 2 | Scatterplot of observations identified as outliers based on the MCD method.

minimizing the determinant of the covariance matrix, which is an estimate of the generalized variance in a multivariate set of data (Rousseeuw and van Driessen, 1999). The dataset with the smallest determinant will be the one least influenced by outliers and which can then be used for future statistical analyses. Statistics calculated on data to which MCD and MVE have been applied will typically have high breakdown points (Rousseeuw and van Driessen, 1999).

As with MVE, the logistics of searching every possible subset of the data of size h to find the one that yields the smallest determinant are not practical in the vast majority of situations. As a consequence Rousseeuw and van Driessen (1999) developed a multiple step algorithm to approximate the MCD, obviating the need to examine all possible subsets of the data. This approach, known as Fast MCD involves the random selection of an initial subsample from the data of size h, for which the values of D_i^2 are calculated and ordered from smallest to largest. The h smallest D_i^2 values (and thus the data points associated with them) are then retained into a new subset of the data, after which individuals from the full dataset are randomly added and the value of the determinant calculated. The algorithm stops when it attains a subsample (size h) of the full data that yields the smallest determinant. Variants of this algorithm involve the selection of multiple subsamples in the initial step, and with several minimization procedures running parallel to one another simultaneously (Hardin and Rocke, 2004). **Figure 2** includes a scatterplot identifying individuals as outliers based on the MCD method. Again, outliers are marked with a star.

MINIMUM GENERALIZED VARIANCE

One potential difficulty with both MVE and MCD is that they tend to identify a relatively large number of outliers when the variables under examination are not independent of one another (Wilcox, 2005). A third approach for outlier detection that was designed to avoid this problem is the Minimum Generalized Variance (MGV). MGV is based on a similar principle to MCD in that the set of data with the smallest overall variance is identified. However, rather than relying on the random addition of observations to the core data set to be retained, it includes those individuals whose inclusion increases the generalized variance as little as possible.

As with MVE and MCD, MGV is an iterative procedure. In the first step the p most centrally located points are identified using a non-parametric estimate of D_i which is calculated as

$$D_i = \sum_{j=1}^{n} \sqrt{\sum_{l=1}^{p} \left(\frac{x_{jl} - x_{il}}{\mathrm{MAD}_l}\right)^2}, \tag{3}$$

where

$$\mathrm{MAD}_l = \mathrm{MED}\{[x_i - M]\}. \tag{4}$$

In other words, MAD, the median absolute deviation, is the median of the deviations between each individual data point and the median of the data set, M. The most centrally located observations are those with the smallest value of D_i as calculated above. These points are then placed in a new data set, after which the generalized variance associated with adding each of the remaining observations not originally placed in this new data is calculated. The observation with the smallest generalized variance is then added to the new data set. For each data point remaining outside of the new data set, the generalized variance is recalculated, accounting for the new observation that was just added. Once again, the observation with the lowest generalized variance is then added to the new data set. This process is repeated until all of the original data points are included in the new data set; i.e., the new data set is identical in terms of membership to the old one. However, now

each observation has associated with it a value for the generalized variance. Observations that are more distant from the bulk of the data will have larger values of the generalized variance. For $p = 2$ variables, observations with a generalized variance greater than

$$q_3 + 1.5(q_3 - q_1) \tag{5}$$

would be considered outliers, where q_1 and q_2 are the lower and upper quartiles, respectively, of the generalized variances. For more than two variables, the generalized variances are compared with

$$M_G + \sqrt{\chi^2_{0.975.p}}(q_3 - q_1), \tag{6}$$

where M_G is the median of the generalized variance values and $\chi^2_{0.975.p}$. **Figure 3** is a plot of X and Y, with outliers identified using the MGV approach. In this graph, outliers are denoted by 0, which is different than notation used in **Figures 1** and **2**. These graphs are included here exactly as taken from the R software output, which will be used extensively in the following examples.

PROJECTION-BASED OUTLIER DETECTION

Another alternative for identifying multivariate outliers is based on the notion of the depth of one data point among a set of other points. The idea of depth was described by Tukey (1975), and later expanded upon by Donoho and Gasko (1992). In general, depth can be thought of as the relative location of an observation vis-à-vis either edge (upper or lower) in a set of data. In the univariate case, this simply means determining to which edge a given observation more closely lies (i.e., maximum or minimum value), and then calculating the proportion of cases between that observation and its closest edge. The larger this proportion, the deeper the observation lies in the univariate data. While mathematically somewhat more challenging, conceptually projection methods of multivariate outlier detection work in much the same way. However, rather than determining the proximity to a single edge, the algorithm must identify the proximity to the edge of the multivariate space. This process is carried out using the method of projection that is described below.

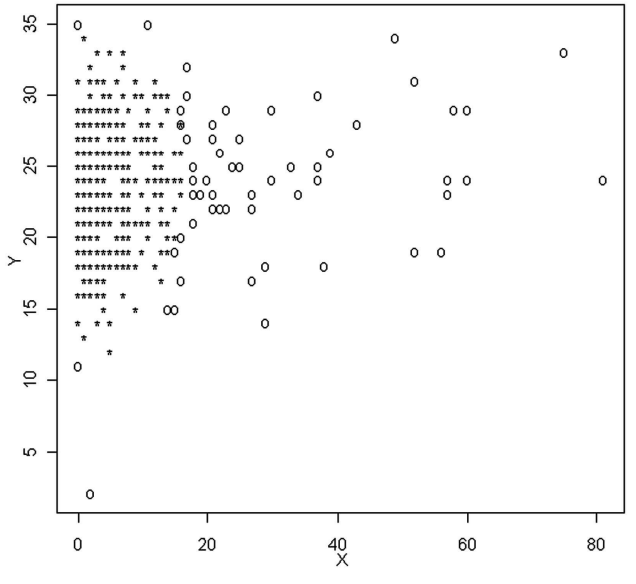


FIGURE 3 | Scatterplot of observations identified as outliers based on the MGV method.

For the purposes of this explanation, we will avoid presenting the mathematical equations that underlie the projection-based outlier detection approach. The interested reader is encouraged to refer to Wilcox (2005) for a more technical treatment of this methodology. Following is a conceptual description of two commonly used approaches to carrying out this technique when $p = 2$. In the first method, the algorithm begins by identifying the multivariate center of the data using an acceptable approach, such as the multivariate mean after application of MCD or MVE. Next, for each point ($\boldsymbol{X_i}$) the following steps are carried out:

(1) A line is drawn connecting the multivariate center and point $\boldsymbol{X_i}$.
(2) A line perpendicular to the line in 1 is then drawn from each of the other observations, $\boldsymbol{X_j}$.
(3) The location where the line in 2 intersects with the line in 1 is the projected depth (d_{ij}) of that data point for the line.
(4) Steps 1–3 are then repeated such that each of the n data points is connected to the multivariate center of the data with a line, and corresponding values of d_{ij} are calculated for each of the other observations.
(5) For a given observation, each of its depth values, d_{ij}, is compared with a standard (to be described below). If for any single projection the observation is an outlier, then it is classified as an outlier for purposes of future analyses.

As mentioned earlier, there is an alternative approach to the projection method, which is not based on finding the multivariate center of the distribution. Rather, all $(n^2 - n)/2$ possible lines are drawn between all pairs of observations in the dataset. Then, the approach outlined above for calculating d_{ij} is used for each of these lines. Thus, rather than having $n-1$ such d_{ij} values, each observation will have $\frac{(n^2-n)}{2} - 1$ indicators of depth. In all other ways, this second approach is identical to the first, however. Prior research has demonstrated that this second method might be more accurate than the first, but it is not clear how great an advantage it actually has in practice (Wilcox, 2005). Furthermore, because it must examine all possible lines in the set of data, method 2 can require quite a bit more computational time, particularly for large datasets.

The literature on multivariate outlier detection using the projection-based method includes two different criteria against which an observation can be judged as an outlier. The first of these is essentially identical to that used for the MGV in Eq. 6, with the exception that M_G is replaced by M_D, the median of the d_{ij} for that projection. Observations associated with values of d_{ij} larger than this cut score are considered to be outliers for that projection. An alternative comparison criterion is

$$M_D + \sqrt{\chi^2_{0.975.p}}\left(\frac{\mathrm{MAD_i}}{0.6745}\right) \tag{7}$$

where MAD_i is the median of all $|d_{ij} - M_D|$. Here, MAD_i is scaled by the divisor 0.6745 so that it approximates the standard deviation obtained when sampling from a normal distribution. Regardless of the criterion used, an observation is declared to be an outlier in general if it is an outlier for any of the projections.

GOALS OF THE CURRENT STUDY

The primary goal of this study was to describe alternatives to D^2 for multivariate outlier detection. As noted above, D^2 has a number of weaknesses in this regard, making it less than optimal for many research situations. Researchers need outlier detection methods on which they can depend, particularly given the sensitivity of many multivariate statistical techniques to the presence of outliers. Those described here may well fill that niche. A secondary goal of this study was to demonstrate, for a well known dataset, the impact of the various outlier detection methods on measures of location, variation and covariation. It is hoped that this study will serve as a useful reference for applied researchers working in the multivariate context who need to ascertain whether any observations are outliers, and if so which ones.

MATERIALS AND METHODS

The Women's Health and Drug study that is described in detail in Tabachnick and Fidell (2007) was used for demonstrative purposes. This dataset was selected because it appears in this very popular text in order to demonstrate data screening and as such was deemed an excellent source for demonstrating the methods studied here. A subset of the variables were used in the study, including number of visits to a health care provider (TIMEDRS), attitudes toward medication (ATTDRUG) and attitudes toward housework (ATTHOUSE). These variables were selected because they were featured in Tabachnick and Fidell's own analysis of the data. The sample used in this study consisted of 465 females aged 20–59 years who were randomly sampled from the San Fernando Valley in California and interviewed in 1976. Further description of the sample and the study from which it was drawn can be found in Tabachnick and Fidell.

In order to explore the impact of the various outlier detection methods included here, a variety of statistical analyses were conducted subsequent to the application of each approach. In particular, distributions of the three variables were examined for the

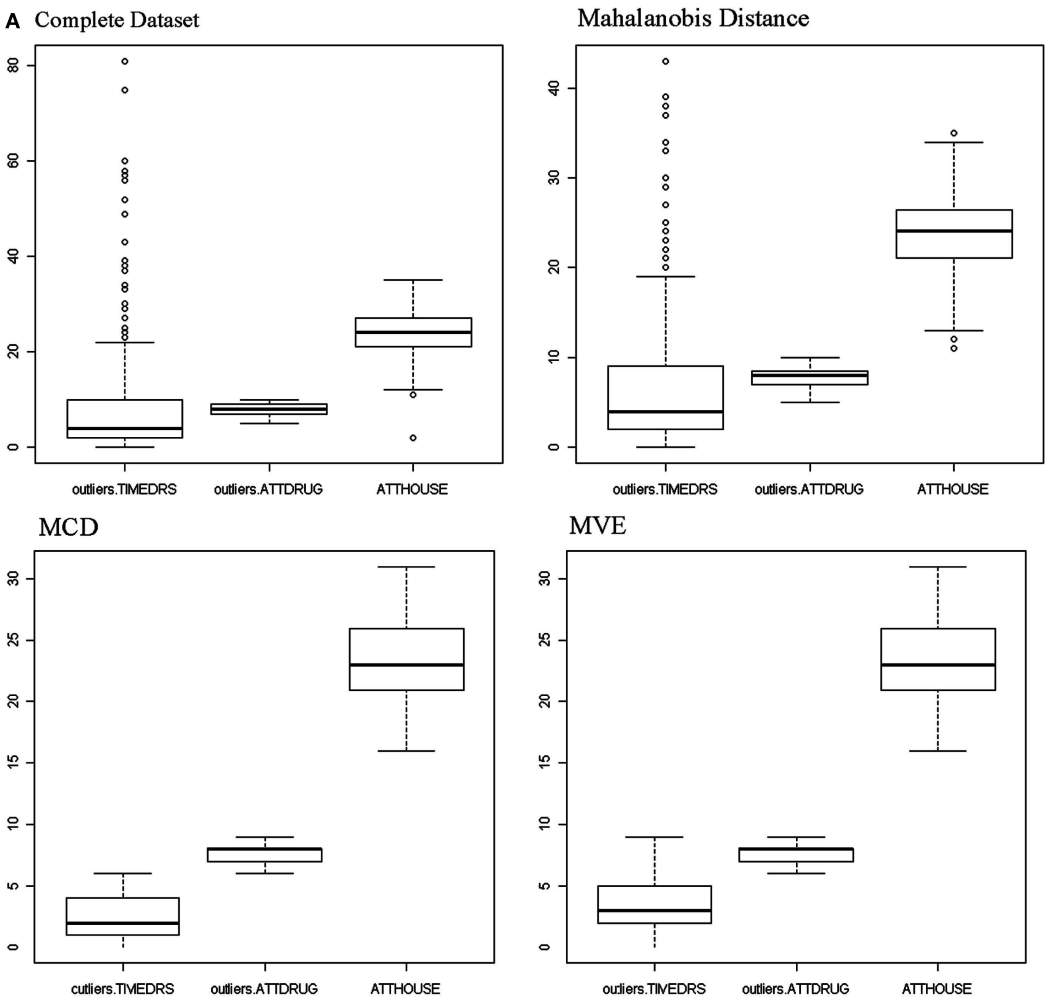


FIGURE 4 | Continued

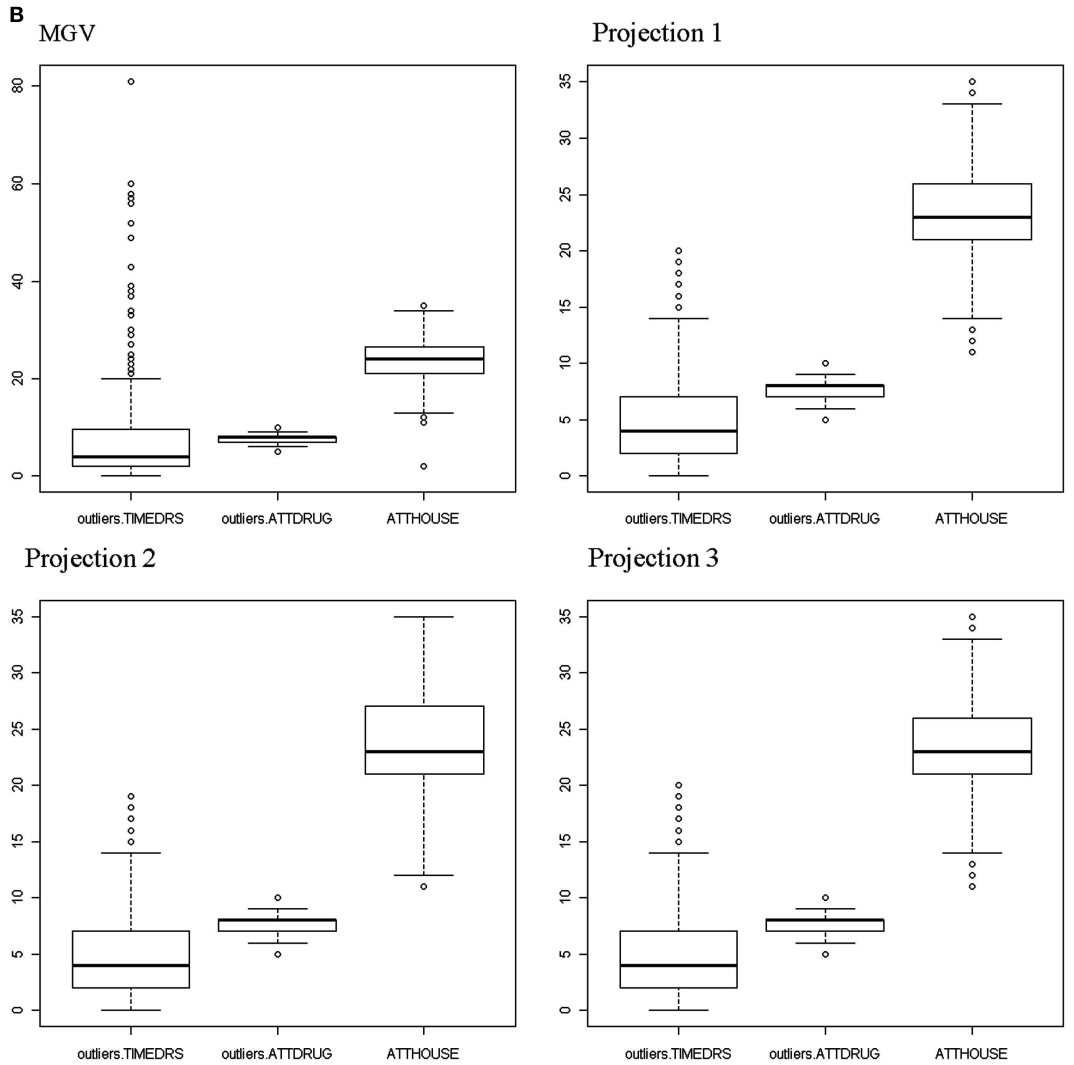


FIGURE 4 | Boxplots.

datasets created by the various outlier detection methods, as well as the full dataset. The strategy in this study was to remove all observations that were identified as outliers by each method, thus creating datasets for each approach that included only those not deemed to be outliers. It is important to note that this is not typically recommended practice, nor is it being suggested here. Rather, the purpose of this study was to demonstrate the impact of each method on the data itself. Therefore, rather than take the approach of examining each outlier carefully to ascertain whether it was truly part of the target population, the strategy was to remove those cases identified as outliers prior to conducting statistical analyses. In this way, it was hoped that the reader could clearly see the way in which each detection method worked and how this might impact resulting analyses. In terms of the actual data analysis, the focus was on describing the resulting datasets. Therefore, distributional characteristics of each variable within each method were calculated, including the mean, median, standard deviation, skewness, kurtosis, and first and third quartiles. In addition, distributions of the variables were examined using the boxplot. Finally, in order to demonstrate the impact of these approaches on relational measures, Pearson's correlation coefficient was estimated between each pair of variables. All statistical analyses including identification of outliers was carried out using the R software package, version 2.12.1 (R Foundation for Statistical Computing, 2010). The R code used to conduct these analyses appears in the Appendix at the end of the manuscript.

RESULTS

An initial examination of the full dataset using boxplots appears in **Figure 4**. It is clear that in particular the variable TIMEDRS is positively skewed with a number of fairly large values, even while the median is well under 10. Descriptive statistics for the full dataset (**Table 2**) show that indeed, all of the variables are fairly kurtotic, particularly TIMEDRS, which also displays a strong positive skew. Finally, the correlations among the three variables for the full dataset appear in **Table 3**. All of these

Table 2 | Descriptive statistics.

Variable	Mean							
	Full ($N = 465$)	**D^2 ($N = 452$)**	**MCD ($N = 235$)**	**MVE ($N = 235$)**	**MGV ($N = 463$)**	**P1 ($N = 425$)**	**P2 ($N = 422$)**	**P3 ($N = 425$)**
TIMEDRS	7.90	6.67	2.45	3.37	7.64	5.27	5.17	5.27
ATTDRUG	7.69	7.67	7.69	7.68	7.68	7.65	7.64	7.65
ATTHOUSE	23.53	23.56	23.36	23.57	23.50	23.54	23.54	23.54
MEDIAN								
TIMEDRS	4.00	4.00	2.00	3.00	2.00	2.00	2.00	2.00
ATTDRUG	8.00	8.00	8.00	8.00	7.00	7.00	7.00	7.00
ATTHOUSE	24.00	24.00	23.00	23.00	21.00	21.00	21.00	21.00
Q1								
TIMEDRS	2.00	2.00	1.00	2.00	2.00	2.00	2.00	2.00
ATTDRUG	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
ATTHOUSE	21.00	21.00	21.00	21.00	21.00	21.00	21.00	21.00
Q3								
TIMEDRS	10.00	9.00	4.00	5.00	9.50	7.00	7.00	7.00
ATTDRUG	9.00	8.25	8.00	8.00	8.00	8.00	8.00	8.00
ATTHOUSE	27.00	26.25	26.00	26.00	26.50	26.00	26.75	26.00
STANDARD DEVIATION								
TIMEDRS	10.95	7.35	1.59	2.22	10.23	4.72	4.58	4.72
ATTDRUG	1.16	1.16	0.89	0.81	1.15	1.56	1.52	1.56
ATTHOUSE	4.48	4.22	3.67	3.40	4.46	4.25	4.26	4.25
SKEWNESS								
TIMEDRS	3.23	2.07	0.23	0.46	3.15	1.12	1.08	1.12
ATTDRUG	−0.12	−0.11	−0.16	0.01	−0.12	−0.09	−0.09	−0.09
ATTHOUSE	−0.45	−0.06	−0.03	0.06	−0.46	−0.03	−0.03	−0.03
KURTOSIS								
TIMEDRS	15.88	7.92	2.32	2.49	15.77	3.42	3.29	3.42
ATTDRUG	2.53	2.51	2.43	2.31	2.54	2.51	2.51	2.51
ATTHOUSE	4.50	2.71	2.16	2.17	4.54	2.69	2.68	2.69

are below 0.15, indicating fairly weak relationships among the measures. However, it is not clear to what extent these correlations may be impacted by the distributional characteristics just described.

Given these distributional issues, the researcher working with this dataset would be well advised to investigate the possibility that outliers are present. For this example, we can use R to calculate D^2 for each observation, with appropriate program code appearing in the Appendix. In order to identify an observation as an outlier, we compare the D^2 value to the chi-square distribution with degrees of freedom equal to the number of variables (three in this case), and $\alpha = 0.001$. Using this criterion, 13 individuals were identified as outliers. In order to demonstrate the impact of using D^2 for outlier detection, these individuals were removed, and descriptive graphics and statistics were generated for the remaining 452 observations. An examination of the boxplot for the Mahalanobis data reveals that the range of values is more truncated than for the original, particularly for TIMEDRS. A similar result is evident in the descriptive statistics found in **Table 1**, where we can see that the standard deviation, skewness, kurtosis and mean are all smaller in the Mahalanobis data for TIMEDRS. In contrast, the removal of the 13 outliers identified by D^2 did not result in great changes for the distributional characteristics of ATTDRUG or ATTHOUSE. The correlations among the three variables were comparable to those in the full dataset, if not slightly smaller.

As discussed previously, there are some potential problems with using D^2 as a method for outlier detection. For this reason, other approaches have been suggested for use in the context of multivariate data in particular. A number of these, including MCD, MVE, MGV, and three projection methods, were applied to this dataset, followed by generation of graphs and descriptive statistics as was done for both the full dataset and the Mahalanobis data. Boxplots for the three variables after outliers were identified and removed by each of the methods appear in **Figure 4**. As can be seen, the MCD and MVE approaches resulted in data that appear to be the least skewed, particularly for TIMEDRS. In contrast, the data from MGV was very similar to that of the full dataset, while the three projection methods resulted in data that appeared to lie between MCD/MVE and the Mahalanobis data in terms of the distributions of the three variables. An examination of **Table 1** confirms that making use of the different outlier detection methods results in datasets with markedly different distributional characteristics. Of particular note are differences between MCD/MVE as compared to the full dataset, and the Mahalanobis and MGV data. Specifically, the skewness and kurtosis evident in these two samples was markedly lower than that of any of the datasets, particularly the full

Table 3 | Correlations.

Variable	TIMEDRS	ATTDRUG	ATTHOUSE
FULL DATA SET ($N=465$)			
TIMEDRS	1.00	0.10	0.13
ATTDRUG	0.10	1.00	0.03
ATTHOUSE	0.13	0.03	1.00
MAHALANOBIS DISTANCE ($N=452$)			
TIMEDRS	1.00	0.07	0.08
ATTDRUG	0.07	1.00	0.02
ATTHOUSE	0.08	0.02	1.00
MCD ($N=235$)			
TIMEDRS	1.00	0.25	0.19
ATTDRUG	0.25	1.00	0.26
ATTHOUSE	0.19	0.26	1.00
MVE ($N=235$)			
TIMEDRS	1.00	0.33	0.05
ATTDRUG	0.33	1.00	0.32
ATTHOUSE	0.05	0.32	1.00
MGV ($N=463$)			
TIMEDRS	1.00	0.07	0.10
ATTDRUG	0.07	1.00	0.02
ATTHOUSE	0.10	0.02	1.00
PROJECTION 1 ($N=425$)			
TIMEDRS	1.00	0.06	0.10
ATTDRUG	0.06	1.00	0.03
ATTHOUSE	0.10	0.03	1.00
PROJECTION 2 ($N=422$)			
TIMEDRS	1.00	0.04	0.11
ATTDRUG	0.04	1.00	0.03
ATTHOUSE	0.11	0.03	1.00
PROJECTION 3 ($N=425$)			
TIMEDRS	1.00	0.06	0.10
ATTDRUG	0.06	1.00	0.03
ATTHOUSE	0.10	0.03	1.00

and MGV datasets. In addition, probably as a result of removing a number of individuals with large values, the mean of TIMEDRS was substantially lower for the MCD and MVE data than for the full, Mahalanobis, and MGV datasets. As noted with the boxplot, the three projection methods produced means, skewness, and kurtosis values that generally fell between those of MCD/MVE and the other approaches. Also of interest in this regard is the relative similarity in distributional characteristics of the ATTDRUG variable across outlier detection methods. This would suggest that there were few if any outliers present in the data for this variable. Finally, the full and MGV datasets had kurtosis values that were somewhat larger than those of the other methods included here.

Finally, in order to ascertain how the various outlier detection methods impacted relationships among the variables we estimated correlations for each approach, with results appearing in **Table 3**. For the full data, correlations among the three variables were all low, with the largest being 0.13. When outliers were detected and removed using D^2, MGV and the three projection methods, the correlations were attenuated even more. In contrast, correlations calculated using the MCD data were uniformly larger than those of any method, except for the value between TIMEDRS and ATTDRUG, which was larger for the MVE data. On the other hand, the correlation between TIMEDRS and ATTHOUSE was smaller for MVE than for any of the other methods used here.

DISCUSSION

The purpose of this study was to demonstrate seven methods of outlier detection designed especially for multivariate data. These methods were compared based upon distributions of individual variables, and relationships among them. The strategy involved first identification of outlying observations followed by their removal prior to data analysis. A brief summary of results for each methodology appears in **Table 4**. These results were markedly different across methods, based upon both distributional and correlational measures. Specifically, the MGV and Mahalanobis distance approaches resulted in data that was fairly similar to the full data. In contrast, MCD and MVE both created datasets that were very different, with variables more closely adhering to the normal distribution, and with generally (though not universally) larger relationships among the variables. It is important to note that these latter two methods each removed 230 observations, or nearly half of the data, which may be of some concern in particular applications and which will be discussed in more detail below. Indeed, this issue is not one to be taken lightly. While the MCD/MVE approaches produced datasets that were more in keeping with standard assumptions underlying many statistical procedures (i.e., normality), the representativeness of the sample may be called into question. Therefore, it is important that researchers making use of either of these methods closely investigate whether the resulting sample resembles the population. Indeed, it is possible that identification of such a large proportion of the sample as outliers is really more of an indication that the population is actually bimodal. Such major differences in performance depending upon the methodology used point to the need for researchers to be familiar with the panoply of outlier detection approaches available. This work provides a demonstration of the methods, comparison of their relative performance with a well known dataset, and computer software code so that the reader can use the methods with their own data.

There is not one universally optimal approach for identifying outliers, as each research problem presents the data analyst with specific challenges and questions that might be best addressed using a method that is not optimal in another scenario. This study helps researchers and data analysts to see the range of possibilities available to them when they must address outliers in their data. In addition, these results illuminate the impact of using the various methods for a representative dataset, while the R code in the Appendix provides the researcher with the software tools necessary to use each technique. A major issue that researchers must consider is the tradeoff between a method with a high breakdown point (i.e., that is impervious to the presence of many outliers) and the desire to retain as much of the data as possible. From this example, it is clear that the methods with the highest breakdown points, MCD/MVE, retained data that more clearly conformed to the normal distribution than did the other approaches, but at the cost of approximately half of the original data. Thus, researchers must consider the purpose of their efforts to detect outliers. If they

Table 4 | Summary of results for outlier detection methods.

Method	Outliers removed	Impact on distributions	Impact on correlations	Comments
D_i^2	13	Reduced skewness and kurtosis when compared to full data set, but did not fully eliminate them. Reduced variation in TIMEDRS	Comparable correlations to the full dataset	Resulted in a sample with somewhat less skewed and kurtotic variables, though they did remain clearly non-normal in nature. The correlations among the variables remained low, as with the full dataset
MVE	230	Largely eliminated skewness and greatly lowered kurtosis in TIMEDRS. Also reduced kurtosis in ATTHOUSE when compared to full data. Greatly lowered both the mean and standard deviation of TIMEDRS	Resulted in markedly higher correlations for two pairs of variables, than was seen with the other methods, except for MCD	Reduced the sample size substantially, but also yielded variables with distributional characteristics much more favorable to use with common statistical analyses; i.e., very little skewness or kurtosis. In addition, correlation coefficients were generally larger than for the other methods, suggesting greater linearity in relationships among the variables
MCD	230	Very similar pattern to that displayed by MVE	Yielded relatively higher correlation values than any of the other methods, except MVE, and no very low values	Provided a sample with very characteristics to that of MVE
MGV	2	Yielded distributional results very similar to those of the full dataset	Very similar correlation structure as found in the full dataset and for D^2	Identified very few outliers, leading to a sample that did not differ meaningfully from the original
P1	40	Resulted in lower mean, standard deviation, skewness, and kurtosis values for TIMEDRS when compared to the full data, D^2, and MGV, though not when compared to MVE and MCD. Yielded comparable skewness and kurtosis to other methods for ATTDRUG and ATTHOUSE, and somewhat greater variation for these other variables, as well	Very comparable correlation results to the full dataset, as well as D^2 and MGV	Appears to find a "middle ground" between MVE/MCD and D^2/MGV in terms of the number of outliers identified and the resulting impact on variable distributions and correlations
P2	43	Very similar results to P1	Very similar results to P1	Provided a sample yielding essentially the same results as P1
P3	40	Identical results to P1	Identical results to P1	In this case, resulted in an identical sample to that of P1

are seeking a "clean" set of data upon which they can run a variety of analyses with little or no fear of outliers having an impact, then methods with a high breakdown point, such as MCD and MVE are optimal. On the other hand, if an examination of outliers reveals that they are from the population of interest, then a more careful approach to dealing with them is necessary. Removing such outliers could result in a dataset that is more tractable with respect to commonly used parametric statistical analyses but less representative of the general population than is desired. Of course, the converse is also true in that a dataset replete with outliers might produce statistical results that are not generalizable to the population of real interest, when the outlying observations are not part of this population.

FUTURE RESEARCH

There are a number of potential areas for future research in the area of outlier detection. Certainly, future work should use these methods with other extant datasets having different characteristics than the one featured here. For example, the current set of data consisted of only three variables. It would be interesting to compare the relative performance of these methods when more variables are present. Similarly, examining them for much smaller groups would also be useful, as the current sample is fairly large when compared to many that appear in social science research. In addition, a simulation study comparing these methods with one another would also be warranted. Specifically, such a study could be based upon the generation of datasets with known outliers and distributional characteristics of the non-outlying cases. The various detection methods could then be used and the resulting retained datasets compared to the known non-outliers in terms of these various characteristics. Such a study would be quite useful in informing researchers regarding approaches that might be optimal in practice.

CONCLUSION

This study should prove helpful to those faced with a multivariate outlier problem in their data. Several methods of outlier detection were demonstrated and great differences among them were observed, in terms of the characteristics of the observations retained. These findings make it clear that researchers must be very thoughtful in their treatment of outlying observations. Simply relying on Mahalanobis Distance because it is widely used

might well yield statistical results that continue to be influenced by the presence of outliers. Thus, other methods described here should be considered as viable options when multivariate outliers are present. In the final analysis, such an approach must be based on the goals of the data analysis and the study as a whole. The removal of outliers, when done, must be carried out thoughtfully and with purpose so that the resulting dataset is both representative of the population of interest and useful with the appropriate statistical tools to address the research questions.

REFERENCES

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research.* New York: The Guilford Press.

Donoho, D. L., and Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. *Ann. Stat.* 20, 1803–1827.

Evans, V. P. (1999). "Strategies for detecting outliers in regression analysis: an introductory primer," in *Advances in Social Science Methodology*, ed. B. Thompson (Stamford, CT: JAI Press), 271–286.

Genton, M. G., and Lucas, A. (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 65, 81–94.

Hardin, J., and Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Comput. Stat. Data Anal.* 44, 625–638.

Huberty, C. J., and Olejnik, S. (2006). *Applied MANOVA and Discriminant Analysis.* Hoboken, NJ: John Wiley & Sons, Inc.

Johnson, R. A., and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis.* New York: Prentice Hall.

Kaufman, L., and Rousseeuw, P. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis.* Hoboken, NJ: John Wiley & Sons, Inc.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences.* Pacific Grove, CA: Brooks/Cole.

Kruskal, W. (1988). Miracles and statistics: the causal assumption of independence. *J. Am. Stat. Assoc.* 83, 929–940.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Indian Natl. Sci. Acad. B Biol. Sci.* 2, 49–55.

Marascuilo, L. A., and Serlin, R. C. (1988). *Statistical Methods for the Social and Behavioral Sciences.* New York: W. H. Freeman.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods.* Hoboken, NJ: John Wiley & Sons, Inc.

Mourão-Miranda, J., Hardoon, D. R., Hahn, T., Marquand, A. F., Williams, S. C. R., Shawe-Taylor, J., and Brammer, M. (2011). Patient classification as an outlier detection problem: an application of the one-class support vector machine. *Neuroimage* 58, 793–804.

Osborne, J. W., and Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Pract. Assess. Res. Eval.* 9. Available at: http://PAREonline.net/getvn.asp?v=9&n=6

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction.* Orlando, FL: Harcourt Brace College Publishers.

R Foundation for Statistical Computing. (2010). *R Software, Version 2.12.1*, Vienna: The R Foundation.

Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

Rousseeuw, P. J., and van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41, 212–223.

Stevens, J. P. (2009). *Applied Multivariate Statistics for the Social Sciences.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Tabachnick, B. G., and Fidell, L. S. (2007). *Using Multivariate Statistics.* Boston: Pearson Education, Inc.

Tukey, J. W. (1975). "Mathematics and the picturing of data," in *Proceeding of the International Congress of Mathemeticians*, Vol. 2, 523–531.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing.* Burlington, MA: Elsevier Academic Press.

Wilcox, R. R. (2010). *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy.* New York: Springer.

Zijlstra, W. P., van der Ark, L. A., and Sijtsmal, K. (2007). Robust Mokken scale analysis by means of a forward search algorithm for outlier detection. *Multivariate Behav. Res.* 46, 58–89.

Zijlstra, W. P., van der Ark, L. A., and Sijtsmal, K. (2011). Outliers in questionnaire data: can they be detected and should they be removed? *J. Educ. Behav. Stat.* 36, 186–212.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 December 2011; paper pending published: 17 January 2012; accepted: 06 June 2012; published online: 05 July 2012.

*Citation: Finch WH (2012) Distribution of variables by method of outlier detection. Front. Psychology **3**:211. doi: 10.3389/fpsyg.2012.00211*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

APPENDIX

```
library(MASS)
mahalanobis.out < -mahalanobis(full.data,colMeans(full.data),
cov(full.data))
mcd.output < -cov.rob(full.data,method = "mcd," nsamp =
"best")
mcd.keep < -full.data[mcd.output$best,]
mve.output < -cov.rob(full.data,method = "mve," nsamp =
"best")
mve.keep < -full.data[mve.output$best,]
mgv.output < -outmgv(full.data,y = NA,outfun = outbox)
mgv.keep < -full.data[mgv.output$keep,]
projection1.output < -outpro(full.data,cop = 2)
projection1.keep < -full.data[projection1.output$keep,]
projection2.output < -outpro(full.data.matrix,cop = 3)
projection2.keep < -full.data[projection2.output$keep,]
projection3.output < -outpro(full.data,cop = 4)
projection3.keep < -full.data[projection3.output$keep,]
```

*Note that for the projection methods, the center of the distribution may be determined using one of four possible approaches. The choice of method is specified in the **cop** command. For this study, three of these were used, including MCD (**cop = 2**), median of the marginal distributions (**cop = 3**), and MVE (**cop = 4**).

The functions for obtaining the mahalanobis distance (mahalanobis**) and MCD/MVE (**cov.rob**) are part of the **MASS** library in R, and will be loaded when it is called. The **outmgv** and **outpro** functions are part of a suite of functions written by Rand Wilcox with information for obtaining them available at his website (http://dornsife.usc.edu/cf/faculty-and-staff/faculty.cfm?pid = 1003819&CFID = 259154&CFTOKEN = 86756889) and his book, *Introduction to Robust Estimation and Hypothesis Testing*.

frontiers in
PSYCHOLOGY

MINI REVIEW ARTICLE
published: 28 August 2012
doi: 10.3389/fpsyg.2012.00322

Statistical assumptions of substantive analyses across the general linear model: a mini-review

Kim F. Nimon*

Learning Technologies, University North Texas, Denton, TX, USA

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Anne C. Black, Yale University School of Medicine, USA
Megan Welsh, University of Connecticut, USA
Cherng-Jyh Yen, Old Dominion University, USA

****Correspondence:***
Kim F. Nimon, Learning Technologies, University North Texas, 3940 North Elm Street, G150 Denton, 76207 TX, USA.
e-mail: kim.nimon@gmail.com

The validity of inferences drawn from statistical test results depends on how well data meet associated assumptions. Yet, research (e.g., Hoekstra et al., 2012) indicates that such assumptions are rarely reported in literature and that some researchers might be unfamiliar with the techniques and remedies that are pertinent to the statistical tests they conduct. This article seeks to support researchers by concisely reviewing key statistical assumptions associated with substantive statistical tests across the general linear model. Additionally, the article reviews techniques to check for statistical assumptions and identifies remedies and problems if data do not meet the necessary assumptions.

Keywords: assumptions, robustness, analyzing data, normality, homogeneity

The degree to which valid inferences may be drawn from the results of inferential statistics depends upon the sampling technique and the characteristics of population data. This dependency stems from the fact that statistical analyses assume that sample(s) and population(s) meet certain conditions. These conditions are called statistical assumptions. If violations of statistical assumptions are not appropriately addressed, results may be interpreted incorrectly. In particular, when statistical assumptions are violated, the probability of a test statistic may be inaccurate, distorting Type I or Type II error rates.

This article focuses on the assumptions associated with substantive statistical analyses across the general linear model (GLM), as research indicates they are reported with more frequency in educational and psychological research than analyses focusing on measurement (cf. Kieffer et al., 2001; Zientek et al., 2008). This review is organized around **Table 1**, which relates key statistical assumptions to associated analyses and classifies them into the following categories: randomization, independence, measurement, normality, linearity, and variance. Note that the assumptions of independence, measurement, normality, linearity, and variance apply to population data and are tested by examining sample data and using test statistics to draw inferences about the population(s) from which the sample(s) were selected.

RANDOMIZATION

A basic statistical assumption across the GLM is that sample data are drawn randomly from the population. However, much social science research is based on unrepresentative samples (Thompson, 2006) and many quantitative researchers select a sample that suits the purpose of the study and that is convenient (Gall et al., 2007). When the assumption of random sampling is not met, inferences to the population become difficult. In this case, researchers should describe the sample and population in sufficient detail to justify that the sample was at least representative of the intended population (Wilkinson and APA Task Force on Statistical Inference, 1999). If such a justification is not made, readers are left to their own interpretation as to the generalizability of the results.

INDEPENDENCE

Across GLM analyses, it is assumed that observations are independent of each other. In quantitative research, data often do not meet the independence assumption. The simplest case of non-independent data is paired sample or repeated measures data. In these cases, only pairs of observations (or sets of repeated data) can be independent because the structure of the data is by design paired (or repeated). More complex data structures that do not meet the assumption of independence include nested data (e.g., employees within teams and teams within departments) and cross-classified data (e.g., students within schools and neighborhoods).

When data do not meet the assumption of independence, the accuracy of the test statistics (e.g., t, F, χ^2) resulting from a GLM analysis depends on the test conducted. For data that is paired (e.g., pretest-posttest, parent-child), paired samples t test is an appropriate statistical analysis as long as the pairs of observations are independent and all other statistical assumptions (see **Table 1**) are met. Similarly, for repeated measures data, repeated measures ANOVA is an appropriate statistical analysis as long as sets of repeated measures data are independent and all other statistical assumptions (see **Table 1**) are met. For repeated measures and/or non-repeated measures data that are nested or cross-classified, multilevel modeling (MLM) is an appropriate statistical analytic strategy because it models non-independence. Statistical tests that do not model the nested or cross-classified structure of data will lead to a higher probability of rejecting the null hypotheses (i.e.,

Table 1 | Statistical assumptions associated with substantive analyses across the general linear model.

Statistical analysis[a]	Independence	Measurement Level of variable(s)[b]		Normality	Linearity	Variance
		Dependent	Independent			
CHI-SQUARE						
Single sample	Independent observations	Nominal+	N/A			
2+ samples	Independent observations	Nominal+	Nominal+			
***T*-TEST**						
Single sample	Independent observations	Continuous	N/A	Univariate		
Dependent sample	Independent paired observations	Continuous	N/A	Univariate		
Independent sample	Independent observations	Continuous	Dichotomous	Univariate		Homogeneity of variance
OVA-RELATED TESTS						
ANOVA	Independent observations	Continuous	Nominal	Univariate		Homogeneity of variance
ANCOVA[c]	Independent observations	Continuous	Nominal	Univariate	✓	Homogeneity of variance
RM ANOVA	Independent repeated observations	Continuous	Nominal (opt.)	Multivariate	✓	Sphericity
MANOVA	Independent observations	Continuous	Nominal	Multivariate	✓	Homogeneity of covariance matrix
MANCOVA[c]	Independent observations	Continuous	Nominal	Multivariate	✓	Homogeneity of covariance matrix
REGRESSION						
Simple linear	Independent observations	Continuous	Continuous	Bivariate	✓	
Multiple linear	Independent observations	Continuous	Continuous	Multivariate	✓	Homoscedasticity
Canonical correlation	Independent observations	Continuous	Continuous	Multivariate	✓	Homoscedasticity

[a]Across all analyses, data are assumed to be randomly sampled from the population. [b]Data are assumed to be reliable. [c]ANCOVA and MANCOVA also assumes homogeneity of regression and continuous covariate(s). Continuous refers to data that may be dichotomous, ordinal, interval, or ratio (cf. Tabachnick and Fidell, 2001).

Type I error) than if an appropriate statistical analysis were performed or if the data did not violate the independence assumption (Osborne, 2000).

Presuming that statistical assumptions can be met, MLM can be used to conduct statistical analyses equivalent to those listed in **Table 1** (see Garson, 2012 for a guide to a variety of MLM analyses). Because multilevel models are generalizations of multiple regression models (Kreft and de Leeuw, 1998), MLM analyses have assumptions similar to analyses that do not model multilevel data. When violated, distortions in Type I and Type II error rates are imminent (Onwuegbuzie and Daniel, 2003).

MEASUREMENT

RELIABILITY

Another basic assumption across the GLM is that population data are measured without error. However, in psychological and educational research, many variables are difficult to measure and yield observed data with imperfect reliabilities (Osborne and Waters, 2002). Unreliable data stems from systematic and random errors of measurement where systematic errors of measurement are "those which consistently affect an individual's score because of some particular characteristic of the person or the test that has nothing to do with the construct being measured (Crocker and Algina, 1986, p. 105) and random errors of measurement are those which "affect an individual's score because of purely chance happenings" (Crocker and Algina, 1986, p. 106).

Statistical analyses on unreliable data may cause effects to be underestimated which increase the risk of Type II errors (Onwuegbuzie and Daniel, 2003). Alternatively in the presence of correlated error, unreliable data may cause effects to be overestimated which increase the risk of Type I errors (Nimon et al., 2012).

To satisfy the assumption of error-free data, researchers may conduct and report analyses based on latent variables in lieu of observed variables. Such analyses are based on a technique called structural equation modeling (SEM). In SEM, latent variables are formed from item scores, the former of which become the unit of analyses (see Schumacker and Lomax, 2004 for an accessible introduction). Analyses based on latent-scale scores yield statistics as if multiple-item scale scores had been measured without error. All of the analyses in **Table 1** as well as MLM analyses can be conducted with SEM. The remaining statistical assumptions apply when latent-scale scores are analyzed through SEM.

Since SEM is a large sample technique (see Kline, 2005), researchers may alternatively choose to delete one or two items in order to raise the reliability of an observed score. Although "extensive revisions to prior scale dimensionality are questionable. . . one or a few items may well be deleted" in order to increase reliability (Dillon and Bearden, 2001, p. 69). The process of item deletion should be reported, accompanied by estimates of the reliability of the data with and without the deleted items (Nimon et al., 2012).

MEASUREMENT LEVEL

Table 1 denotes measurement level as a statistical assumption. Whether level of measurement is considered a statistical assumption is a point of debate in statistical literature. For example, proponents of Stevens (1946, 1951) argue that the dependent variable in parametric tests such as *t* tests and analysis-of-variance related tests should be scaled at the interval or ratio level (Maxwell

and Delaney, 2004). Others (e.g., Howell, 1992; Harris, 2001) indicate that the validity of statistical conclusions depends only on whether data meet distributional assumptions not on the scaling procedures used to obtain data (Warner, 2008). Because measurement level plays a pivotal role in statistical analyses decision trees (e.g., Tabachnick and Fidell, 2001, pp. 27–29), **Table 1** relates measurement level to statistical analyses from a pragmatic perspective. It is important to note that lowering the measurement level of data (e.g., dichotomizing intervally scaled data) is ill-advised unless data meet certain characteristics (e.g., multiple modes, serious skewness; Kerlinger, 1986). Although such a transformation makes data congruent with statistics that assume only the nominal measurement level, it discards important information and may produce misleading or erroneous information (see Thompson, 1988).

NORMALITY

For many inferential statistics reported in educational and psychological research (cf. Kieffer et al., 2001; Zientek et al., 2008), there is an assumption that population data are normally distributed. The requirement for, type of, and loci of normality assumption depend on the analysis conducted. Univariate group comparison tests (*t* tests, ANOVA, ANCOVA) assume univariate normality (Warner, 2008). Simple linear regression assumes bivariate normality (Warner, 2008). Multivariate analyses (repeated measures ANOVA, MANOVA, MANCOVA, multiple linear regression, and canonical correlation) assume multivariate normality (cf. Tabachnick and Fidell, 2001; Stevens, 2002). For analysis-of-variance type tests (OVA-type tests) involving multiple samples, the normality assumption applies to each level of the IV.

UNIVARIATE

The assumption of univariate normality is met when a distribution of scores is symmetrical and when there is an appropriate proportion of distributional height to width (Thompson, 2006). To assess univariate normality, researchers may conduct graphical or non-graphical tests (Stevens, 2002): Non-graphical tests include the chi-square goodness of fit test, the Kolmogorov–Smirnov test, the Shapiro–Wilks test, and the evaluation of kurtosis and skewness values. Graphical tests include the normality probability plot and the histogram (or stem-and-leave plot).

Non-graphical tests are preferred for small to moderate sample sizes, with the Shapiro–Wilks test and the evaluation of kurtosis and skewness values being preferred methods for sample sizes of less than 20 (Stevens, 2002). The normal probability plot in which observations are ordered in increasing degrees of magnitude and then plotted against expected normal distribution values is preferred over histograms (or stem-and leave plots). Evaluating normality by examining the shape of histogram scan be problematic (Thompson, 2006), because there are *infinitely* different distribution shapes that may be normal (Bump, 1991). The bell-shaped distribution that many educational professionals are familiar with is not the only normal distribution (Henson, 1999).

BIVARIATE

The assumption of bivariate normality is met when the linear relationship between two variables follows a normal distribution (Burdenski, 2000). A necessary but insufficient condition for bivariate normality is univariate normality for each variable. Bivariate normality can be evaluated graphically (e.g., scatterplots). However, in practice, even large datasets ($n > 200$) have insufficient data points to evaluate bivariate normality (Warner, 2008) which may explain why this assumption often goes unchecked and unreported.

MULTIVARIATE

The assumption of multivariate normality is met when each variable in a set is normally distributed around fixed values on all other variables in the set (Henson, 1999). Necessary but insufficient conditions for multivariate normality include univariate normality for each variable along with bivariate normality for each variable pair. Multivariate normality can be assessed graphically or with statistical tests.

To assess multivariate normality graphically, a scatterplot of Mahalanobis distances and paired χ^2-values may be examined, where Mahalanobis distance indicates how far each "set of scores is from the group means adjusting for correlation of the variables" (Burdenski, 2000, p. 20). If the plot approximates a straight-line, data are considered multivariate normal. Software to produce the Mahalanobis distance by χ^2 scatterplot can be found in Thompson (1990); Henson (1999), and Fan (1996).

Researchers may also assess multivariate normality by testing Mardia's (1985) coefficient of multivariate kurtosis and examining its critical ratio. If the critical ratio of Mardia's coefficient of multivariate kurtosis is less than 1.96, a sample can be considered multivariate normal at the 0.05 significance level, indicating that the multivariate kurtosis is not statistically significantly different than zero. Mardia's coefficient of multivariate kurtosis is available in statistical software packages including AMOS, EQS, LISREL, and PASW (see DeCarlo, 1997).

VIOLATIONS

The effect of violating the assumption of normality depends on the level of non-normality and the statistical test examined. As long the assumption of normality is not severely violated, the actual Type I error rates approximate nominal rates for *t* tests and OVA-tests (cf. Boneau, 1960; Glass et al., 1972; Stevens, 2002). However, in the case of data that are severely platykurtic, power is reduced in *t* tests and OVA-type tests (cf. Boneau, 1960; Glass et al., 1972; Stevens, 2002). Non-normal variables that are highly skewed or kurtotic distort relationships and significance tests in linear regression (Osborne and Waters, 2002). Similarly, proper inferences regarding statistical significance tests in canonical correlation depend on multivariate normality (Tabachnick and Fidell, 2001). If the normality assumption is violated, researchers may delete outlying cases, transform data, or conduct non-parametric tests (see Conover, 1999; Osborne, 2012), as long as the process is clearly reported.

LINEARITY

For parametric statistics involving two or more continuous variables (ANCOVA, repeated measures ANOVA, MANOVA, MANCOVA, linear regression, and canonical correlation) linearity between pairs of continuous variables is assumed (cf. Tabachnick and Fidell, 2001; Warner, 2008). The assumption of linearity is that there is a straight-line relationship between two variables.

Linearity is important in a practical sense because Pearson's *r*, which is fundamental to the vast majority of parametric statistical procedures (Graham, 2008), captures *only* the linear relationship among variables (Tabachnick and Fidell, 2001). Pearson's *r* underestimates the true relationship between two variables that is non-linear (i.e., curvilinear; Warner, 2008).

Unless there is strong theory specifying non-linear relationships, researchers may assume linear relationships in their data (Cohen et al., 2003). However, linearity is not guaranteed and should be validated with graphical methods (see Tabachnick and Fidell, 2001). Non-linearity reduces the power of statistical tests such as ANCOVA, MANOVA, MANCOVA, linear regression, and canonical correlation (Tabachnick and Fidell, 2001). In the case of ANCOVA and MANCOVA, non-linearity results in improper adjusted means (Stevens, 2002). If non-linearity is detected, researchers may transform data, incorporate curvilinear components, eliminate the variable producing non-linearity, or conduct a non-linear analysis (cf. Tabachnick and Fidell, 2001; Osborne and Waters, 2002; Stevens, 2002; Osborne, 2012), as long as the process is clearly reported.

VARIANCE

Across parametric statistical procedures commonly used in quantitative research, at least five assumptions relate to variance. These are: homogeneity of variance, homogeneity of regression, sphericity, homoscedasticity, and homogeneity of variance-covariance matrix.

Homogeneity of variance applies to univariate group analyses (independent samples *t* test, ANOVA, ANCOVA) and assumes that the variance of the DV is roughly the same at all levels of the IV (Warner, 2008). The Levene's test validates this assumption, where smaller statistics indicate greater homogeneity. Research (Boneau, 1960; Glass et al., 1972) indicates that univariate group analyses are generally robust to moderate violations of homogeneity of variance as long as the sample sizes in each group are approximately equal. However, with unequal sample sizes, heterogeneity may compromise the validity of null hypothesis decisions. Large sample variances from small-group sizes increase the risk of Type I error. Large sample variances from large-group sizes increase the risk of Type II error. When the assumption of homogeneity of variance is violated, researchers may conduct and report non-parametric tests such as the Kruskal–Wallis. However, Maxwell and Delaney (2004) noted that the Kruskal–Wallis test also assumes equal variances and suggested that data be either transformed to meet the assumption of homogeneity of variance or analyzed with tests such as Brown–Forsythe F* or Welch's W.

Homogeneity of regression applies to group analyses with covariates, including ANCOVA and MANCOVA, and assumes that the regression between covariate(s) and DV(s) in one group is the same as the regression in other groups (Tabachnick and Fidell, 2001). This assumption can be examined graphically or by conducting a statistical test on the interaction between the COV(s) and the IV(s). Violation of this assumption can lead to very misleading results if covariance is used (Stevens, 2002). For example, in the case of heterogeneous slopes, group means that have been adjusted by a covariate could indicate no difference when, in fact, group differences might exist at different values of the covariate. If heterogeneity of regression exists, ANCOVA and MANCOVA are inappropriate analytic strategies (Tabachnick and Fidell, 2001).

Sphericity applies to repeated measures analyses that involve three or more measurement occasions (repeated measures ANOVA) and assumes that the variances of the differences for all pairs of repeated measures are equal (Stevens, 2002). Presuming that data are multivariate normal, the Mauchly test can be used to test this assumption, where smaller statistics indicate greater levels of sphericity (Tabachnick and Fidell, 2001). Violating the sphericity assumption increases the risk of Type I error (Box, 1954). To adjust for this risk and provide better control for Type I error rate, the degrees of freedom for the repeated measures *F* test may be corrected using and reporting one of three adjustments: (a) Greenhouse–Geisser, (b) Huynh–Feldt, and (c) Lower-bound (see Nimon and Williams, 2009). Alternatively, researchers may conduct and report analyses that do not assume sphericity (e.g., MANOVA).

Homoscedasticity applies to multiple linear regression and canonical correlation and assumes that the variability in scores for one continuous variable is roughly the same at all values of another continuous variable (Tabachnick and Fidell, 2001). Scatterplots are typically used to test homoscedasticity. Linear regression is generally robust to slight violations of homoscedasticity; however, marked heteroscedasticity increases the risk of Type I error (Osborne and Waters, 2002). Canonical correlation performs best when relationships among pairs of variables are homoscedastic (Tabachnick and Fidell, 2001). If the homoscedasticity assumption is violated, researchers may delete outlying cases, transform data, or conduct non-parametric tests (see Conover, 1999; Osborne, 2012), as long as the process is clearly reported.

Homogeneity of variance-covariance matrix is a multivariate generalization of homogeneity of variance. It applies to multivariate group analyses (MANOVA and MANCOVA) and assumes that the variance-covariance matrix is roughly the same at all levels of the IV (Stevens, 2002). The Box M test tests this assumption, where smaller statistics indicate greater homogeneity. Tabachnick and Fidell (2001) provided the following guidelines for interpreting violations of this assumption: if sample sizes are equal, heterogeneity is not an issue. However, with unequal sample sizes, heterogeneity may compromise the validity of null hypothesis decisions. Large sample variances from small-group sizes increase the risk of Type I error whereas large sample variances from large-group sizes increase the risk of Type II error. If sample sizes are unequal and the Box M test is significant at $p < 0.001$, researchers should conduct the Pillai's test or equalize sample sizes by random deletion of cases if power can be retained.

DISCUSSION

With the advances in statistical software, it is easy for researchers to use point and click methods to conduct a wide variety of statistical analyses on their datasets. However, the output from statistical software packages typically does not fully indicate if necessary statistical assumptions have been met. I invite editors and reviewers to use the information presented in this article as a basic checklist of the statistical assumptions to be reported in scholarly reports. The references cited in this article should also be helpful to researchers who are unfamiliar with a particular assumption or how to test it.

For example, Osborne's (2012) book provides an accessible treatment of a wide variety of data transformation techniques while Burdenski's (2000) article review graphics procedures to evaluate univariate, bivariate, and multivariate normality. Finally, the information presented in this article should be helpful to readers of scholarly reports. Readers cannot presume that just because an article has survived peer review, the interpretation of the findings is methodologically sound (cf. Henson et al., 2010). Readers must make their own judgment as to the quality of the study if information that could affect the validity of the data presented is not reported. With the information presented in this article and others, I invite readers to take an active role in evaluating the findings of quantitative research reports and become informed consumers of the data presented.

REFERENCES

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychol. Bull.* 57, 49–64.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Ann. Math. Statist.* 25, 484–498.

Bump, W. (1991). The normal curve takes many forms: a review of skewness and kurtosis. *Paper Presented at the Annual Meeting of the Southwest Educational Research Association*, San Antonio. [ERIC Document Reproduction Service No. ED 342 790].

Burdenski, T. K. (2000). Evaluating univariate, bivariate, and multivariate normality using graphical procedures. *Mult. Linear Regression Viewp.* 26, 15–28.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edn. Mahwah, NJ: Erlbaum.

Conover, W. J. (1999). *Practical Nonparametric Statistics*, 3rd Edn. New York: Wiley.

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadsworth.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychol. Methods* 2, 292–307.

Dillon, W. R., and Bearden, W. (2001). Writing the survey question to capture the concept. *J. Consum. Psychol.* 10, 67–69.

Fan, X. (1996). A SAS program for assessing multivariate normality. *Educ. Psychol. Meas.* 56, 668–674.

Gall, M. D., Gall, J. P., and Borg, W. R. (2007). *Educational Research: An Introduction*, 8th Edn. Boston, MA: Allyn and Bacon.

Garson, G. D. (2012). *Hierarchical Linear Modeling: Guide and Applications*. Thousand Oaks, CA: Sage Publications, Inc.

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev. Educ. Res.* 42, 237–288.

Graham, J. M. (2008). The general linear model as structural equation modeling. *J. Educ. Behav. Stat.* 33, 485–506.

Harris, R. J. (2001). *A Primer of Multivariate Statistics*, 3rd Edn. Mahwah, NJ: Erlbaum.

Henson, R. K. (1999). "Multivariate normality: what is it and how is it assessed?" in *Advances in Social Science Methodology*, Vol. 5, ed. B. Thompson (Stamford, CT: JAI Press), 193–211.

Henson, R. K., Hull, D. M., and Williams, C. S. (2010). Methodology in our education research culture: toward a stronger collective quantitative proficiency. *Educ. Res.* 39, 229–240.

Hoekstra, R., Kiers, H. A. L., and Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Front. Psychol.* 3:137. doi:10.3389/fpsyg.2012.00137

Howell, D. D. (1992). *Statistical Methods for Psychology*, 3rd Edn. Boston: PWS-Kent.

Kerlinger, F. N. (1986). *Foundations of Behavioral Research*, 3rd Edn. New York: Holt, Rinehart and Winston.

Kieffer, K. M., Reese, R. J., and Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: a methodological review. *J. Exp. Educ.* 69, 280–309.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*, 2nd Edn. New York, NY: The Guilford Press.

Kreft, I. G. G., and de Leeuw, J. (1998). *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage.

Mardia, K. V. (1985). "Mardia's test of multinormality," in *Encyclopedia of Statistical Sciences*, Vol. 5, eds S. Kotz and N. L. Johnson (New York: Wiley), 217–221.

Maxwell, S. E., and Delaney, H. D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 2nd Edn. Mahwah, NJ: Erlbaum.

Nimon, K., and Williams, C. (2009). Performance improvement through repeated measures: a primer for educators considering univariate and multivariate design. *Res. High. Educ. J.* 2, 117–136.

Nimon, K., Zientek, L. R., and Henson, R. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Front. Psychol.* 3:102. doi:10.3389/fpsyg.2012.00102

Onwuegbuzie, A. J., and Daniel, L. G. (2003). "Typology of analytical and interpretational errors in quantitative and qualitative educational research," in *Current Issues in Education*, Vol. 6. Available at: http://cie.ed.asu.edu/volume6/number2/ [accessed February 19, 2003].

Osborne, J., and Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Pract. Assess. Res. Eval.* 8. Available at: http://PAREonline.net/getvn.asp?v=8&n=2

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Pract. Assess. Res. Eval.* 71. Available at: http://pareonline.net/getvn.asp?v=7&n=1

Osborne, J. W. (2012). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Thousand Oaks, CA: Sage.

Schumacker, R. E., and Lomax, R. G. (2004). *A Beginner's Guide to Structural Equation Modeling*. Mahwah, NJ: Erlbaum.

Stevens, J. (2002). *Applied Multivariate Statistics for the Social Sciences*, 4th Edn. Mahwah, NJ: Erlbaum.

Stevens, S. (1946). On the theory of scales of measurement. *Science* 103, 677–680.

Stevens, S. (1951). "Mathematics, measurement, and psychophysics," in *Handbook of Experimental Psychology*, ed. S. Stevens (New York: Wiley), 1–49.

Tabachnick, B. G., and Fidell, L. S. (2001). *Using Multivariate Statistics*, 4th Edn. Needham Heights, MA: Allyn and Bacon.

Thompson, B. (1988). Discard variance: a cardinal since in research. *Meas. Eval. Couns. Dev.* 21, 3–4.

Thompson, B. (1990). Multinor: a Fortran program that assists in evaluating multivariate normality. *Educ. Psychol. Meas.* 50, 845–848.

Thompson, B. (2006). *Foundations of Behavioral Statistics: An Insight-Based Approach*. New York: Guilford Press.

Warner, R. M. (2008). *Applied Statistics: From Bivariate through Multivariate Techniques*. Thousand Oaks, CA: Sage.

Wilkinson, L., and APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanation. *Am. Psychol.* 54, 594–604.

Zientek, L. R., Capraro, M. M., and Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: one look at the evidence cited in the AERA panel report. *Educ. Res.* 37, 208–216.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 June 2012; paper pending published: 26 July 2012; accepted: 12 August 2012; published online: 28 August 2012.

*Citation: Nimon KF (2012) Statistical assumptions of substantive analyses across the general linear model: a mini-review. Front. Psychology **3**:322. doi: 10.3389/fpsyg.2012.00322*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

METHODS ARTICLE
published: 14 March 2012
doi: 10.3389/fpsyg.2012.00044

Tools to support interpreting multiple regression in the face of multicollinearity

Amanda Kraha[1]*, Heather Turner[2], Kim Nimon[3], Linda Reichwein Zientek[4] and Robin K. Henson[2]

[1] *Department of Psychology, University of North Texas, Denton, TX, USA*
[2] *Department of Educational Psychology, University of North Texas, Denton, TX, USA*
[3] *Department of Learning Technologies, University of North Texas, Denton, TX, USA*
[4] *Department of Mathematics and Statistics, Sam Houston State University, Huntsville, TX, USA*

Edited by:
Jason W Osborne, Old Dominion University, USA

Reviewed by:
Elizabeth Stone, Educational Testing Service, USA
James Stamey, Baylor University, USA

****Correspondence:***
Amanda Kraha, Department of Psychology, University of North Texas, 1155 Union Circle No. 311280, Denton, TX 76203, USA.
e-mail: amandakraha@my.unt.edu

While multicollinearity may increase the difficulty of interpreting multiple regression (MR) results, it should not cause undue problems for the knowledgeable researcher. In the current paper, we argue that rather than using one technique to investigate regression results, researchers should consider multiple indices to understand the contributions that predictors make not only to a regression model, but to each other as well. Some of the techniques to interpret MR effects include, but are not limited to, correlation coefficients, beta weights, structure coefficients, all possible subsets regression, commonality coefficients, dominance weights, and relative importance weights. This article will review a set of techniques to interpret MR effects, identify the elements of the data on which the methods focus, and identify statistical software to support such analyses.

Keywords: multicollinearity, multiple regression

Multiple regression (MR) is used to analyze the variability of a dependent or criterion variable using information provided by independent or predictor variables (Pedhazur, 1997). It is an important component of the general linear model (Zientek and Thompson, 2009). In fact, MR subsumes many of the quantitative methods that are commonly taught in education (Henson et al., 2010) and psychology doctoral programs (Aiken et al., 2008) and published in teacher education research (Zientek et al., 2008). One often cited assumption for conducting MR is minimal correlation among predictor variables (cf. Stevens, 2009). As Thompson (2006) explained, "Collinearity (or multicollinearity) refers to the extent to which the predictor variables have non-zero correlations with each other" (p. 234). In practice, however, predictor variables are often correlated with one another (i.e., multicollinear), which may result in combined prediction of the dependent variable.

Multicollinearity can lead to increasing complexity in the research results, thereby posing difficulty for researcher interpretation. This complexity, and thus the common admonition to avoid multicollinearity, results because the combined prediction of the dependent variable can yield regression weights that are poor reflections of variable relationships. Nimon et al. (2010) noted that correlated predictor variables can "complicate result interpretation. . . a fact that has led many to bemoan the presence of multicollinearity among observed variables" (p. 707). Indeed, Stevens (2009) suggested "Multicollinearity poses a real problem for the researcher using multiple regression" (p. 74).

Nevertheless, Henson (2002) observed that multicollinearity should not be seen as a problem if additional analytic information is considered:

> The bottom line is that multicollinearity is not a problem in multiple regression, and therefore not in any other [general linear model] analysis, if the researcher invokes structure coefficients in addition to standardized weights. In fact, in some multivariate analyses, multicollinearity is actually encouraged, say, for example, when multi-operationalizing a dependent variable with several similar measures. (p. 13)

Although multicollinearity is not a direct statistical assumption of MR (cf. Osborne and Waters, 2002), it complicates interpretation as a function of its influence on the magnitude of regression weights and the potential inflation of their standard error (SE), thereby negatively influencing the statistical significance tests of these coefficients. Unfortunately, many researchers rely heavily on standardized (beta, β) or unstandardized (slope) regression weights when interpreting MR results (Courville and Thompson, 2001; Zientek and Thompson, 2009). In the presence of multicollinear data, focusing solely on regression weights yields at best limited information and, in some cases, erroneous interpretation. However, it is not uncommon to see authors argue for the *importance* of predictor variables to a regression model based on the results of null hypothesis statistical significance tests of these regression weights without consideration of the multiple complex relationships between predictors and predictors with their outcome.

PURPOSE

The purpose of the present article is to discuss and demonstrate several methods that allow researchers to fully interpret and understand the contributions that predictors play in forming regression effects, even when confronted with collinear relationships among the predictors. When faced with multicollinearity in MR (or other general linear model analyses), researchers should be aware of and judiciously employ various techniques available for interpretation. These methods, when used correctly, allow researchers to reach better and more comprehensive understandings of their data than

would be attained if only regression weights were considered. The methods examined here include inspection of zero-order correlation coefficients, β weights, structure coefficients, commonality coefficients, all possible subsets regression, dominance weights, and relative importance weights (RIW). Taken together, the various methods will highlight the complex relationships between predictors themselves, as well as between predictors and the dependent variables. Analysis from these different standpoints allows the researcher to fully investigate regression results and lessen the impact of multicollinearity. We also concretely demonstrate each method using data from a heuristic example and provide reference information or direct syntax commands from a variety of statistical software packages to help make the methods accessible to readers.

In some cases multicollinearity may be desirable and part of a well-specified model, such as when multi-operationalizing a construct with several similar instruments. In other cases, particularly with poorly specified models, multicollinearity may be so high that there is unnecessary redundancy among predictors, such as when including both subscale and total scale variables as predictors in the same regression. When unnecessary redundancy is present, researchers may reasonably consider deletion of one or more predictors to reduce collinearity. When predictors are related and theoretically meaningful as part of the analysis, the current methods can help researchers parse the roles related predictors play in predicting the dependent variable. Ultimately, however, the degree of collinearity is a judgement call by the researcher, but these methods allow researchers a broader picture of its impact.

PREDICTOR INTERPRETATION TOOLS

CORRELATION COEFFICIENTS

One method to evaluate a predictor's contribution to the regression model is the use of correlation coefficients such as Pearson *r*, which is the zero-order bivariate linear relationship between an independent and dependent variable. Correlation coefficients are sometimes used as validity coefficients in the context of construct measurement relationships (Nunnally and Bernstein, 1994). One advantage of *r* is that it is the fundamental metric common to all types of correlational analyses in the general linear model (Henson, 2002; Thompson, 2006; Zientek and Thompson, 2009). For interpretation purposes, Pearson *r* is often squared (r^2) to calculate a variance-accounted-for effect size.

Although widely used and reported, *r* is somewhat limited in its utility for explaining MR relationships in the presence of multicollinearity. Because *r* is a zero-order bivariate correlation, it does not take into account any of the MR variable relationships except that between a single predictor and the criterion variable. As such, *r* is an inappropriate statistic for describing regression results as it does not consider the complicated relationships between predictors themselves and predictors and criterion (Pedhazur, 1997; Thompson, 2006). In addition, Pearson *r* is highly sample specific, meaning that *r* might change across individual studies even when the population-based relationship between the predictor and criterion variables remains constant (Pedhazur, 1997).

Only in the hypothetical (and unrealistic) situation when the predictors are perfectly uncorrelated is *r* a reasonable representation of predictor contribution to the regression effect. This is because the overall R^2 is simply the sum of the squared correlations between each predictor (X) and the outcome (Y):

$$R^2 = r_{Y-X1}2 + r_{Y-X2}2 + \ldots + r_{Y-Xk}2, \text{ or}$$
$$R^2 = (r_{Y-X1})(r_{Y-X1}) + (r_{Y-X2})(r_{Y-X2}) + \ldots + (r_{Y-Xk})(r_{Y-Xk}). \quad (1)$$

This equation works only because the predictors explain different and unique portions of the criterion variable variance. When predictors are correlated and explain some of the same variance of the criterion, the sum of the squared correlations would be greater than 1.00, because *r* does not consider this multicollinearity.

BETA WEIGHTS

One answer to the issue of predictors explaining some of the same variance of the criterion is standardized regression (β) weights. Betas are regression weights that are applied to standardized (z) predictor variable scores in the linear regression equation, and they are commonly used for interpreting predictor contribution to the regression effect (Courville and Thompson, 2001). Their utility lies squarely with their function in the standardized regression equation, which speaks to how much credit each predictor variable is receiving in the equation for predicting the dependent variable, while holding all other independent variables constant. As such, a β weight coefficient informs us as to how much change (in standardized metric) in the criterion variable we might expect with a one-unit change (in standardized metric) in the predictor variable, again holding all other predictor variables constant (Pedhazur, 1997). This interpretation of a β weight suggests that its computation must simultaneously take into account the predictor variable's relationship with the criterion as well as the predictor variable's relationships with all other predictors.

When predictors are correlated, the sum of the squared bivariate correlations no longer yields the R^2 effect size. Instead, βs can be used to adjust the level of correlation credit a predictor gets in creating the effect:

$$R^2 = (\beta_1)(r_{Y-X1}) + (\beta_2)(r_{Y-X2}) + \ldots + (\beta_k)(r_{Y-Xk}). \quad (2)$$

This equation highlights the fact that β weights are not direct measures of relationship between predictors and outcomes. Instead, they simply reflect how much credit is being given to predictors in the regression equation in a particular context (Courville and Thompson, 2001). The accuracy of β weights are theoretically dependent upon having a perfectly specified model, since adding or removing predictor variables will inevitably change β values. The problem is that the true model is rarely, if ever, known (Pedhazur, 1997).

Sole interpretation of β weights is troublesome for several reasons. To begin, because they must account for *all* relationships among *all* of the variables, β weights are heavily affected by the variances and covariances of the variables in question (Thompson, 2006). This sensitivity to covariance (i.e., multicollinear) relationships can result in very sample-specific weights which can dramatically change with slight changes in covariance relationships in future samples, thereby decreasing generalizability. For example, β weights can even change in sign as new variables are added or as old variables are deleted (Darlington, 1968).

When predictors are multicollinear, variance in the criterion that can be explained by multiple predictors is often not equally divided among the predictors. A predictor might have a large correlation with the outcome variable, but might have a near-zero β weight because another predictor is receiving the credit for the variance explained (Courville and Thompson, 2001). As such, β weights are context-specific to a given specified model. Due to the limitation of these standardized coefficients, some researchers have argued for the interpretation of structure coefficients in addition to β weights (e.g., Thompson and Borrello, 1985; Henson, 2002; Thompson, 2006).

STRUCTURE COEFFICIENTS

Like correlation coefficients, structure coefficients are also simply bivariate Pearson *r*s, but they are not zero-order correlations between two observed variables. Instead, a structure coefficient is a correlation between an observed predictor variable and the predicted criterion scores, often called "Yhat" ($\widehat{Y}$) scores (Henson, 2002; Thompson, 2006). These $\widehat{Y}$ scores are the predicted estimate of the outcome variable based on the synthesis of all the predictors in regression equation; they are also the primary focus of the analysis. The variance of these predicted scores represents the portion of the total variance of the criterion scores that can be explained by the predictors. Because a structure coefficient represents a correlation between a predictor and the $\widehat{Y}$ scores, a squared structure coefficient informs us as to how much variance the predictor can explain *of the* R^2 effect observed (not of the total dependent variable), and therefore provide a sense of how much each predictor could contribute to the explanation of the entire model (Thompson, 2006).

Structure coefficients add to the information provided by β weights. Betas inform us as to the credit given to a predictor in the regression equation, while structure coefficients inform us as to the bivariate relationship between a predictor and the effect observed without the influence of the other predictors in the model. As such, structure coefficients are useful in the presence of multicollinearity. If the predictors are perfectly uncorrelated, the sum of all squared structure coefficients will equal 1.00 because each predictor will explain its own portion of the total effect (R^2). When there is shared explained variance of the outcome, this sum will necessarily be larger than 1.00. Structure coefficients also allow us to recognize the presence of suppressor predictor variables, such as when a predictor has a large β weight but a disproportionately small structure coefficient that is close to zero (Courville and Thompson, 2001; Thompson, 2006; Nimon et al., 2010).

ALL POSSIBLE SUBSETS REGRESSION

All possible subsets regression helps researchers interpret regression effects by seeking a smaller or simpler solution that still has a comparable R^2 effect size. All possible subsets regression might be referred to by an array of synonymous names in the literature, including regression weights for submodels (Braun and Oswald, 2011), all possible regressions (Pedhazur, 1997), regression by leaps and bounds (Pedhazur, 1997), and all possible combination solution in regression (Madden and Bottenberg, 1963).

The concept of all possible subsets regression is a relatively straightforward approach to explore for a regression equation until the *best* combination of predictors is used in a single equation (Pedhazur, 1997). The exploration consists of examining the variance explained by each predictor individually and then in all possible combinations up to the complete set of predictors. The best subset, or model, is selected based on judgments about the largest R^2 with the fewest number of variables relative to the full model R^2 with all predictors. All possible subsets regression is the skeleton for commonality and dominance analysis (DA) to be discussed later.

In many ways, the focus of this approach is on the total effect rather than the particular contribution of variables that make up that effect, and therefore the concept of multicollinearity is less directly relevant here. Of course, if variables are redundant in the variance they can explain, it may be possible to yield a similar effect size with a smaller set of variables. A key strength of all possible subsets regression is that no combination or subset of predictors is left unexplored.

This strength, however, might also be considered the biggest weakness, as the number of subsets requiring exploration is exponential and can be found with $2^k - 1$, where k represents the number of predictors. Interpretation might become untenable as the number of predictor variables increases. Further, results from an all possible subset model should be interpreted cautiously, and only in an exploratory sense. Most importantly, researchers must be aware that the model with the highest R^2 might have achieved such by chance (Nunnally and Bernstein, 1994).

COMMONALITY ANALYSIS

Multicollinearity is explicitly addressed with regression commonality analysis (CA). CA provides separate measures of unique variance explained for each predictor in addition to measures of shared variance for all combinations of predictors (Pedhazur, 1997). This method allows a predictor's contribution to be related to other predictor variables in the model, providing a clear picture of the predictor's role in the explanation by itself, as well as with the other predictors (Rowell, 1991, 1996; Thompson, 2006; Zientek and Thompson, 2006). The method yields all of the uniquely and commonly explained parts of the criterion variable which always sum to R^2. Because CA identifies the unique contribution that each predictor and all possible combinations of predictors make to the regression effect, it is particularly helpful when suppression or multicollinearity is present (Nimon, 2010; Zientek and Thompson, 2010; Nimon and Reio, 2011). It is important to note, however, that commonality coefficients (like other MR indices) can change as variables are added or deleted from the model because of fluctuations in multicollinear relationships. Further, they cannot overcome model misspecification (Pedhazur, 1997; Schneider, 2008).

DOMINANCE ANALYSIS

Dominance analysis was first introduced by Budescu (1993) and yields weights that can be used to determine dominance, which is a qualitative relationship defined by one predictor variable dominating another in terms of variance explained based upon pairwise variable sets (Budescu, 1993; Azen and Budescu, 2003). Because dominance is roughly determined based on which predictors explain the most variance, even when other predictors

explain some of the same variance, it tends to de-emphasize redundant predictors when multicollinearity is present. DA calculates weights on three levels (complete, conditional, and general), within a given number of predictors (Azen and Budescu, 2003).

Dominance levels are hierarchical, with complete dominance as the highest level. Complete dominance is inherently both conditional and generally dominant. The reverse, however, is not necessarily true; a generally dominant variable is not necessarily conditionally or completely dominant. Complete dominance occurs when a predictor has a greater dominance weight, or average additional R^2, in all possible pairwise (and combination) comparisons. However, complete dominance does not typically occur in real data. Because predictor dominance can present itself in more practical intensities, two lower levels of dominance were introduced (Azen and Budescu, 2003).

The middle level of dominance, referred as conditional dominance, is determined by examining the additional contribution to R^2 within specific number of predictors (k). A predictor might conditionally dominate for $k = 2$ predictors, but not necessarily $k = 0$ or 1. The conditional dominance weight is calculated by taking the average R^2 contribution by a variable for a specific k. Once the conditional dominance weights are calculated, the researcher can interpret the averages in pairwise fashion across all k predictors.

The last and lowest level of dominance is general. General dominance averages the overall additional contributions of R^2. In simple terms, the average weights from each k group ($k = 0$, 1, 2) for each predictor ($X1$, $X2$, and $X3$) are averaged for the entire model. General dominance is *relaxed* compared to the complete and conditional dominance weights to alleviate the number of undetermined dominance in data analysis (Azen and Budescu, 2003). General dominance weights provide similar results as *RIWs*, proposed by Lindeman et al. (1980) and Johnson (2000, 2004). RIWs and DA are deemed the superior MR interpretation techniques by some (Budescu and Azen, 2004), almost always producing consistent results between methods (Lorenzo-Seva et al., 2010). Finally, an important point to emphasize is that the sum of the general dominance weights will equal the multiple R^2 of the model.

Several strengths are noteworthy with a full DA. First, dominance weights provide information about the contribution of predictor variables across all possible subsets of the model. In addition, because comparisons can be made across all pairwise comparisons in the model, DA is sensitive to patterns that might be present in the data. Finally, complete DA can be a useful tool for detection and interpretation of suppression cases (Azen and Budescu, 2003).

Some weaknesses and limitations of DA exist, although some of these weaknesses are not specific to DA. DA is not appropriate in path analyses or to test a specific hierarchical model (Azen and Budescu, 2003). DA is also not appropriate for mediation and indirect effect models. Finally, as is true with all other methods of variable interpretation, model misspecification will lead to erroneous interpretation of predictor dominance (Budescu, 1993). Calculations are also thought by some to be laborious as the number of predictors increases (Johnson, 2000).

RELATIVE IMPORTANCE WEIGHTS

Relative importance weights can also be useful in the presence of multicollinearity, although like DA, these weights tend to focus on attributing general credit to primary predictors rather than detailing the various parts of the dependent variable that are explained. More specifically, RIWs are the proportionate contribution from each predictor to R^2, after correcting for the effects of the intercorrelations among predictors (Lorenzo-Seva et al., 2010). This method is recommended when the researcher is examining the relative contribution each predictor variable makes to the dependent variable rather than examining predictor ranking (Johnson, 2000, 2004) or having concern with specific unique and commonly explained portions of the outcome, as with CA. RIWs range between 0 and 1, and their sum equals R^2 (Lorenzo-Seva et al., 2010). The weights most always match the values given by general dominance weights, despite being derived in a different fashion.

Relative importance weights are computed in four major steps (see full detail in Johnson, 2000; Lorenzo-Seva et al., 2010). Step one transforms the original predictors (X) into orthogonal variables (Z) to achieve the highest similarity of prediction compared to the original predictors but with the condition that the transformed predictors must be uncorrelated. This initial step is an attempt to simplify prediction of the criterion by removing multicollinearity. Step two involves regressing the dependent variable (Y) onto the orthogonalized predictors (Z), which yields the standardized weights for each Z. Because the Zs are uncorrelated, these β weights will equal the bivariate correlations between Y and Z, thus making equations (1) and (2) above the same. In a three predictor model, for example, the result would be a 3×1 weight matrix (β) which is equal to the correlation matrix between Y and the Zs. Step three correlates the orthogonal predictors (Z) with the original predictors (X) yielding a 3×3 matrix (**R**) in a three predictor model. Finally, step four calculates the RIWs (ε) by multiplying the *squared ZX* correlations (**R**) with the *squared YZ* weights (β).

Relative importance weights are perhaps more efficiently computed as compared to computation of DA weights which requires all possible subsets regressions as building blocks (Johnson, 2004; Lorenzo-Seva et al., 2010). RIWs and DA also yield almost identical solutions, despite different definitions (Johnson, 2000; Lorenzo-Seva et al., 2010). However, these weights do not allow for easy identification of suppression in predictor variables.

HEURISTIC DEMONSTRATION

When multicollinearity is present among predictors, the above methods can help illuminate variable relationships and inform researcher interpretation. To make their use more accessible to applied researchers, the following section demonstrates these methods using a heuristic example based on the classic suppression correlation matrix from Azen and Budescu (2003), presented in **Table 1**. **Table 2** lists statistical software or secondary syntax programs available to run the analyses across several commonly used of software programs – blank spaces in the table reflect an absence of a solution for that particular analysis and solution, and should be seen as an opportunity for future development. Sections "Excel For All Available Analyses, R Code For All Available Analyses, SAS Code For All Available Analyses, and SPSS Code For All

Analyses" provide instructions and syntax commands to run various analyses in Excel, R, SAS, and SPSS, respectively. In most cases, the analyses can be run after simply inputting the correlation matrix from **Table 1** ($n = 200$ cases was used here). For SPSS (see SPSS Code For All Analyses), some analyses require the generation of data ($n = 200$) using the syntax provided in the first part of the appendix (International Business Machines Corp, 2010). Once the data file is created, the generic variable labels (e.g., var1) can be changed to match the labels for the correlation matrix (i.e., Y, $X1$, $X2$, and $X3$).

All of the results are a function of regressing Y on $X1$, $X2$, and $X3$ via MR. **Table 3** presents the summary results of this analysis, along with the various coefficients and weights examined here to facilitate interpretation.

Table 1 | Correlation matrix for classical suppression example (Azen and Budescu, 2003).

	Y	*X1*	*X2*	*X3*
Y	1.000			
X1	0.500	1.000		
X2	0.000	0.300	1.000	
X3	0.250	0.250	0.250	1.000

Reprinted with permission from Azen and Budescu (2003). Copyright 2003 by Psychological Methods.

CORRELATION COEFFICIENTS

Examination of the correlations in **Table 1** indicate that the current data indeed have collinear predictors ($X1$, $X2$, and $X3$), and therefore some of the explained variance of Y ($R^2 = 0.301$) may be attributable to more than one predictor. Of course, the bivariate correlations tell us nothing directly about the nature of shared explained variance. Here, the correlations between Y and $X1$, $X2$, and $X3$ are 0.50, 0, and 0.25, respectively. The squared correlations (r^2) suggest that $X1$ is the strongest predictor of the outcome variable, explaining 25% ($r^2 = 0.25$) of the criterion variable variance by itself. The zero correlation between Y and $X2$ suggests that there is no relationship between these variables. However, as we will see through other MR indices, interpreting the regression effect based only on the examination of correlation coefficients would provide, at best, limited information about the regression model as it ignores the relationships between predictors themselves.

BETA WEIGHTS

The β weights can be found in **Table 3**. They form the standardized regression equation which yields predicted Y scores: $\widehat{Y} = (0.517 * X1) + (-0.198 * X2) + (0.170 * X3)$, where all predictors are in standardized (Z) form. The squared correlation between Y

Table 2 | Tools to support interpreting multiple regression.

Program	Beta weights	Structure coefficients	All possible subsets	Commonality analysis[c]	Relative weights	Dominance analysis
Excel	Base	$r_s = r_{y \cdot x1}/R$	Braun and Oswald (2011)[a]		Braun and Oswald (2011)[a]	Braun and Oswald (2011)[a]
R	Nimon and Roberts (2009)	Nimon and Roberts (2009)	Lumley (2009)	Nimon et al. (2008)		
SAS	Base	base	base[b]		Tonidandel et al. (2009)[d]	Azen and Budescu (2003)[b]
SPSS	Base	Lorenzo-Seva et al. (2010)	Nimon (2010)	Nimon (2010)	Lorenzo-Seva et al. (2010), Lorenzo-Seva and Ferrando (2011), LeBreton and Tonidandel (2008)	

[a] *Up to 9 predictors,* [b] *up to 10 predictors,* [c] *A FORTRAN IV computer program to accomplish commonality analysis was developed by Morris (1976). However, the program was written for a mainframe computer and is now obsolete,* [d] *The Tonidandel et al. (2009) SAS solution computes relative weights with a bias correction, and thus results do not mirror those in the current paper. As such, we have decided not to demonstrate the solution here. However, the macro can be downloaded online (http://www1.davidson.edu/academic/psychology/Tonidandel/TonidandelProgramsMain.htm) and provides user-friendly instructions.*

Table 3 | Multiple regression results.

Predictor	β	r_s	r_s^2	r	R^2	Unique[a]	Common[a]	General dominance weights[b]	Relative importance weights
$X1$	0.517	0.911	0.830	0.500	0.250	0.234	0.016	0.241	0.241
$X2$	−0.198	0.000	0.000	0.000	0.000	0.034	−0.034	0.016	0.015
$X3$	0.170	0.455	0.207	0.250	0.063	0.026	0.037	0.044	0.045

$R^2 = 0.301$. The primary predictor suggested by a method is underlined. r is correlation between predictor and outcome variable.
r_s = structure coefficient = r/R. $r_s^2 = r^2 / R^2$. Unique = proportion of criterion variance explained uniquely by the predictor. Common = proportion of criterion variance explained by the predictor that is also explained by one or more other predictors. Unique + Common = r^2. Σ General dominance weights = Σ relative importance weights = R^2. [a] *See **Table 5** for full CA.* [b] *See **Table 6** for full DA.*

and $\widehat{Y}$ equals the overall R^2 and represents the amount of variance of Y that can be explained by $\widehat{Y}$, and therefore by the predictors collectively. The β weights in this equation speak to the amount of credit each predictor is receiving in the creation of $\widehat{Y}$, and therefore are interpreted by many as indicators of variable importance (cf. Courville and Thompson, 2001; Zientek and Thompson, 2009).

In the current example, $r^2_{Y.\widehat{Y}t} = R^2 = 0.301$, indicating that about 30% of the criterion variance can be explained by the predictors. The β weights reveal that $X1$ ($\beta = 0.517$) received more credit in the regression equation, compared to both $X2$ ($\beta = -0.198$) and $X3$ ($\beta = 0.170$). The careful reader might note that $X2$ received considerable credit in the regression equation predicting Y even though its correlation with Y was 0. This oxymoronic result will be explained later as we examine additional MR indices. Furthermore, these results make clear that the βs are not direct measures of relationship in this case since the β for $X2$ is negative even though the zero-order correlation between the $X2$ and Y is positive. This difference in sign is a good first indicator of the presence of multicollinear data.

STRUCTURE COEFFICIENTS

The structure coefficients are given in **Table 3** as r_s. These are simply the Pearson correlations between $\widehat{Y}$ and each predictor. When squared, they yield the proportion of variance in the effect (or, of the $\widehat{Y}$ scores) that can be accounted for by the predictor alone, irrespective of collinearity with other predictors. For example, the squared structure coefficient for $X1$ was 0.830 which means that of the 30.1% (R^2) effect, $X1$ can account for 83% of the explained variance by itself. A little math would show that 83% of 30.1% is 0.250, which matches the r^2 in **Table 3** as well. Therefore, the interpretation of a (squared) structure coefficient is *in relation to the explained effect* rather than to the dependent variable as a whole.

Examination of the β weights and structure coefficients in the current example suggests that $X1$ contributed most to the variance explained with the largest absolute value for both the β weight and structure coefficient ($\beta = 0.517$, $r_s = 0.911$ or $r_s^2 = 83.0\%$). The other two predictors have somewhat comparable βs but quite dissimilar structure coefficients. Predictor $X3$ can explain about 21% of the obtained effect by itself ($\beta = 0.170$, $r_s = 0.455$, $r_s^2 = 20.7\%$), but $X2$ shares no relationship with the $\widehat{Y}$ scores ($\beta = -0.198$, r_s and $r_s^2 = 0$).

On the surface it might seem a contradiction for $X2$ to explain none of the effect but still be receiving credit in the regression equation for creating the predicted scores. However, in this case $X2$ is serving as a suppressor variable and helping the other predictor variables do a better job of predicting the criterion even though $X2$ itself is unrelated to the outcome. A full discussion of suppression is beyond the scope of this article[1]. However, the current discussion makes apparent that the identification of suppression would be unlikely if the researcher were to only examine β weights when interpreting predictor contributions.

[1]Suppression is apparent when a predictor has a beta weight that is disproportionately large (thus receiving predictive credit) relative to a low or near-zero structure coefficient (thus indicating no relationship with the predicted scores). For a broader discussion of suppression, see Pedhazur (1997) and Thompson (2006).

Because a structure coefficient speaks to the bivariate relationship between a predictor and an observed effect, it is not directly affected by multicollinearity among predictors. If two predictors explain some of the same part of the $\widehat{Y}$ score variance, the squared structure coefficients do not arbitrarily divide this variance explained among the predictors. Therefore, if two or more predictors explain some of the same part of the criterion, the sum the squared structure coefficients for all predictors will be greater than 1.00 (Henson, 2002). In the current example, this sum is 1.037 ($0.830 + 0 + 0.207$), suggesting a small amount of multicollinearity. Because $X2$ is unrelated to Y, the multicollinearity is entirely a function of shared variance between $X1$ and $X3$.

ALL POSSIBLE SUBSETS REGRESSION

We can also examine how each of the predictors explain Y both uniquely and in all possible combinations of predictors. With three variables, seven subsets are possible ($2^k - 1$ or $2^3 - 1$). The R^2 effects from each of these subsets are given in **Table 4**, which includes the full model effect of 30.1% for all three predictors. Predictors $X1$ and $X2$ explain roughly 27.5% of the variance in the outcome. The difference between a three predictor versus this two predictor model is a mere 2.6% (30.1−27.5), a relatively small amount of variance explained. The researcher might choose to drop $X3$, striving for parsimony in the regression model. A decision might also be made to drop $X2$ given its lack of prediction of Y independently. However, careful examination of the results speaks again to the suppression role of $X2$, which explains none of Y directly but helps $X1$ and $X3$ explain more than they could by themselves when $X2$ is added to the model. In the end, decisions about variable contributions continue to be a function of thoughtful researcher judgment and careful examination of existing theory. While all possible subsets regression is informative, this method generally lacks the level of detail provided by both βs and structure coefficients.

COMMONALITY ANALYSIS

Commonality analysis takes all possible subsets further and divides all of the explained variance in the criterion into unique and common (or shared) parts. **Table 5** presents the commonality coefficients, which represent the proportions of variance explained in the dependent variable. The unique coefficient for $X1$ (0.234) indicates that $X1$ uniquely explains 23.4% of the variance in the

Table 4 | All possible subsets regression.

Predictor set	R^2
$X1$	0.250
$X2$	0.000
$X3$	0.063
$X1$, $X2$	0.275
$X1$, $X3$	0.267
$X2$, $X3$	0.067
$X1$, $X2$, $X3$	0.301

Predictor contribution is determined by researcher judgment. The model with the highest R^2 value, but with the most ease of interpretation, is typically chosen.

dependent variable. This amount of variance is more than any other partition, representing 77.85% of the R^2 effect (0.301). The unique coefficient for $X3$ (0.026) is the smallest of the unique effects and indicates that the regression model only improves slightly with the addition of variable $X3$, which is the same interpretation provided by the all possible subsets analysis. Note that $X2$ uniquely accounts for 11.38% of the variance in the regression effect. Again, this outcome is counterintuitive given that the correlation between $X2$ and Y is zero. However, as the common effects will show, $X2$ serves as a suppressor variable, yielding a unique effect greater than its total contribution to the regression effect and negative commonality coefficients.

The common effects represent the proportion of criterion variable variance that can be jointly explained by two or more predictors together. At this point the issue of multicollinearity is explicitly addressed with an estimate of each part of the dependent variable that can be explained by more than one predictor. For example, $X1$ and $X3$ together explain 4.1% of the outcome, which represents 13.45% of the total effect size.

It is also important to note the presence of negative commonality coefficients, which seem anomalous given that these coefficients are supposed to represent variance explained. Negative commonality coefficients are generally indicative of suppression (cf. Capraro and Capraro, 2001). In this case, they indicate that $X2$ suppresses variance in $X1$ and $X3$ that is irrelevant to explaining variance in the dependent variable, making the predictive power of their unique contributions to the regression effect larger than they would be if $X2$ was not in the model. In fact, if $X2$ were not in the model, $X1$ and $X3$ would respectively only account for 20.4% (0.234−0.030) and 1.6% (0.026−0.010) of unique variance in the dependent variable. The remaining common effects indicate that, as noted above, multicollinearity between $X1$ and $X3$ accounts for 13.45% of the regression effect and that there is little variance in the dependent variable that is common across all three predictor variables. Overall, CA can help to not only identify the most parsimonious model, but also quantify the location and amount of variance explained by suppression and multicollinearity.

Table 5 | Commonality coefficients.

Predictor(s)	*X1*	*X2*	*X3*	Coefficient	Percent
X1	0.234			0.234	77.845
X2		0.034		0.034	11.381
X3			0.026	0.026	8.702
X1, *X2*	−0.030	−0.030		−0.030	−10.000
X1, *X3*	0.041		0.041	0.041	13.453
X2, *X3*		−0.010	−0.010	−0.010	−3.167
X1, *X2*, *X3*	0.005	0.005	0.005	0.005	1.779
Total	0.250	0.000	0.063	0.301	100.000

Commonality coefficients identifying suppression underlined.

ΣX_k Commonality coefficients equals r^2 between predictor (k) and dependent variable.

Σ Commonality coefficients equals Multiple $R^2 = 30.1\%$. Percent = coefficient/multiple R^2.

DOMINANCE WEIGHTS

Referring to **Table 6**, the conditional dominance weights for the null or $k = 0$ subset reflects the r^2 between the predictor and the dependent variable. For the subset model where $k = 2$, note that the additional contribution each variable makes to R^2 is equal to the unique effects identified from CA. In the case when $k = 1$, DA provides new information to interpreting the regression effect. For example, when $X2$ is added to a regression model with $X1$, DA shows that the change (Δ) in R^2 is 0.025.

The DA weights are typically used to determine if variables have complete, conditional, or general dominance. When evaluating for complete dominance, all pairwise comparisons must be considered. Looking across all rows to compare the size of dominance weights, we see that $X1$ consistently has a larger conditional dominance weight. Because of this, it can be said that predictor $X1$ completely dominates the other predictors. When considering conditional dominance, however, only three rows must be considered: these are labeled null and $k = 0$, $k = 1$, and $k = 2$ rows. These rows provide information about which predictor dominates when there are 0, 1, and 2 *additional* predictors present. From this, we see that $X1$ conditionally dominates in all model sizes with weights of 0.250 ($k = 0$), 0.240 ($k = 1$), and 0.234 ($k = 2$). Finally, to evaluate for general dominance, only one row must be attended to. This is the overall average row. General dominance weights are the average conditional dominance weight (additional contribution of R^2) for each variable across situations. For example, $X1$ generally dominates with a weight of 0.241 [i.e., $(0.250 + 0.240 + 0.234)/3$]. An important observation is the sum of the general dominance weights ($0.241 + 0.016 + 0.044$) is also equal to 0.301, which is the total model R^2 for the MR analysis.

RELATIVE IMPORTANCE WEIGHTS

Relative importance weights were computed using the Lorenzo-Seva et al. (2010) SPSS code using the correlation matrix provided in **Table 1**. Based on RIW (Johnson, 2001), $X1$ would

Table 6 | Full dominance analysis (Azen and Budescu, 2003).

Subset model	$R^2_{Y \cdot Xi}$	Additional contribution of:		
		X1	*X2*	*X3*
Null and $k = 0$ average	0	0.250	0.000	0.063
X1	0.250		0.025	0.017
X2	0.000	0.275		0.067
X3	0.063	0.204	0.004	
$k = 1$ average		0.240	0.015	0.044[a]
X1, *X2*	0.275			0.026
X1, *X3*	0.267		0.034	
X2, *X3*	0.067	0.234		
$k = 2$ average		0.234	0.034	0.026
X1, *X2*, *X3*	0.301			
Overall average		0.241	0.016	0.044

X1 is completely dominant (underlined). Blank cells are not applicable. [a]Small differences are noted in the hundredths decimal place for X3 between Braun and Oswald (2011) and Azen and Budescu (2003).

be considered the most important variable (RIW = 0.241), followed by $X3$ (RIW = 0.045) and $X2$ (RIW = 0.015). The RIWs offer an additional representation of the individual effect of each predictor while simultaneously considering the combination of predictors as well (Johnson, 2000). The sum of the weights ($0.241 + 0.045 + 0.015 = 0.301$) is equal to R^2. Predictor $X1$ can be interpreted as the most important variable relative to other predictors (Johnson, 2001). The interpretation is consistent with a full DA, because both the individual predictor contribution with the outcome variable ($r_{X}1{\cdot}Y$), and the potential multicollinearity ($r_{X}1{\cdot}X2$ and $r_{X}1{\cdot}X3$) with other predictors are accounted for. While the RIWs may differ slightly compared to general dominance weights (e.g., 0.015 and 0.016, respectively, for $X2$), the conclusions are the consistent with those from a full DA. This method rank orders the variables with $X1$ as the most important, followed by $X3$ and $X2$. The suppression role of $X2$, however, is not identified by this method, which helps explain its rank as third in this process.

DISCUSSION

Predictor variables are more commonly correlated than not in most practical situations, leaving researchers with the necessity to addressing such multicollinearity when they interpret MR results. Historically, views about the impact of multicollinearity on regression results have ranged from challenging to highly problematic. At the extreme, avoidance of multicollinearity is sometimes even considered a prerequisite assumption for conducting the analysis. These perspectives notwithstanding, the current article has presented a set of tools that can be employed to effectively interpret the roles various predictors have in explaining variance in a criterion variable.

To be sure, traditional reliance on standardized or unstandardized weights will often lead to poor or inaccurate interpretations when multicollinearity or suppression is present in the data. If researchers choose to rely solely on the null hypothesis statistical significance test of these weights, then the risk of interpretive error is noteworthy. This is primarily because the weights are heavily affected by multicollinearity, as are their SE which directly impact the magnitude of the corresponding p values. It is this reality that has led many to suggest great caution when predictors are correlated.

Advances in the literature and supporting software technology for their application have made the issue of multicollinearity much less critical. Although predictor correlation can certainly complicate interpretation, use of the methods discussed here allow for a much broader and more accurate understanding of the MR results regarding which predictors explain how much variance in the criterion, both uniquely and in unison with other predictors.

In data situations with a small number of predictors or very low levels of multicollinearity, the interpretation method used might not be as important as results will most often be very similar. However, when the data situation becomes more complicated (as is often the case in real-world data, or when suppression exists as exampled here), more care is needed to fully understand the nature and role of predictors.

CAUSE AND EFFECT, THEORY, AND GENERALIZATION

Although current methods are helpful, it is very important that researchers remain aware that MR is ultimately a correlational-based analysis, as are all analyses in the general linear model. Therefore, variable correlations should not be construed as evidence for cause and effect relationships. The ability to claim cause and effect are predominately issues of research design rather than statistical analysis.

Researchers must also consider the critical role of theory when trying to make sense of their data. Statistics are mere tools to help understand data, and the issue of predictor importance in any given model must invoke the consideration of the theoretical expectations about variable relationships. In different contexts and theories, some relationships may be deemed more or less relevant.

Finally, the pervasive impact of sampling error cannot be ignored in any analytical approach. Sampling error limits the generalizability of our findings and can cause any of the methods described here to be more unique to our particular sample than to future samples or the population of interest. We should not assume too easily that the predicted relationships we observe will necessarily appear in future studies. Replication continues to be a key hallmark of good science.

INTERPRETATION METHODS

The seven approaches discussed here can help researchers better understand their MR models, but each has its own strengths and limitations. In practice, these methods should be used to inform each other to yield a better representation of the data. Below we summarize the key utility provided by each approach.

Pearson r correlation coefficient

Pearson r is commonly employed in research. However, as illustrated in the heuristic example, r does not take into account the multicollinearity between variables and they do not allow detection of suppressor effects.

Beta weights and structure coefficients

Interpretations of *both* β weights and structure coefficients provide a complementary comparison of predictor contribution to the regression equation and the variance explained in the effect. Beta weights alone should not be utilized to determine the contribution predictor variables make to a model because a variable might be denied predictive credit in the presence of multicollinearity. Courville and Thompson, 2001; see also Henson, 2002) advocated for the interpretation of (a) both β weights and structure coefficients or (b) both β weights and correlation coefficients. When taken together, β and structure coefficients can illuminate the impact of multicollinearity, reflect more clearly the ability of predictors to explain variance in the criterion, and identify suppressor effects. However, they do not necessarily provide detailed information about the nature of unique and commonly explained variance, nor about the magnitude of the suppression.

All possible subsets regression

All possible subsets regression is exploratory and comes with increasing interpretive difficulty as predictors are added to

the model. Nevertheless, these variance portions serve as the foundation for unique and common variance partitioning and full DA.

Commonality analysis, dominance analysis, and relative importance weights

Commonality analysis decomposes the regression effect into unique and common components and is very useful for identifying the magnitude and loci of multicollinearity and suppression. DA explores predictor contribution in a variety of situations and provides consistent conclusions with RIWs. Both general dominance and RIWs provide alternative techniques to decomposing the variance in the regression effect and have the desirable feature that there is only one coefficient per independent variable to interpret. However, the existence of suppression is not readily understood by examining general dominance weights or RIWs. Nor do the indices yield information regarding the magnitude and loci of multicollinearity.

CONCLUSION

The real world can be complex – and correlated. We hope the methods summarized here are useful for researchers using regression to confront this multicollinear reality. For both multicollinearity and suppression, multiple pieces of information should be consulted to understand the results. As such, these data situations should not be shunned, but simply handled with appropriate interpretive frameworks. Nevertheless, the methods are not a panacea, and require appropriate use and diligent interpretation. As correctly stated by Wilkinson and the APA Task Force on Statistical Inference (1999), "Good theories and intelligent interpretation advance a discipline more than rigid methodological orthodoxy. . . Statistical methods should guide and discipline our thinking but should not determine it" (p. 604).

REFERENCES

Aiken, L. S., West, S. G., and Millsap, R. E. (2008). Doctorial training in statistics, measurement, and methodology in psychology: replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *Am. Psychol.* 63, 32–50.

Azen, R., and Budescu, D. V. (2003). The dominance analysis approach to comparing predictors in multiple regression. *Psychol. Methods* 8, 129–148.

Braun, M. T., and Oswald, F. L. (2011). Exploratory regression analysis: a tool for selecting models and determining predictor importance. *Behav. Res. Methods* 43, 331–339.

Budescu, D. V. (1993). Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychol. Bull.* 114, 542–551.

Budescu, D. V., and Azen, R. (2004). Beyond global measures of relative importance: some insights from dominance analysis. *Organ. Res. Methods* 7, 341–350.

Capraro, R. M., and Capraro, M. M. (2001). Commonality analysis: understanding variance contributions to overall canonical correlation effects of attitude toward mathematics on geometry achievement. *Mult. Lin. Regression Viewpoints* 27, 16–23.

Courville, T., and Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: is not enough. *Educ. Psychol. Meas.* 61, 229–248.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychol. Bull.* 69, 161–182.

Henson, R. K. (2002). The logic and interpretation of structure coefficients in multivariate general linear model analyses. *Paper Presented at the Annual Meeting of the American Educational Research Association*, New Orleans.

Henson, R. K., Hull, D. M., and Williams, C. (2010). Methodology in our education research culture: toward a stronger collective quantitative proficiency. *Educ. Res.* 39, 229–240.

International Business Machines Corp. (2010). *Can SPSS Help me Generate a File of Raw Data with a Specified Correlation Structure?* Available at: https://www-304.ibm.com/support/docview.wss?uid=swg21480900

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behav. Res.* 35, 1–19.

Johnson, J. W. (2001). "Determining the relative importance of predictors in multiple regression: practical applications of relative weights," in *Advances in Psychology Research*, Vol. V, eds F. Columbus and F. Columbus (Hauppauge, NY: Nova Science Publishers), 231–251.

Johnson, J. W. (2004). Factors affecting relative weights: the influence of sampling and measurement error. *Organ. Res. Methods* 7, 283–299.

LeBreton, J. M., and Tonidandel, S. (2008). Multivariate relative importance: relative weight analysis to multivariate criterion spaces. *J. Appl. Psychol.* 93, 329–345.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis.* Glenview, IL: Scott Foresman.

Lorenzo-Seva, U., and Ferrando, P. J. (2011). FIRE: an SPSS program for variable selection in multiple linear regression via the relative importance of predictors. *Behav. Res. Methods* 43, 1–7.

Lorenzo-Seva, U., Ferrando, P. J., and Chico, E. (2010). Two SPSS programs for interpreting multiple regression results. *Behav. Res. Methods* 42, 29–35.

Lumley, T. (2009). *Leaps: Regression Subset Selection. R Package Version 2.9.* Available at: http://CRAN.R-project.org/package=leaps

Madden, J. M., and Bottenberg, R. A. (1963). Use of an all possible combination solution of certain multiple regression problems. *J. Appl. Psychol.* 47, 365–366.

Morris, J. D. (1976). A computer program to accomplish commonality analysis. *Educ. Psychol. Meas.* 36, 721–723.

Nimon, K. (2010). Regression commonality analysis: demonstration of an SPSS solution. *Mult. Lin. Regression Viewpoints* 36, 10–17.

Nimon, K., Henson, R., and Gates, M. (2010). Revisiting interpretation of canonical correlation analysis: a tutorial and demonstration of canonical commonality analysis. *Multivariate Behav. Res.* 45, 702–724.

Nimon, K., Lewis, M., Kane, R., and Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: an introduction to the package and a practical example. *Behav. Res. Methods* 40, 457–466.

Nimon, K., and Reio, T. (2011). Regression commonality analysis: a technique for quantitative theory building. *Hum. Resour. Dev. Rev.* 10, 329–340.

Nimon, K., and Roberts, J. K. (2009). *Yhat: Interpreting Regression effects. R Package Version 1.0-3.* Available at: http://CRAN.R-project.org/package=yhat

Nunnally, J.C., and Bernstein, I. H. (1994). *Psychometric Theory*, 3rd Edn. New York: McGraw-Hill.

Osborne, J., and Waters, E. (2002). *Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research & Evaluation, 8(2).* Available at: http://PAREonline.net/getvn.asp?v=8&n=2 [accessed December 12, 2011]

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction*, 3rd Edn. Fort Worth, TX: Harcourt Brace.

Rowell, R. K. (1991). Partitioning predicted variance into constituent parts: how to conduct commonality analysis. *Paper Presented at the Annual Meeting of the Southwest Educational Research Association*, San Antonio.

Rowell, R. K. (1996). "Partitioning predicted variance into constituent parts: how to conduct commonality analysis," in *Advances in Social science Methodology,* Vol. 4, ed. B. Thompson (Greenwich, CT: JAI Press), 33–44.

Schneider, W. J. (2008). Playing statistical ouija board with commonality analysis: good questions, wrong assumptions. *Appl. Neuropsychol.* 15, 44–53.

Stevens, J. P. (2009). *Applied Multivariate Statistics for the Social Sciences*, 4th Edn. New York: Routledge.

Thompson, B. (2006). *Foundations of Behavioral Statistics: An Insight-Based Approach.* New York: Guilford Press.

Thompson, B., and Borrello, G. M. (1985). The importance of structure coefficients in regression research. *Educ. Psychol. Meas.* 45, 203–209.

Tonidandel, S., LeBreton, J. M., and Johnson, J. W. (2009). Determining the statistical significance of relative weights. *Psychol. Methods* 14, 387–399.

UCLA: Academic Technology Services, Statistical Consulting Group. (n.d.). *Introduction to SAS*. Available at: http://www.ats.ucla.edu/stat/sas

Wilkinson, L., and APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanation. *Am. Psychol.* 54, 594–604.

Zientek, L. R., Capraro, M. M., and Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: one look at the evidence cited in the AERA panel report. *Educ. Res.* 37, 208–216.

Zientek, L. R., and Thompson, B. (2006). Commonality analysis: partitioning variance to facilitate better understanding of data. *J. Early Interv.* 28, 299–307.

Zientek, L. R., and Thompson, B. (2009). Matrix summaries improve research reports: secondary analyses using published literature. *Educ. Res.* 38, 343–352.

Zientek, L. R., and Thompson, B. (2010). Using commonality analysis to quantify contributions that self-efficacy and motivational factors make in mathematics performance. *Res. Sch.* 17, 1–12.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 December 2011; paper pending published: 17 January 2012; accepted: 07 February 2012; published online: 14 March 2012.

*Citation: Kraha A, Turner H, Nimon K, Zientek LR and Henson RK (2012) Tools to support interpreting multiple regression in the face of multicollinearity. Front. Psychology **3**:44. doi: 10.3389/fpsyg.2012.00044*

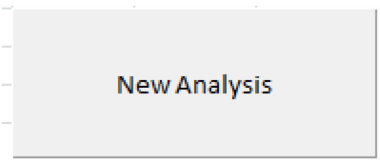
This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

APPENDIX

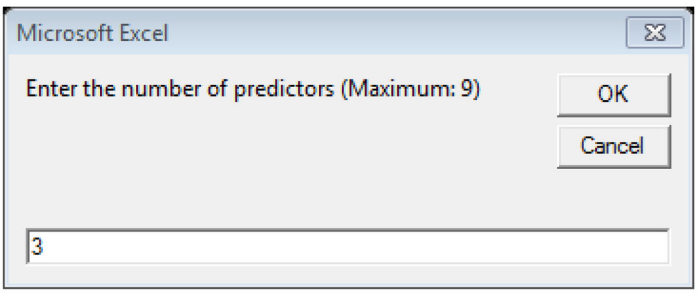
EXCEL FOR ALL AVAILABLE ANALYSES

Note. Microsoft Excel version 2010 is demonstrated. The following will yield all possible subsets, relative importance weights, and dominance analysis results.

Download the Braun and Oswald (2011) Excel file (ERA.xlsm) from
http://dl.dropbox.com/u/2480715/ERA.xlsm?dl = 1
Save the file to your desktop
Click **Enable Editing**, if prompted
Click **Enable Macros,** if prompted
Step 1: Click on **New Analysis**



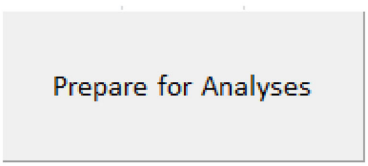
Step 2: Enter the number of predictors and click **OK**



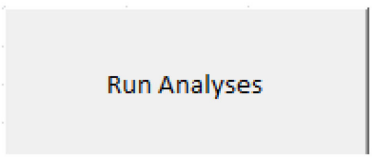
Step 3: Enter the correlation matrix as shown

	X1	X2	X3	Y
X1	1			
X2	0.3	1		
X3	0.25	0.25	1	
Y	0.5	0	0.25	1

Step 4: Click **Prepare for Analyses** to complete the matrix



Step 5: Click **Run Analyses**

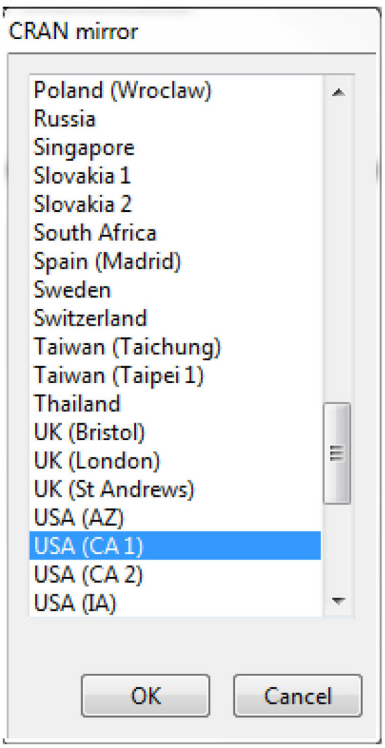


Step 6: Review output in the **Results** worksheet

R CODE FOR ALL AVAILABLE ANALYSES

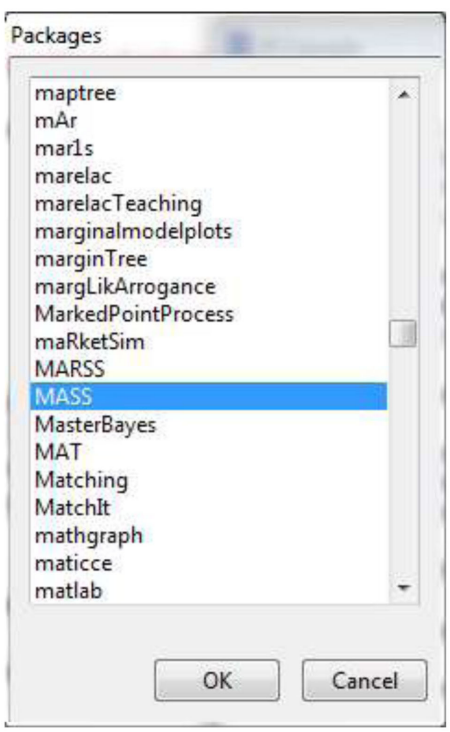
Note. R Code for Versions 2.12.1 and 2.12.2 are demonstrated.

Open R
Click on Packages → **Install package(s)**
Select the one package from a user-selected CRAN mirror site (e.g., USA CA 1)



Packages
1. MASS
2. corpcor
3. yhat
4. leaps

Repeat installation for all four packages
Click on Packages → Load for each package (for a total of four times)



Step 1: Copy and paste the following code to Generate Data from Correlation Matrix

```
library(MASS)
library(corpcor)
covm<-c(1.00,0.5,0.00,0.25,
      0.5, 1,0.3,0.25,
      0.0,0.30, 1,0.25,
      0.25,0.25,0.25, 1)
covm<-matrix(covm,4,4)
covm<-make.positive.definite(covm)
varlist<-c("DV", "IV1", "IV2", "IV3")
dimnames(covm)<-list(varlist,varlist)
data1<-mvrnorm(n=200,rep(0,4),covm,empirical=TRUE)
data1<-data.frame(data1)
```

Step 2: Copy and paste the following code to **Produce Beta Weights, Structure Coefficients, and Commonality Coefficients**

```
library(yhat)
lmOut<-lm(DV∼IV1+IV2+IV3,data1)
regrOut<-regr(lmOut)
regrOut$Beta_Weights
regrOut$Structure_Coefficients
regrOut$Commonality_Data
```

Step 3: All Possible Subset Analysis

```
library(leaps)
a<-regsubsets(data1[,(2:4)],data1[,1],method='exhaustive',nbest=7)
cbind(summary(a)$which,rsq = summary(a)$rsq)
```

SAS CODE FOR ALL AVAILABLE ANALYSES

Note. SAS Code is demonstrated in SAS Version 9.2.

Open SAS
Click on File → New Program

Step 1: Copy and paste the following code to **Generate Data from Correlation Matrix**

```
options pageno = min nodate formdlim = '-';
DATA corr (TYPE=CORR);
LENGTH _NAME_ $ 2;
INPUT _TYPE_$_NAME_$Y X1 X2 X3;
CARDS;
corr Y 1.00 .500 .000 .250
corr X1 .500 1.00 .300 .250
corr X2 .000 .300 1.00 .250
corr X3 .250 .250 .250 1.00
;
```

Step 2: **Download** a SAS macro from UCLA (n.d.) Statistics http://www.ats.ucla.edu/stat/sas/macros/corr2data.sas and save the file as **"Corr2Data.sas"** to a working directory such as "My Documents"

Step 3: **Copy and paste** the code below.

```
%include 'C:\My Documents\corr2data.sas';
%corr2data(mycorr, corr, 200, FULL = 'T', corr = 'T') ;
```

Step 4: Copy and paste the code below to rename variables with the macro (referenced from http://www.ats.ucla.edu/stat/sas/code/renaming_variables_dynamically.htm)

```
%macro rename1(oldvarlist, newvarlist);
%let k=1;
%let old = %scan(&oldvarlist, &k);
%let new = %scan(&newvarlist, &k);
%do %while(("&old" NE "") & ("&new" NE ""));
rename &old = &new;
```

```
%let k=%eval(&k+1);
%let old=%scan(&oldvarlist, &k);
%let new=%scan(&newvarlist, &k);
%end;
%mend;
```

COMMENT set dataset;
data Azen;
set mycorr;
COMMENT set (old, new) variable names;
`%rename1(col1 col2 col3 col4, Y X1 X2 X3);`
run;

Step 5: Copy and paste the code below to **Conduct Regression Analyses and Produce Beta Weights and Structure Coefficients.**

```
proc reg data=azen;model Y =X1 X2 X3;
output r=resid p=pred;
run;
```

COMMENT structure coefficients;

```
proc corr; VAR pred X1 X2 X3;
run;
```

Step 6: Copy and paste the code below to conduct an **All Possible Subset Analysis**

```
proc rsquare; MODEL Y =X1 X2 X3;
run;
```

Step 7: Link to https://pantherfile.uwm.edu/azen/www/DAonly.txt (Azen and Budescu, 1993)

Step 8: Copy and paste the text below the line of asterisks (i.e., the code beginning at run; option nosource;).

Step 9: Save the SAS file as "dom.sas" to a working directory such as "My Documents."

Step 8: Copy and paste the code below to conduct a full **Dominance Analysis**

```
%include 'C:\My Documents\dom.sas'; *** CHANGE TO PATH WHERE MACRO IS SAVED ***;
%dom(p = 3);
```

SPSS CODE FOR ALL ANALYSES

Notes. SPSS Code demonstrated in Version 19.0. SPSS must be at least a graduate pack with syntax capabilities.

Reprint Courtesy of International Business Machines Corporation, ©(2010) International Business Machines Corporation. The syntax was retrieved from https://www-304.ibm.com/support/docview.wss?uid = swg21480900.

Open SPSS
If a dialog box appears, click **Cancel** and open SPSS data window.
Click on **File** → **New** → **Syntax**

Step 1: Generate Data from Correlation Matrix. Be sure to specify a valid place to save the correlation matrix. **Copy and paste** syntax below into the SPSS syntax editor.

```
matrix data variables=v1 to v4
/contents=corr.
begin data.
1.000
0.500 1.000
0.000 .300 1.000
0.250 .250 .250 1.000
end data.
save outfile="C:\My Documents\corrmat.sav"
/keep=v1 to v4.
```

Step 2: Generate raw data. Change #i from 200 to your desired N. Change $x(4)$ and #j from 4 to the size of your correlation matrix, if different. Double Check the filenames and locations.

```
new file.
input program.
```

```
loop #i=1 to 200.
vector x(4).
loop #j=1 to 4.
compute x(#j)=rv.normal(0,1).
end loop.
end case.
end loop.
end file.
end input program.
execute.
factor var=x1 to x4
/criteria=factors(4)
/save=reg(all z).
matrix.
get z/var=z1 to z4.
get r/file='C:\My Documents\corrmat.sav'.
compute out=z*chol(r).
save out/outfile='C:\My Documents\AzenData.sav'.
end matrix.
```

Step 3: Retrieve file generated from the syntax above. **Copy and paste** the syntax below Highlight the syntax and run the selection by clicking on the  button.

```
get file='C:\My Documents\AzenData.sav'.
```

Step 4: Rename variables if desired. Replace "var1 to var10" with appropriate variable names. **Copy and paste** the syntax below and run the selection by highlighting one line. Be sure to save changes.

```
rename variables(col1 col2 col3 col4=Y X1 X2 X3).
```

Step 5: Copy and paste the syntax into the syntax editor to confirm correlations are correct.

```
CORRELATIONS
/VARIABLES = Y X1 X2 X3
/PRINT = TWOTAIL NOSIG
/MISSING = PAIRWISE.
```

Step 6: **Copy and paste** the syntax into the syntax editor to **Conduct Regression and Produce Beta Weights.**

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA = PIN(0.05) POUT(0.10)
/NOORIGIN
/DEPENDENT Y
/METHOD = ENTER X1 X2 X3
/SAVE PRED.
```

Step 7: **Copy and paste** the syntax into the syntax editor to ***Compute Structure Coefficients.**

```
CORRELATIONS
/VARIABLES=X1 X2 X3 WITH PRE_1
/PRINT = TWOTAIL NOSIG
/MISSING = PAIRWISE.
```

Step 8: All Subset Analysis and Commonality analysis (step 8 to step 11). Before executing, **download** cc.sps (commonality coefficients macro) from http://profnimon.com/CommonalityCoefficients.sps to working directory such as My Documents.

Step 9: Copy data file to working directory (e.g., C:\My Documents)

Step 10: Copy and paste syntax below in the SPSS syntax editor

```
CD "C:\My Documents".
INCLUDE FILE="CommonalityCoefficients.sps".
```

```
!cc dep=Y
Db=AzenData.sav
Set=Azen
Ind=X1 X2 X3.
```

Step 11: Retrieve commonality results. Commonality files are written to **AzenCommonalityMatrix.sav** and **AzenCCByVariable.sav.** APS files are written to**Azenaps.sav.**

Step 12: Relative Weights (Step 13 to Step 16).

Step 13: Before executing, download mimr_raw.sps and save to working directory from http://www.springerlink.com/content/06112u8804155th6/supplementals/

Step 14: Open or **activate** the AzenData.sav dataset file, by clicking on it.

Step 15: If applicable, change the reliabilities of the predictor variables as indicated (4 in the given example).

Step 16: Highlight all the syntax and run; these steps will yield relative importance weights.

frontiers in
PSYCHOLOGY

METHODS ARTICLE
published: 09 July 2012
doi: 10.3389/fpsyg.2012.00231

A simple statistic for comparing moderation of slopes and correlations

Michael Smithson*

Psychology, The Australian National University, Canberra, ACT, Australia

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Fernando Marmolejo-Ramos, University of Adelaide, Australia
Kristopher Jason Preacher, Vanderbilt University, USA

****Correspondence:***
Michael Smithson, Psychology, The Australian National University, Building 39 Science Road, Canberra, ACT, Australia.
e-mail: michael.smithson@anu.edu.au

Given a linear relationship between two continuous random variables X and Y that may be moderated by a third, Z, the extent to which the correlation ρ is (un)moderated by Z is equivalent to the extent to which the regression coefficients β_Y and β_X are (un)moderated by Z iff the variance ratio σ_Y^2/σ_X^2 is constant over the range or states of Z. Otherwise, moderation of slopes and of correlations must diverge. Most of the literature on this issue focuses on tests for heterogeneity of variance in Y, and a test for this ratio has not been investigated. Given that regression coefficients are proportional to ρ via this ratio, accurate tests, and estimations of it would have several uses. This paper presents such a test for both a discrete and continuous moderator and evaluates its Type I error rate and power under unequal sample sizes and departures from normality. It also provides a unified approach to modeling moderated slopes and correlations with categorical moderators via structural equations models.

Keywords: moderator effects, interaction effects, heteroscedasticity, regression, correlation

INTRODUCTION

Let X and Y have a bivariate normal distribution, $X \sim N(\mu_x, \sigma_x^2)$, and $Y \sim N(\mu_y, \sigma_y^2)$. Suppose also that the correlation between X and Y is a function of a moderator variable Z. Under homogeneity of variance (HoV), moderation of correlations implies moderation of regression coefficients (or means, in ANOVA), and vice versa. For example, establishing the existence of a moderator effect from Z in a linear regression model with X and Z predicting Y by finding a significant regression coefficient for the product term $X \times Z$ suffices to infer a corresponding moderator effect of Z on the correlation between X and Y.

Heterogeneity of variance (HeV) due to Z, however, can alter moderator effects so that correlation and regression coefficients are not equivalently moderated. We may have moderation of slopes, for instance, without moderation of correlations, moderation of correlations with no moderation of slopes, moderation of slopes and correlations in opposite directions, or even moderation of regression coefficients in opposite directions (e.g., what appears to be a positive moderator effect when X predicts Y becomes a negative effect when Y predicts X).

Although some scholars have warned about the impacts of heteroscedasticity on the analysis of variance (e.g., Grissom, 2000) and linear regression, most contemporary textbook advice and published evidence on this matter comforts researchers with the notion that ANOVA and regression are fairly robust against it. Howell (2007, p. 316), for instance, states that despite Grissom's pessimistic outlook "the homogeneity of variance assumption can be violated without terrible consequences" and advises that for symmetrically distributed populations and equal numbers of participants in each cell, the validity of ANOVA is likely if the ratio of the largest to the smallest variance is no greater than 4. Tabachnick and Fidell (2007, pp. 121–123) are even more relaxed, recommending an upper limit on this ratio of 10 before raising an alarm. A recent investigation into the robustness of one-way ANOVA against violations of normality (Schmider et al., 2010) also is relatively reassuring on that count. A fairly recent comparison of several tests of homogeneity of variance (Correa et al., 2006) generally finds in favor of the Levene test but leaves the issue of the impact of HoV on moderator effects unexamined.

Nevertheless, this problem is well-known. Arnold (1982) drew a distinction between the "form" and "degree" of moderator effects, whereby the "form" is indexed by moderation of slopes (or means, in ANOVA) whereas the "degree" is indexed by moderation of correlations. He argued from first principles and demonstrated empirically that it is possible to find a significant difference between correlations from two independent-samples but fail to find a corresponding significant regression interaction term, and vice versa. A related treatment was presented independently by Sharma et al. (1981), who referred to "degree" moderators as "homologizers" (a term taken from Zedeck, 1971). They pointed out that homologizers that act through the error-term in a regression instead of through the predictor itself.

Stone and Hollenbeck (1984) dissented from Arnold (1982), arguing that only moderated regression is needed to assess moderating effects, regardless of whether they are of form or degree. Their primary claims were that moderated slopes also can be interpreted as differing strengths of relationship, and that the subgrouping method advocated by Arnold raises concerns about how subgroups are created if the moderator is not categorical. Arnold (1984) rebutted their claim regarding the slope as a measure of relationship strength, reiterating the position that slopes, and correlations convey different types of information about such relationships. He also declared that both moderated regression

and tests of differences between correlation coefficients are essentially "subgroup" methods. At the time there was no way to unify the examination of moderation of correlations and slopes. The present paper describes and demonstrates such an approach for categorical moderators, via structural equations models.

In a later paper, Stone and Hollenbeck (1989) reprised this debate and recommended variance-stabilizing and homogenizing transformations as a way to eliminate the apparent disagreement between moderation of correlations and moderation of slopes. These include not only transformations of the dependent variable, but also within-groups standardization and/or normalization. They also, again, recommended abandoning the distinction between degree and form moderation and focusing solely on form (i.e., moderated regression). The usual cautions against routinely transforming variables and objections to applying different transformations to subsamples aside, we shall see that transforming the dependent variable is unlikely to eliminate the non-equivalence between moderation of slopes and correlations. Moreover, other investigators of this issue do not arrive at the same recommendation as Stone and Hollenbeck when it comes to a "best" test.

Apparently independently of the aforementioned work, and extending the earlier work of Dretzke et al. (1982), Alexander and DeShon (1994) demonstrated severe effects from heterogeneity of error-variance (HeEV) on power and Type I error rates for the F-test of equality of regression slopes. In contrast to Stone and Hollenbeck (1989), they concluded that for a categorical moderator, the "test of choice" is the test for equality of correlations across the moderator categories, provided that the hypotheses of equal correlations and equal slopes are approximately identical.

These hypotheses are equivalent if and only if the ratio of the variance in X to the variance in Y is equal across moderator categories (Arnold, 1982; Alexander and DeShon, 1994). The reason for this is clear from the textbook equation between correlations and unstandardized regression coefficients. For the ith category of the moderator,

$$\beta_{yi} = \rho_i \frac{\sigma_{yi}}{\sigma_{xi}} \tag{1}$$

For example, a simple algebraic argument shows that if the σ_{yi}/σ_{xi} ratio is not constant for, say, $i=1$ and $i=2$ then $\beta_1 = \beta_2 \Rightarrow \rho_1 \neq \rho_2$, and likewise $\rho_1 = \rho_2 \Rightarrow \beta_1 \neq \beta_2$. More generally,

$$\frac{\sigma_{y1}\sigma_{x2}}{\sigma_{x1}\sigma_{y2}} > (<)\, 1 \Leftrightarrow \left|\frac{\beta_1}{\beta_2}\right| > (<) \left|\frac{\rho_1}{\rho_2}\right|. \tag{2}$$

The condition for correlations and slopes to be moderated in opposite directions follows immediately: $\beta_1 > \beta_2$ but $\rho_2 > \rho_1$ if when $\rho_2 > \rho_1$, it is also true that

$$\frac{\sigma_{y1}\sigma_{x2}}{\sigma_{x1}\sigma_{y2}} > \frac{\rho_2}{\rho_1}.$$

The same implication holds if the inequalities are changed from $>$ to $<$.

The position taken in this paper is that in multiple linear regression there are three distinct and valid types of moderator effects. First, in multiple regression equation (1) generalizes to a version where standardized regression coefficients replace correlation coefficients:

$$\beta_{yi} = B_{yi} \frac{\sigma_{yi}}{\sigma_{xi}} \tag{3}$$

where B_{yi} is a standardized regression coefficient. Thus, we have moderation of unstandardized versus standardized regression coefficients (or correlations when there is only one predictor), which are equivalent if and only if the aforementioned variance ratio is equal across moderator categories. Otherwise, the assumption that moderation of one implies equivalent moderation of the other is mistaken. This is a simple generalization of Arnold's (1982) and Sharma et al.'s (1981) distinction.

Second, the semi-partial correlation coefficient, ν_{xi}, is a simple function of B_{yi} and tolerance. In the ith moderator category, the tolerance of a predictor, X, is $T_{xi} = 1 - R^2_{xi}$, where R^2_{xi} is the squared multiple correlation for X regressed on the other predictors included in the multiple regression model. The standardized regression coefficient, semi-partial correlation, and tolerance are related by

$$\nu_{xi} = B_{yi}\sqrt{T_{xi}}.$$

Equation (3) therefore may be rewritten as

$$\beta_{yi} = \nu_{xi} \frac{\sigma_{yi}}{\sigma_{xi}\sqrt{T_{xi}}}. \tag{4}$$

Thus, we have a distinction between the moderation of the unique contribution of a predictor to the explained variance of a dependent variable and moderation of regression coefficients (whether standardized or not). Equivalence with moderation of standardized coefficients (or simple correlations) hinges on whether tolerance is constant across moderator categories (an issue not dealt with in this paper), while equivalence with moderation of unstandardized coefficients depends on both constant tolerance and constant variance ratios.

In a later paper, DeShon and Alexander (1996) proposed alternative procedures for testing equality of regression slopes under HeEV, but they and both earlier and subsequent researchers appear to have neglected the idea of testing for equal variance ratios (EVR) across moderator categories. This is understandable, given that HeEV is a more general concern in some respects and the primary object of most regression (and ANOVA) models is prediction.

Nevertheless, it is possible for HeEV to be satisfied when EVR is not. An obvious example is when there is HoV for Y and equality of correlations across moderator categories but HeV for X. These conditions entail HeEV but also imply that slopes cannot be equal across categories. This case seems to have been largely overlooked in the literature on moderators. More generally, HeEV is ensured when, for all i and j,

$$\frac{\sigma^2_{yi}}{\sigma^2_{yj}} = \frac{1-\rho^2_j}{1-\rho^2_i}, \tag{5}$$

which clearly has no bearing on whether EVR holds or not.

Thus, a test of EVR would provide a guide for determining when equality of slopes and equality of correlations are equivalent null hypotheses and when not. Given that it is not uncommon for researchers to be interested in both moderation of slopes (or means) and moderation of correlations, this test could be a useful addition to data screening procedures.

It might seem that if researchers are going to test for both moderation of slopes and correlations, a test of EVR is superfluous. However, the joint outcome of the tests of equal correlations and equal slopes does not render the question of EVR moot or irrelevant. The reason this should interest applied researchers is that the tests of equal correlations and equal slopes will not inform them of whether the moderation of slopes is equivalent to the moderation of correlations, whereas a test of EVR would do exactly that. Suppose, for example, the test for equality of slopes yields $p = 0.04$ (so we reject the null hypothesis) whereas the corresponding test for correlations yields $p = 0.06$ (so we fail to reject). An EVR test would tell us whether these two outcomes are genuinely unequal or whether their apparent difference may be illusory. Thus, an EVR test logically should take place *before* tests of equality of slopes or correlations, because it will indicate whether both of the latter tests need to be conducted or just one will suffice.

Furthermore, an estimate of the ratio of the variance ratios along with its standard error provides an estimate of (and potentially a confidence interval for) a ratio comparison between moderation of slopes and moderation of correlations. From equations (1) and (2), for the ith and jth moderator categories, we immediately have

$$\frac{\sigma_{yi}/\sigma_{xi}}{\sigma_{yj}/\sigma_{xj}} = \frac{\beta_{yi}/\beta_{yj}}{\rho_i/\rho_j}. \tag{6}$$

Finally, equation (3) tells us that an EVR test can be used to assess the equivalence between the moderation of standardized and unstandardized regression coefficients, thereby expanding its domain of application into multiple regression.

All said and done, it is concerning that numerous articles in the foremost journals in psychology routinely report tests of interactions in ANOVAs, ANCOVAs, and regressions with no mention of prior testing for either HeV or HeEV. Moreover, reviews of the literature on metric invariance by Vandenberg and Lance (2000) and DeShon (2004) indicated considerable disagreement on the importance of HeEV for assessments of measurement invariance across samples in structural equations models. Researchers are unlikely to be strongly motivated to use a test for EVR unless it is simple, readily available in familiar computing environments, robust, and powerful. We investigate such a test with these criteria in mind.

A TEST OF EVR FOR CATEGORICAL MODERATORS

An obvious candidate for a test of EVR is a parametric test based on the log-likelihood of a bivariate normal distribution for X and Y conditional on a categorical moderator Z. We employ submodels for the standard deviations using the log link. Using the first category of the moderator as the "base" category, the submodels may be written as

$$\sigma_{xi} = \exp\left(\sum_i z_i \delta_{xi}\right), \tag{7}$$
$$\sigma_{yi} = \exp\left(\sum_i z_i \delta_{yi}\right),$$

where $z_1 = 1$ and for $i > 1$ z_i is an indicator variable for the ith category of Z, and the δ parameters are regression coefficients. Under the hypothesis that EVR holds between the ith and first categories, the relevant test statistic is

$$\theta_i = \delta_{yi} - \delta_{xi}, \tag{8}$$

for $i > 1$, with

$$\mathrm{var}(\theta_i) = \mathrm{var}(\delta_{yi}) + \mathrm{var}(\delta_{xi}) - 2\mathrm{cov}(\delta_{yi}, \delta_{xi}), \tag{9}$$

and the assumption that δ_{yi} and δ_{yi} are asymptotically bivariate normally distributed. Immediately we have a confidence interval for θ_i, namely $\hat{\theta}_i \pm t_{\alpha/2}\sqrt{\hat{\mathrm{var}}(\theta_i)}$, where $t_{\alpha/2}$ is the $1-\alpha/2$ quantile of the t distribution with the appropriate degrees of freedom for an independent-samples test. We also have

$$\exp(\theta_i) = \frac{\beta_{yi}/\beta_{y1}}{\rho_i/\rho_1}, \tag{10}$$

and we may exponentiate the limits of this confidence interval to obtain a confidence interval for the right-hand expression in this equation, i.e., for the ratio comparison between the ratio of moderated regression coefficients and the ratio of moderated correlations.

The hypothesis that $\theta_i = 0$ is equivalent to a restricted model in which, for $i > 1$, $\delta_{xi} = \delta_{yi}$. The modeling approaches outlined later in this paper make use of this equivalence. More complex EVR hypotheses may require different design matrices from the setup proposed in this introductory treatment. First, however, we shall examine the properties of θ, including Type I error rates and power under unequal sample sizes, and the effects of departures from normality for X and Y.

ASSESSING TYPE I ERROR ACCURACY AND POWER

We begin with simulations testing null hypothesis rejection rates for EVR when the null hypotheses of EVR and unmoderated correlations and slopes are true. Simulations using a two-category moderator (20,000 runs for each condition) were based on DeShon and Alexander, 1996; **Table 1**), with constant variance ratio of 2, $\rho_{xy} = 1/\sqrt{2}$, and $\beta_y = 1$ for both categories. Three pairs of sample sizes were used (again based on DeShon and Alexander, 1996): 70 for both samples, 45 for one and 155 for the second, and 90 for one and 180 for the second. Three pairs of variances also were used, to ascertain any impact from the sizes of the variances. All runs used a Type I error criterion of $\alpha = 0.05$.

The top half of **Table 1** shows the EVR rejection rates for random samples from normally distributed X and Y. Unequal sample

Table 1 | Type I error: two-groups simulations.

Skew	N_1	N_2	$\sigma_{xi} = 1,1$ $\sigma_{yi} = 2,2$	$\sigma_{xi} = 1,2$ $\sigma_{yi} = 2,4$	$\sigma_{xi} = 1,4$ $\sigma_{yi} = 2,8$
NORMAL					
0	70	70	0.0518	0.0532	0.0511
0	45	155	0.0578	0.0553	0.0547
0	90	180	0.0513	0.0531	0.0503
SKEWED					
2	70	70	0.0710	0.0704	0.0680
4	70	70	0.0768	0.0767	0.0713
2	45	155	0.0681	0.0686	0.0687
4	45	155	0.0778	0.0776	0.0774
2	90	180	0.0679	0.0708	0.0714
4	90	180	0.0755	0.0728	0.0723

sizes have little impact on rejection rates, with the effect appearing to diminish in the larger-sample (90–180) condition. The rates are slightly higher than 0.05, but are unaffected by the sizes of the variances.

The lower half of **Table 1** shows simulations under the same conditions, but this time with X and Y sampled from the Azzalini skew-normal distribution (Azzalini, 1985). The standard skew-normal pdf is

$$f(x, \lambda) = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \left(1 + Erf\left[\frac{\lambda x}{\sqrt{2}}\right]\right).$$

The simulations had the skew parameter λ set to 2 and 4, the pdfs for which are shown in **Figure 1**. Skew increased the rejection rates to 0.068–0.078, rendering the test liberal but not dramatically so.

We now turn to investigating the power of the EVR test. Simulations testing its power were conducted for two situations: moderated slopes but unmoderated correlations, and moderated correlations but unmoderated slopes. Both batches of simulations were run with four combinations of sample sizes (70–70, 40–140, 140–140, and 80–280) and three variance ratio combinations (1–1.5, 1–2, 1–4). In the unmoderated correlations setup $\rho = 0.5$ for all conditions, and in the unmoderated slopes setup $\beta_y = 0.5$ for all conditions. These tests also require modeling the moderation of correlations. The correlation submodel uses the Fisher link, i.e.

$$\log\left(\frac{1+\rho_i}{1-\rho_i}\right) = \sum_i w_i \delta_{ri}. \quad (11)$$

Note that we allow a different set of predictors for the correlation from those in equation (7). However, in this paper we will impose the restriction $w_i = z_i$.

Table 2 shows the simulation results for unequal variance ratios with unmoderated correlations. The table contains rejection rates of the EVR and moderation of correlation null hypotheses. The resultant moderated slopes and error-variances are displayed for each condition. Note that HeV and HeEV do not have discernible effects on either of the rejection rates. As in the preceding simulations, the rejection rates for the unmoderated correlations are only slightly above the 0.05 criterion. The rejection rates for the EVR test l and in the 0.85–1.0 range in the conditions where the combined sample sizes are 280 and the ratio of the variance ratios is 2:1 or for both combined sizes when the ratio is 4:1.

Table 3 shows the rejection rates of the EVR and moderation of correlation null hypotheses when there are unequal variance ratios and moderated correlations. The resultant moderated correlations and error-variances are displayed for each condition. As before, HeV and HeEv do not affect either of the rejection rates. Likewise, as expected, the EVR rejection rates are very similar to those in **Table 2**. It is noteworthy that rejection rates for the unmoderated correlations hypothesis are considerably smaller than those for the EVR hypothesis, even though the correlations differ fairly substantially. It is well-known that tests for moderation of slopes and correlations have rather low power. These results, and the fact that the ratios of the variance ratios do not exceed Howell's benchmark of 4:1, suggest that the EVR test has relatively high power.

STRUCTURAL EQUATIONS MODEL APPROACH

When the moderator variable is categorical, the EVR test can be incorporated in a structural equations model (SEM) approach that permits researchers not only to compare an EVR model against one that relaxes this assumption, but also to test simultaneously for HeV, HeEV, moderation of correlations and moderation of slopes. **Figure 2** shows the regression (left-hand side) and correlation (right-hand side) versions of this model. The latter follows Preacher's (2006) strategy for a multi-group SEM for correlations. The regression version models the error-variances σ^2_{ei} rather than the variances σ^2_{yi}. Instead, σ^2_{yi} is modeled in the correlation version. The only addition to the correlation SEM required for incorporating EVR tests is to explicitly model variance ratios for each of the moderator variable categories. Two SEM package that can do so are lavaan (Rosseel, 2012) in R and MPlus (Muthén and Muthén, 2010). Examples in lavaan and MPlus are available at http://dl.dropbox.com/u/1857674/EVR_moderator/EVR.html, as are EVR test scripts in SPSS and SAS.

Simulations were run using lavaan in model comparisons for samples with moderated slopes but unmoderated correlations, and samples with moderated correlations but unmoderated slopes. As before, each simulation had 20,000 runs.

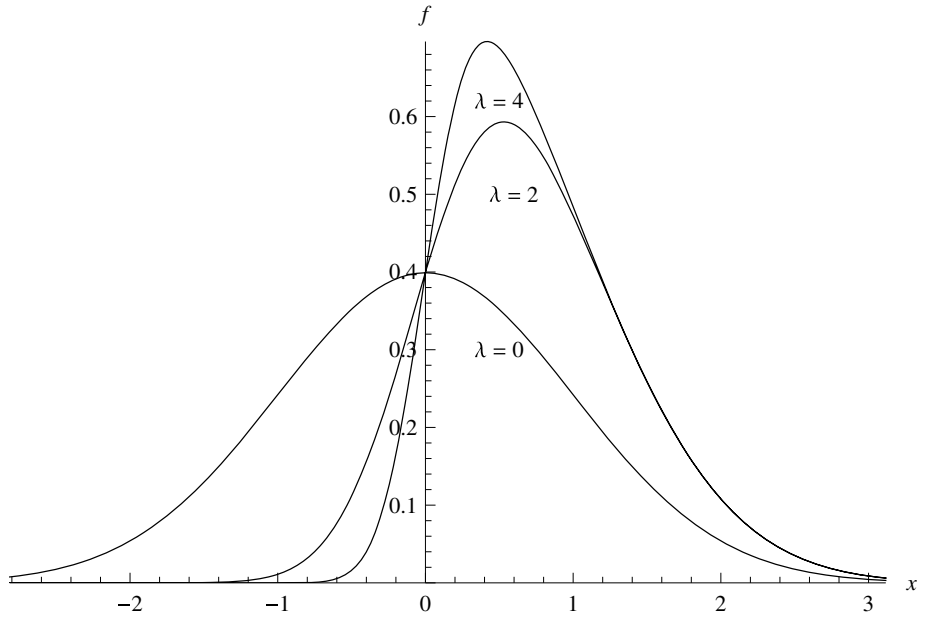


FIGURE 1 | Azzalini Skew-normal distributions with $\lambda = 0,2,4$.

Table 2 | Power: moderated slopes and unmoderated correlations.

N_1	N_2	$\sigma_x^2 = 2$ $\sigma_y^2 = 2$ $\sigma_{xy} = 1$ $\sigma_e^2 = 1.5$ $\beta_y = 0.5$ δ_r	$\sigma_x^2 = 2$ $\sigma_y^2 = 3$ $\sigma_{xy} = \sqrt{3/2}$ $\sigma_e^2 = 2.25$ $\beta_y = \sqrt{3/8}$ θ	$\sigma_x^2 = 2$ $\sigma_y^2 = 2$ $\sigma_{xy} = 1$ $\sigma_e^2 = 1.5$ $\beta_y = 0.5$ δ_r	$\sigma_x^2 = 2$ $\sigma_y^2 = 4$ $\sigma_{xy} = \sqrt{2}$ $\sigma_e^2 = 3$ $\beta_y = \sqrt{2}/2$ θ	$\sigma_x^2 = 2$ $\sigma_y^2 = 2$ $\sigma_{xy} = 1$ $\sigma_e^2 = 1.5$ $\beta_y = 0.5$ δ_r	$\sigma_x^2 = 2$ $\sigma_y^2 = 8$ $\sigma_{xy} = 2$ $\sigma_e^2 = \sqrt{8}/2$ $\beta_y = 1$ θ
70	70	0.0556	0.2810	0.0603	0.6321	0.0576	0.9939
40	100	0.0566	0.2478	0.0569	0.5706	0.0566	0.9875
140	140	0.0549	0.4841	0.0532	0.9032	0.0537	1.000
80	200	0.0529	0.4311	0.0497	0.8524	0.0522	0.9999

Table 3 | Power: unmoderated slopes and moderated correlations.

N_1	N_2	$\sigma_x^2 = 2$ $\sigma_y^2 = 2$ $\sigma_{xy} = 1$ $\sigma_e^2 = 1.5$ $\rho_{xy} = 0.5$ δ_r	$\sigma_x^2 = 2$ $\sigma_y^2 = 3$ $\sigma_{xy} = 1$ $\sigma_e^2 = 2.5$ $\rho_{xy} = 1/\sqrt{6}$ θ	$\sigma_x^2 = 2$ $\sigma_y^2 = 2$ $\sigma_{xy} = 1$ $\sigma_e^2 = 1.5$ $\rho_{xy} = 0.5$ δ_r	$\sigma_x^2 = 2$ $\sigma_y^2 = 4$ $\sigma_{xy} = 1$ $\sigma_e^2 = 3.5$ $\rho_{xy} = 1/\sqrt{8}$ θ	$\sigma_x^2 = 2$ $\sigma_y^2 = 2$ $\sigma_{xy} = 1$ $\sigma_e^2 = 1.5$ $\rho_{xy} = 0.5$ δ_r	$\sigma_x^2 = 2$ $\sigma_y^2 = 8$ $\sigma_{xy} = 1$ $\sigma_e^2 = 6$ $\rho_{xy} = 0.25$ θ
70	70	0.0944	0.2635	0.1525	0.5925	0.3426	0.9864
40	100	0.0992	0.2394	0.1700	0.5476	0.3558	0.9826
140	140	0.1296	0.4575	0.2483	0.8771	0.5844	1.000
80	200	0.1444	0.4201	0.2878	0.8326	0.6036	0.9999

Simulations from bivariate normal distributions with $\rho_{xy} = 0.05$ for both groups (**Table 4**) indicated that moderately large samples and slope differences are needed for reasonable power. However, there was little impact on power from unequal group sizes. Rejection rates for the unmoderated correlations hypothesis were at appropriate levels, 0.0493–0.0559.

Likewise, simulations from bivariate normal distributions with $\beta_y = 0.5$ for both groups (**Table 5**) indicated that moderately large samples and correlation differences are needed for reasonable power. There was a slight to moderate impact from unequal group sizes, somewhat greater than the impact in **Table 4**. Rejection rates for the unmoderated slopes hypothesis were appropriately 0.0484–0.0538.

SEM EXAMPLE

Consider a population with two normally distributed variables *X*, political liberalism, and *Y*, degree of belief in global warming. Suppose that they are measured on scales with means of 0 and standard deviations of 1, and the correlation between these two scales is $\rho = 0.45$. Suppose also that if members of this population are exposed to a video debate highlighting the arguments for and against the reality of global warming, it polarizes belief in global warming by increasing the degree of belief of those who already tend to believe it and decreasing the degree of belief of those who already are skeptical. Thus, the standard deviation doubles from 1 to 2. However, the mean remains at 0 and the correlation between belief in global warming and political liberalism also is unchanged, remaining at 0.45.

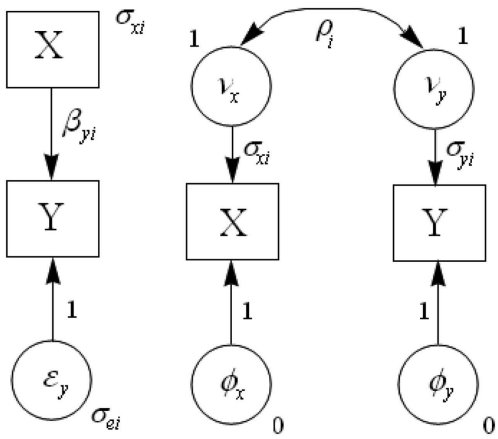


FIGURE 2 | Moderated regression and correlation structural equations models.

In a two-condition experiment with half the participants from this population assigned to a condition where they watch the video and half to a "no-video" condition, the experimental conditions may be regarded as a two-category moderator variable *Z*. We have $\rho = 0.45$ and $\sigma_x = 1$ regardless of *Z*, and $\sigma_{y1} = 2$ whereas $\sigma_{y2} = 1$. It is also noteworthy that when *X* predicts *Y* HeEV is violated whereas when *Y* predicts *X* it is not.

We randomly sample 600 people from this population and randomly assign 300 to each condition, representing the video condition with $Z = 1$ and the no-video condition with $Z = -1$. As expected, the sample correlations in each subsample do not differ significantly: $r_1 = 0.458$, $r_2 = 0.463$, and Fisher's test yields $z = 0.168$ ($p = 0.433$). However, a linear regression with *Y* predicted by *X* and *Z* that includes an interaction term ($Z \times X$) finds a significant positive interaction coefficient ($z = 3.987$, $p < 0.0001$). Taking the regression on face value could mislead us into believing that because the slope between *X* and *Y* differs significantly between the two categories of *Z*, *Z* also moderates the association between *X* and *Y*. Of course, it does not. Seemingly more puzzling is the fact that linear regression with *Y* predicting *X* yields a significant *negative* interaction term ($Z \times Y$) with $z = -3.859$ ($p = 0.0001$). So the regression coefficient is moderated in opposite directions, depending on whether we predict *Y* or *X*.

The scatter plots in **Figure 3** provide an intuitive idea of what is going on. Clearly the slope for *Y* (belief in global warming) predicted by *X* (liberalism) appears steeper when $Z = 1$ than when $Z = -1$. Just as clearly, the slope for *X* predicted by *Y* appears less steep when $Z = 1$ than when $Z = -1$. The oval shapes of the

Table 4 | Moderated regression coefficients.

N_1	N_2	$\beta_y = 0.50$ $\beta_y = 0.61$	$\beta_y = 0.50$ $\beta_y = 0.71$	$\beta_y = 0.50$ $\beta_y = 1.00$
70	70	0.1086	0.2030	0.5659
40	100	0.1012	0.2026	0.6031
140	140	0.1633	0.3668	0.8566
80	200	0.1499	0.3549	0.8875

Table 5 | Moderated correlations.

N_1	N_2	$\rho_{xy} = 0.50$ $\rho_{xy} = 0.41$	$\rho_{xy} = 0.50$ $\rho_{xy} = 0.35$	$\rho_{xy} = 0.50$ $\rho_{xy} = 0.25$	$\rho_{xy} = 0.50$ $\rho_{xy} = 0.17$
70	70	0.1159	0.1984	0.4406	0.6390
40	100	0.1031	0.1728	0.3610	0.5487
140	140	0.1760	0.3521	0.7215	0.9008
80	200	0.1541	0.2960	0.6312	0.8395

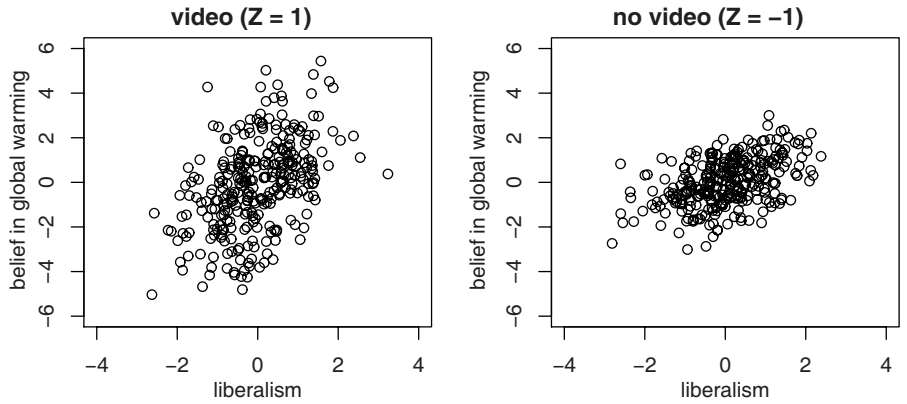


FIGURE 3 | Scatter plots for the two-condition experiment.

data distribution in both conditions appear similar to one another, giving the impression that the correlations are similar.

We now demonstrate that the SEM approach can clarify and validate these impressions, using Mplus 6.12. We begin with the moderation of slopes models. Because $\sigma_{x1} = \sigma_{x2}$ (i.e., X has HoV) we may move from the saturated model to one that restricts those parameters to be equal. The model fit is $\chi^2(1) = 0.370$ ($p = 0.543$). This baseline model also reproduces the slopes estimates in OLS regression. Now, a model removing HoV for X and imposing the EVR restriction yields $\chi^2(1) = 82.246$ ($p < 0.0001$), so clearly we can reject the EVR hypothesis. Fitting another model with HoV in X and HeV in Y but where we set $\beta_{y1} = \beta_{y2}$, the fit is $\chi^2(2) = 15.779$ ($p = 0.0004$), and the model comparison test is $\chi^2(1) = 15.779 - 0.370 = 15.409$ ($p < 0.0001$). We conclude there is moderation of slopes but EVR does not hold, so we expect that the moderation of correlations will differ from that of the slopes, and the moderation of slopes will differ when X predicts Y versus when Y predicts X. Indeed, if we fit models with Y predicting X we also can reject the equal slopes model, and the slopes differ in opposite directions across the categories of Z. When X predicts Y $\beta_{y1} = 0.496$ and $\beta_{y2} = 0.978$, whereas when Y predicts X $\beta_{x1} = 0.219$ and $\beta_{x2} = 0.423$.

Turning to correlations, we start with a model that sets $\sigma_{x1} = \sigma_{x2}$ (i.e., assuming that X has HoV) and leaves all other parameters free. The fit is $\chi^2(1) = 0.370$ ($p = 0.543$), identical to the equivalent baseline model described above. This model closely reproduces the sample correlations (the parameter estimates are 0.452 and 0.469, versus the sample correlations 0.458 and 0.463). Moreover, a model adding the EVR restriction yields $\chi^2(1) = 82.246$, again identical to the equivalent regression model. Now if we set $\rho_1 = \rho_2$, the fit is $\chi^2(2) = 0.453$ ($p = 0.797$) and the model comparison test is $\chi^2(1) = 0.083$ ($p = 0.773$). Thus, there is moderation of slopes but not of correlations.

CONTINUOUS MODERATORS

Continuous moderators pose considerably greater challenges than categorical ones, because of the many forms that HeV and HeEV can take. Arnold (1982) sketched out a treatment of this problem that is not satisfactory, namely correlating correlations between X and Y with values of the continuous moderator Z. In an innovative paper, Allison et al. (1992) extended a standard approach to assessing heteroscedasticity to test for homologizers when the moderator variable, Z, is continuous. Their technique is simply to compute the correlation between Z and the absolute value of the residuals from the regression equation that already includes both the main effect for Z and the interaction term. This is a model of moderated error, akin to modeling error-variance, which is useful in itself but not equivalent to testing for EVR. In their approach and the simulations that tested it, Allison et al. assumed HeV for their predictor, thereby ignoring the fact that EVR can be violated even when HeEV is satisfied.

The approach proposed here generalizes the model defined by equations (7) and (11), with the z_i now permitted to be continuous. This model is

$$\begin{aligned} \log(\sigma_x) &= \sum_i z_i \delta_{xi}, \\ \log(\sigma_y) &= \sum_i z_i \delta_{yi}, \\ \log\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) &= \sum_i z_i \delta_{ri} \end{aligned} \tag{12}$$

where $z_1 = 1$ and for $i > 1$ the z_i are continuous random variables. The δ_{xi}, δ_{yi}, and δ_{ri} coefficients can be simultaneously estimated via maximum likelihood, using the likelihood function of a bivariate normal distribution conditioned by the z_i. Scripts for maximum likelihood estimation in R, SPSS, and SAS are available via the link cited earlier. This model can be made more flexible by introducing polynomial terms in the z_i, but we do not undertake that extension here.

To begin, simulations (20,000 runs each) for a single-moderator model took samples for Z from a $N(0, 1)$ population. X and Y were sampled from bivariate normal distributions with $\delta_{r1} = 0$, $\delta_{x1} = \delta_{y1} = \{0, 0.5, 1.0\}$, and $\delta_{r0} = \{0, 0.5, 1.0\}$. **Table 6** displays their results. Rejection rates are somewhat too high for δ_{r1} but only slightly too high for θ_1 unless sample sizes are over 200 or so.

Simulations also were run under the same conditions as **Table 6** but with samples from a skew-normal distribution with skew parameter $\lambda = 2$. These results are shown in **Table 7**. There, it can be seen that Type I error rates are inflated by skew almost independently of sample size, much more so for δ_{r1} than θ_1. Both are affected by size of the correlation's moderation effect.

To investigate power, simulations were run with $\delta_{r1} = \{0, 0.2007, 0.6190, 1.0986, 1.7346\}$ (correlation differences of {0,0.1,0.3,0.5,0.7} when $z = 1$) and $\theta_1 = \{0.1116, 0.2027, 0.3466, 0.5493, 0.6931, 0.8047\}$ (variance ratios of {1.25, 1.5, 2, 3, 4, 5} when $z = 1$). Thus, there were 30 simulations for each of three sample sizes (70,140, and 280). The results are displayed in **Figure 4**. Power for θ_1 attains high levels even for moderate sample sizes when the variance ratio is 2 or more. However, power also is higher the more strongly correlations are moderated, whereas power for δ_{r1} is unaffected by moderation of the variance ratio. Power for δ_{r1} does not become high unless correlations differ by at least 0.3, and the results for a correlation difference of 0.1 are in line with those for categorical moderators (see **Table 3**).

The simulation results were examined for evidence of estimation bias. Both $\hat{\delta}_{r1}$ and θ_1 were slightly biased upward, and most strongly for smaller samples and larger effect-sizes. The maximum average bias for $\hat{\delta}_{r1}$ and $\hat{\theta}_1$ was 0.04 and 0.03 respectively. For both estimators, doubling the sample size approximately halved the bias.

Table 6 | Unmoderated continuous moderator simulations.

N	$\delta_{r0} = 0.0$	$\delta_{r0} = 0.5$	$\delta_{r0} = 1.0$
δ_{r1}			
70	0.0715	0.0712	0.0685
140	0.0619	0.0589	0.0571
280	0.0543	0.0554	0.0545
θ_1			
70	0.0610	0.0627	0.0616
140	0.0548	0.0564	0.0556
280	0.0533	0.0536	0.0528

Table 7 | Simulations from Azzalini distribution with $\lambda = 2$.

N	$\delta_{r0} = 0.0$	$\delta_{r0} = 0.5$	$\delta_{r0} = 1.0$
δ_{r1}			
70	0.0724	0.0874	0.1001
140	0.0682	0.0801	0.0925
280	0.0665	0.0767	0.0930
θ_1			
70	0.0554	0.0673	0.0679
140	0.0519	0.0689	0.0645
280	0.0514	0.0676	0.0636

DISCUSSION

This paper has introduced a simple test of equal variance ratios (EVR), whose purpose is to determine when moderation of correlations and slopes are not equivalent. The test can be inverted to produce an approximate confidence interval for the ratio comparison of these two kinds of moderator effects. This test also may be extended easily to assessing whether the moderation of standardized and unstandardized regression coefficients are unequal.

Simulation results indicated that when EVR holds, Type I error rates are reasonably accurate but slightly high. Skew inflates Type I error rates somewhat, but not dramatically. When EVR does not hold, moderately large samples and effect-sizes are needed for high power, but HeV, HeEV, and unequal group sizes are not problematic for testing EVR or modeling the moderation of variance ratios. There is evidence that the EVR test has fairly high power, relative to the power to detect moderator effects.

Variance ratios for continuous moderators can be modeled via maximum likelihood methods, although no single model can deal with all forms of variance ratio moderation or HeV. The model presented here uses the log link for the standard deviation submodel and the Fisher link for the correlation submodel, with possibly different predictors in each submodel and, potentially, polynomial terms for the predictors. Bayesian estimation methods

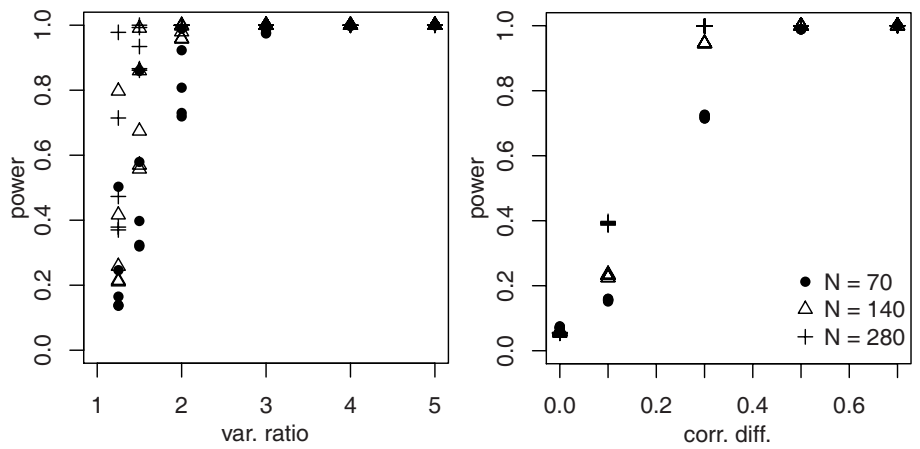


FIGURE 4 | Power for δ_{r1} and θ_{r1}.

also may be used, but that extension is beyond the scope of this paper. When EVR holds and correlations are unmoderated, Type I error rates are somewhat too high for δ_{r1} and slightly too high for θ_1 unless sample sizes are over 200 or so. Skew inflates Type I error rates for δ_{r1} but only slightly for θ_1. For moderated variance ratios and correlations, maximum likelihood estimates are only slightly upward-biased for both δ_{r1} and θ_1, and in the usual fashion this bias decreases with increasing sample size. Moderately large samples and effect-sizes are needed for high power, but apparently no more so than for categorical moderators.

Tests of EVR for categorical moderators can be entirely dealt with using multi-groups SEM, and Mplus and the lavaan package in R are able to incorporate these tests via appropriate model comparisons. It also is possible to fit such models via scripts in computing environments such as SAS and SPSS possessing appropriate inbuilt optimizers. The SEM approach makes it possible to test complex hypotheses regarding the (non)equivalence of moderation of slopes and correlations, and to obtain a clear picture of both kinds of moderator effects. The online supplementary material for this paper includes a four-category moderator example where EVR holds for two pairs of categories but not for all four. In fact, the SEM approach elaborates conventional moderated regression into a combination of models for moderated slopes and moderated correlations. In principle it may be extended to incorporate tests for equality of tolerance across groups, which would enable modeling the moderation of semi-partial correlations.

All told, for categorical moderators the EVR test comes reasonably close to fulfilling the criteria of simplicity, availability, robustness and power. Considerable work remains to be done before the same can be said for continuous moderators. Nevertheless, the EVR test proposed here is highly relevant for both experimental and non-experimental research in mainstream psychology, and would seem to be a worthy addition to the researcher's toolkit.

ACKNOWLEDGMENTS

I thank the two anonymous reviewers for their informative and helpful criticisms and suggestions regarding this paper. Any remaining errors or lacunae are my responsibility.

REFERENCES

Alexander, R. A., and DeShon, R. P. (1994). The effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychol. Bull.* 115, 308–314.

Allison, D. B., Heshka, S., Pierson, R. N. J., Wang, J., and Heymsfield, S. B. (1992). The analysis and identification of homologizer/moderator variables when the moderator is continuous: an illustration with anthropometric data. *Am. J. Hum. Biol.* 4, 775–782.

Arnold, H. J. (1982). Moderator variables: a clarification of conceptual, analytic, and psychometric issues. *Organ. Behav. Hum. Perform.* 29, 143–174.

Arnold, H. J. (1984). Testing moderator variable hypotheses: a reply to Stone and Hollenbeck. *Organ. Behav. Hum. Perform.* 34, 214–224.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stat.* 12, 171–178.

Correa, J. C., Iral, R., and Rojas, L. (2006). Estudio de potencias de prueba de homogeneidad de varianza (a study on homogeneity of variance tests). *Revista Colombiana de Estadistica* 29, 57–76.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychol. Sci.* 46, 137–149.

DeShon, R. P., and Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychol. Methods* 1, 261–277.

Dretzke, B. J., Levin, J. R., and Serlin, R. C. (1982). Testing for regression homogeneity under variance heterogeneity. *Psychol. Bull.* 91, 376–383.

Grissom, R. (2000). Heterogeneity of variance in clinical data. *J. Consult. Clin. Psychol.* 68, 155–165.

Howell, D. (2007). *Statistical Methods for Psychology*, 6th Edn. Belmont, CA: Thomson Wadsworth.

Muthén, L., and Muthén, B. (2010). *Mplus User's Guide*, 6th Edn. Los Angeles, CA: Muthén and Muthén.

Preacher, K. (2006). Testing complex correlational hypotheses with structural equations models. *Struct. Equ. Modeling* 13, 520–543.

Rosseel, Y. (2012). *lavaan: Latent Variable Analysis.* R package version 0.4–13. Available at: http://CRAN.R-project.org/package=lavaan

Schmider, E., Ziegler, M., Danay, E., Beyer, L., and Bühner, M. (2010). Is it really robust? reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology (Gott)* 6, 147–151.

Sharma, S., Durand, R., and Gur-Arie, O. (1981). Identification and analysis of moderator variables. *J. Mark. Res.* 18, 291–300.

Stone, E. F., and Hollenbeck, J. R. (1984). Some issues associated with the use of moderated regression. *Organ. Behav. Hum. Perform.* 34, 195–213.

Stone, E. F., and Hollenbeck, J. R. (1989). Clarifying some controversial issues surrounding statistical procedures for detecting moderator variables: empirical evidence and related matters. *J. Appl. Psychol.* 74, 3–10.

Tabachnick, B., and Fidell, L. (2007). *Experimental Design Using ANOVA*. Belmont, CA: Duxbury.

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Meth.* 3, 4–69.

Zedeck, S. (1971). Problems with the use of "moderator" variables. *Psychol. Bull.* 76, 295–310.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 May 2012; accepted: 19 June 2012; published online: 09 July 2012.
Citation: Smithson M (2012) A simple statistic for comparing moderation of slopes and correlations. Front. Psychology ***3****:231. doi: 10.3389/fpsyg.2012.00231*
This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 01 March 2012
doi: 10.3389/fpsyg.2012.00055

Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis

David B. Flora*, Cathy LaBrish and R. Philip Chalmers

Department of Psychology, York University, Toronto, ON, Canada

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Stanislav Kolenikov, University of Missouri, USA
Cam L. Huynh, University of Manitoba, Canada

****Correspondence:***
David B. Flora, Department of Psychology, York University, 101 BSB, 4700 Keele Street, Toronto, ON M3J 1P3, Canada.
e-mail: dflora@yorku.ca

We provide a basic review of the data screening and assumption testing issues relevant to exploratory and confirmatory factor analysis along with practical advice for conducting analyses that are sensitive to these concerns. Historically, factor analysis was developed for explaining the relationships among many continuous test scores, which led to the expression of the common factor model as a multivariate linear regression model with observed, continuous variables serving as dependent variables, and unobserved factors as the independent, explanatory variables. Thus, we begin our paper with a review of the assumptions for the common factor model and data screening issues as they pertain to the factor analysis of continuous observed variables. In particular, we describe how principles from regression diagnostics also apply to factor analysis. Next, because modern applications of factor analysis frequently involve the analysis of the individual items from a single test or questionnaire, an important focus of this paper is the factor analysis of items. Although the traditional linear factor model is well-suited to the analysis of continuously distributed variables, commonly used item types, including Likert-type items, almost always produce dichotomous or ordered categorical variables. We describe how relationships among such items are often not well described by product-moment correlations, which has clear ramifications for the traditional linear factor analysis. An alternative, non-linear factor analysis using polychoric correlations has become more readily available to applied researchers and thus more popular. Consequently, we also review the assumptions and data-screening issues involved in this method. Throughout the paper, we demonstrate these procedures using an historic data set of nine cognitive ability variables.

Keywords: exploratory factor analysis, confirmatory factor analysis, item factor analysis, structural equation modeling, regression diagnostics, data screening, assumption testing

Like any statistical modeling procedure, factor analysis carries a set of assumptions and the accuracy of results is vulnerable not only to violation of these assumptions but also to disproportionate influence from unusual observations. Nonetheless, the importance of data screening and assumption testing is often ignored or misconstrued in empirical research articles utilizing factor analysis. Perhaps some researchers have an overly indiscriminate impression that, as a large sample procedure, factor analysis is "robust" to assumption violation and the influence of unusual observations. Or, researchers may simply be unaware of these issues. Thus, with the applied psychology researcher in mind, our primary goal for this paper is to provide a general review of the data screening and assumption testing issues relevant to factor analysis along with practical advice for conducting analyses that are sensitive to these concerns. Although presentation of some matrix-based formulas is necessary, we aim to keep the paper relatively non-technical and didactic. To make the statistical concepts concrete, we provide data analytic demonstrations using different factor analyses based on an historic data set of nine cognitive ability variables.

First, focusing on factor analyses of continuous observed variables, we review the common factor model and its assumptions and show how principles from regression diagnostics can be applied to determine the presence of influential observations. Next, we move to the analysis of categorical observed variables, because treating ordered, categorical variables as continuous variables is an extremely common form of assumption violation involving factor analysis in the substantive research literature. Thus, a key aspect of the paper focuses on how the linear common factor model is not well-suited to the analysis of categorical, ordinally scaled item-level variables, such as Likert-type items. We then describe an alternative approach to item factor analysis based on polychoric correlations along with its assumptions and limitations. We begin with a review the linear common factor model which forms the basis for both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

THE COMMON FACTOR MODEL

The major goal of both EFA and CFA is to *model* the relationships among a potentially large number of observed variables using a smaller number of unobserved, or latent, variables. The latent variables are the factors. In EFA, a researcher does not have a strong prior theory about the number of factors or how each observed variable relates to the factors. In CFA, the number of factors is hypothesized *a priori* along with hypothesized relationships between factors and observed variables. With both EFA and CFA, the factors influence the observed variables to account for

their variation and covariation; that is, covariation between any two observed variables is due to them being influenced by the same factor. This idea was introduced by Spearman (1904) and, largely due to Thurstone (1947), evolved into the *common factor model*, which remains the dominant paradigm for factor analysis today. Factor analysis is traditionally a method for fitting models to the bivariate associations among a set of variables, with EFA most commonly using Pearson product-moment correlations and CFA most commonly using covariances. Use of product-moment correlations or covariances follows from the fact that the common factor model specifies a linear relationship between the factors and the observed variables.

Lawley and Maxwell (1963) showed that the common factor model can be formally expressed as a linear model with observed variables as dependent variables and factors as explanatory or independent variables:

$$y_j = \lambda_1\eta_1 + \lambda_2\eta_2 + \lambda_k\eta_k + \varepsilon_j$$

where y_j is the jth observed variable from a battery of p observed variables, η_k is the kth of m common factors, λ_k is the regression coefficient, or *factor loading*, relating each factor to y_j, and ε_j is the residual, or *unique factor*, for y_j. (Often there are only one or two factors, in which case the right hand side of the equation includes only $\lambda_1\eta_1 + \varepsilon_j$ or only $\lambda_1\eta_1 + \lambda_2\eta_2 + \varepsilon_j$.) It is convenient to work with the model in matrix form:

$$\mathbf{y} = \mathbf{\Lambda\eta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is a vector of the p observed variables, $\mathbf{\Lambda}$ is a $p \times m$ matrix of factor loadings, $\boldsymbol{\eta}$ is a vector of m common factors, and $\boldsymbol{\varepsilon}$ is a vector of p unique factors[1]. Thus, each common factor may influence more than one observed variable while each unique factor (i.e., residual) influences only one observed variable. As with the standard regression model, the residuals are assumed to be independent of the explanatory variables; that is, all unique factors are uncorrelated with the common factors. Additionally, the unique factors are usually assumed uncorrelated with each other (although this assumption may be tested and relaxed in CFA).

Given Eq. 1, it is straightforward to show that the covariances among the observed variables can be written as a function of model parameters (factor loadings, common factor variances and covariances, and unique factor variances). Thus, in CFA, the parameters are typically estimated from their covariance structure:

$$\mathbf{\Sigma} = \mathbf{\Lambda\Psi\Lambda'} + \mathbf{\Theta}, \tag{2}$$

where $\mathbf{\Sigma}$ is the $p \times p$ population covariance matrix for the observed variables, $\mathbf{\Psi}$ is the $m \times m$ interfactor covariance matrix, and $\mathbf{\Theta}$ is the $p \times p$ matrix unique factor covariance matrix that often contains only diagonal elements, i.e., the unique factor variances. The covariance structure model shows that the observed covariances are a function of the parameters but not the unobservable scores on the common or unique factors; hence, it is not necessary to observe scores on the latent variables to estimate the model parameters. In EFA, the parameters are most commonly estimated from the correlation structure

$$\mathbf{P} = \mathbf{\Lambda}^*\mathbf{\Psi}^*\mathbf{\Lambda}^{*\prime} + \mathbf{\Theta}^* \tag{3}$$

where $\mathbf{P}$ is the population correlation matrix and, in that $\mathbf{P}$ is simply a re-scaled version of $\mathbf{\Sigma}$, we can view $\mathbf{\Lambda}^*$, $\mathbf{\Psi}^*$, and $\mathbf{\Theta}^*$ as re-scaled versions of $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and $\mathbf{\Theta}$, respectively. This tendency to conduct EFA using correlations is mainly a result of historical tradition, and it is possible to conduct EFA using covariances or CFA using correlations. For simplicity, we focus on the analysis of correlations from this point forward, noting that the principles we discuss apply equivalently to the analysis of both correlation and covariance matrices (MacCallum, 2009; but see Cudeck, 1989; Bentler, 2007; Bentler and Savalei, 2010 for discussions of the analysis of correlations vs. covariances). We also drop the asterisks when referring to the parameter matrices in Eq. 3.

Jöreskog (1969) showed how the traditional EFA model, or an "unrestricted solution" for the general factor model described above, can be constrained to produce the "restricted solution" that is commonly understood as today's CFA model and is well-integrated in the structural equation modeling (SEM) literature. Specifically, in the EFA model, the elements of $\mathbf{\Lambda}$ are all freely estimated; that is, each of the m factors has an estimated relationship (i.e., factor loading) with every observed variable; factor rotation is then used to aid interpretation by making some values in $\mathbf{\Lambda}$ large and others small. But in the CFA model, depending on the researcher's hypothesized model, many of the elements of $\mathbf{\Lambda}$ are restricted, or constrained, to equal zero, often so that each observed variable is determined by one and only one factor (i.e., so that there are no "cross-loadings"). Because the common factors are unobserved variables and thus have an arbitrary scale, it is conventional to define them as standardized (i.e., with variance equal to one); thus $\mathbf{\Psi}$ is the interfactor correlation matrix[2]. This convention is not a testable assumption of the model, but rather imposes necessary identification restrictions that allow the model parameters to be estimated (although alternative identification constraints are possible, such as the marker variable approach often used with CFA). In addition to constraining the factor variances, EFA requires a diagonal matrix for $\mathbf{\Theta}$, with the unique factor variances along the diagonal.

Exploratory factor analysis and CFA therefore share the goal of using the common factor model to represent the relationships among a set of observed variables using a small number of factors. Hence, EFA and CFA should not be viewed as disparate methods, despite that their implementation with conventional software might seem quite different. Instead, they are two approaches to

[1]For traditional factor analysis models, the means of the observed variables are arbitrary and unstructured by the model, which allows omission of an intercept term in Eq. 1 by assuming the observed variables are mean-deviated, or centered (MacCallum, 2009).

[2]In EFA, the model is typically estimated by first setting $\mathbf{\Psi}$ to be an identity matrix, which implies that the factors are uncorrelated, or orthogonal, leading to the initial unrotated factor loadings in $\mathbf{\Lambda}$. Applying an oblique factor rotation obtains a new set of factor loadings along with non-zero interfactor correlations. Although rotation is not a focus of the current paper, we recommend that researchers always use an oblique rotation.

investigating variants of the same general model which differ according to the number of constraints placed on the model, where the constraints are determined by the strength of theoretical expectations (see MacCallum, 2009). Indeed, it is possible to conduct EFA in a confirmatory fashion (e.g., by using "target rotation") or to conduct CFA in an exploratory fashion (e.g., by comparing a series of models that differ in the number or nature of the factors)[3]. Given that EFA and CFA are based on the same common factor model, the principles and methods we discuss in this paper largely generalize to both procedures. The example analyses we present below follow traditional EFA procedures; where necessary, we comment on how certain issues may be different for CFA.

In practice, a sample correlation matrix **R** (or sample covariance matrix **S**) is analyzed to obtain estimates of the unconstrained parameters given some specified value *m* of the number of common factors[4]. The parameter estimates can be plugged into Eq. 3 to derive a model-implied population correlation matrix, $\hat{\mathbf{P}}$. The goal of model estimation is thus to find the parameter estimates that optimize the match of the model-implied correlation matrix to the observed sample correlation matrix. An historically popular method for parameter estimation in EFA is the "principal factors method with prior communality estimates," or "principal axis factor analysis," which obtains factor loading estimates from the eigenstructure of a matrix formed by $\mathbf{R} - \hat{\boldsymbol{\Theta}}$ (see MacCallum, 2009). But given modern computing capabilities, we agree with MacCallum (2009) that this method should be considered obsolete, and instead factor analysis models should be estimated using an iterative algorithm to minimize a model fitting function, such as the unweighted least-squares (ULS) or maximum likelihood (ML) functions. Although principal axis continues to be used in modern applications of EFA, iterative estimation, usually ML, is almost always used with CFA. In short, ULS is preferable to principal axis because it will better account for the observed correlations in **R** (specifically, it will produce a smaller squared difference between a given observed correlation and the corresponding model-implied correlation), whereas ML obtains parameter estimates that give the most likely account of the observed data (MacCallum, 2009; see also Briggs and MacCallum, 2003). Below, we discuss how the assumption of normally distributed observed variables comes into play for ULS and ML.

As shown above, given that the parameters (namely, the factor loadings in $\boldsymbol{\Lambda}$ and interfactor correlations in $\boldsymbol{\Psi}$) are estimated directly from the observed correlations, **R**, using an estimator such as ML or ULS, factor analysis as traditionally implemented is essentially an analysis of the correlations among a set of observed variables. In this sense, the correlations *are* the data. Indeed, any factor analysis software program can proceed if a sample correlation matrix is the only data given; it does not need the complete raw, person by variable ($N \times p$) data set from which the correlations were calculated[5]. Conversely, when the software is given the complete ($N \times p$) data set, it will first calculate the correlations among the p variables to be analyzed and then fit the model to those correlations using the specified estimation method. Thus, it is imperative that the "data" for factor analysis, the correlations, be appropriate and adequate summaries of the relationships among the observed variables, despite that the common factor model makes no explicit assumptions about the correlations themselves. Thus, if the sample correlations are misrepresentative of the complete raw data, then the parameter estimates (factor loadings and interfactor correlations) will be inaccurate, as will model fit statistics and estimated SEs for the parameter estimates. Of course, more explicit assumption violation also can cause these problems. In these situations, information from the complete data set beyond just the correlations becomes necessary to obtain "robust" results.

Before we discuss these issues further, we first present a data example to illustrate the common factor model and to provide a context for demonstrating the main concepts of this article involving data screening and assumption testing for factor analysis. The purpose of this data example is not to provide a comprehensive simulation study for evaluating the effectiveness of different factor analytic methods; such studies have already been conducted in the literature cited throughout this paper. Rather, we use analyses of this data set to illustrate the statistical concepts discussed below so that they may be more concrete for the applied researcher.

DATA EXAMPLE

Our example is based on unpublished data reported in Harman (1960); these data and an accompanying factor analysis are described in the user's guide for the free software package CEFA (Browne et al., 2010). The variables are scores from nine cognitive ability tests. Although the data come from a sample of $N = 696$ individuals, for our purposes we consider the correlation matrix among the nine variables to be a population correlation matrix (see **Table 1**) and thus an example of **P** in Eq. 3. An obliquely rotated (quartimin rotation) factor pattern for a three-factor model is considered the population factor loading matrix (see **Table 2**) and thus an example of $\boldsymbol{\Lambda}$ in Eq. 3. Interpretation of the factor loadings indicates that the first factor (η_1) has relatively strong effects on the variables Word Meaning, Sentence Completion, and Odd words; thus, η_1 is considered a "verbal ability" factor. The second factor (η_2) has relatively strong effects on Mixed Arithmetic, Remainders, and Missing numbers; thus, η_2 is "math ability." Finally, the third factor (η_3) is "spatial ability" given its strong influences on Gloves, Boots, and Hatchets. The interfactor correlation between η_1 and η_2 is $\psi_{12} = 0.59$, the correlation between η_1 and η_3 is $\psi_{13} = 0.43$, and η_2 and η_3 are also moderately correlated with $\psi_{23} = 0.48$.

These population parameters give a standard against which to judge sample-based results for the same factor model that we present throughout this paper. To begin, we created a random sample of $N = 100$ with nine variables from a multivariate standard normal population distribution with a correlation matrix matching that in **Table 1**. We then estimated a three-factor EFA

[3]Asparouhov and Muthén (2009) further show how an unrestricted EFA solution can be used in place of a restricted CFA-type measurement model within a larger SEM.

[4]We will not extensively discuss EFA methods for determining the number of factors except to note where assumption violation or inaccurate correlations may affect decisions about the optimal number of common factors. See MacCallum (2009) for a brief review of methods for determining the number of common factors in EFA.

[5]Certain models, such as multiple-group measurement models, may also include a mean structure, in which case the means of the observed variables also serve as data for the complete model estimation.

Table 1 | Population correlation matrix.

Variable	1	2	3	4	5	6	7	8	9
WrdMean	1								
SntComp	0.75	1							
OddWrds	0.78	0.72	1						
MxdArit	0.44	0.52	0.47	1					
Remndrs	0.45	0.53	0.48	0.82	1				
MissNum	0.51	0.58	0.54	0.82	0.74	1			
Gloves	0.21	0.23	0.28	0.33	0.37	0.35	1		
Boots	0.30	0.32	0.37	0.33	0.36	0.38	0.45	1	
Hatchts	0.31	0.30	0.37	0.31	0.36	0.38	0.52	0.67	1

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers, Hatchts, hatchets.

Table 2 | Population factor loading matrix.

Variable	Factor		
	η_1	η_2	η_3
WrdMean	**0.94**	−0.05	−0.03
SntComp	**0.77**	0.14	−0.03
OddWrds	**0.83**	0.00	0.08
MxdArit	−0.05	**1.01**	−0.05
Remndrs	0.04	**0.80**	0.06
MissNum	0.14	**0.75**	0.06
Gloves	−0.06	0.13	**0.56**
Boots	0.05	−0.01	**0.74**
Hatchts	0.03	−0.09	**0.91**

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

model using ULS and applied a quartimin rotation. The estimated rotated factor loading matrix, $\hat{\mathbf{\Lambda}}$, is in **Table 3**. Not surprisingly, these factor loading estimates are similar, but not identical, to the population factor loadings.

REGRESSION DIAGNOSTICS FOR FACTOR ANALYSIS OF CONTINUOUS VARIABLES

Because the common factor model is just a linear regression model (Eq. 1), many of the well-known concepts about regression diagnostics generalize to factor analysis. Regression diagnostics are a set of methods that can be used to reveal aspects of the data that are problematic for a model that has been fitted to that data (see Belsley et al., 1980; Fox, 1991, 2008, for thorough treatments). Many data characteristics that are problematic for ordinary multiple regression are also problematic for factor analysis; the trick is that in the common factor model, the explanatory variables (the factors) are unobserved variables whose values cannot be determined precisely. In this section, we illustrate how regression diagnostic principles can be applied to factor analysis using the example data presented above.

Table 3 | Sample factor loading matrix (multivariate normal data, no outlying cases).

Variable	Factor		
	η_1	η_2	η_3
WrdMean	**0.96**	−0.08	0.04
SntComp	**0.81**	0.15	−0.13
OddWrds	**0.75**	0.05	0.12
MxdArit	−0.06	**1.01**	−0.03
Remndrs	0.09	**0.75**	0.09
MissNum	0.09	**0.80**	0.06
Gloves	−0.03	0.20	**0.56**
Boots	0.07	0.00	**0.68**
Hatchts	−0.01	−0.01	**0.85**

N = 100. WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

But taking a step back, given that factor analysis is basically an analysis of correlations, all of the principles about correlation that are typically covered in introductory statistics courses are also relevant to factor analysis. Foremost is that a product-moment correlation measures the *linear* relationship between two variables. Two variables may be strongly related to each other, but the actual correlation between them might be close to zero if their relationship is poorly approximated by a straight line (for example, a U-shaped relationship). In other situations, there may be a clear curvilinear relationship between two variables, but a researcher decides that a straight line is still a reasonable model for that relationship. Thus, some amount of subjective judgment may be necessary to decide whether a product-moment correlation is an adequate summary of a given bivariate relationship. If not, variables involved in non-linear relationships may be transformed prior to fitting the factor model or alternative correlations, such as Spearman rank-order correlations, may be factor analyzed (see Gorsuch, 1983, pp. 297–309 for a discussion of these strategies, but see below for methods for item-level categorical variables).

Visual inspection of simple scatterplots (or a scatterplot matrix containing many bivariate scatterplots) is an effective method for assessing linearity, although formal tests of linearity are possible (e.g., the RESET test of Ramsey, 1969). If there is a large number of variables, then it may be overly tedious to inspect every bivariate relationship. In this situation, one might focus on scatterplots involving variables with odd univariate distributions (e.g., strongly skewed or bimodal) or randomly select several scatterplots to scrutinize closely. Note that it is entirely possible for two variables to be linearly related even when one or both of them is non-normal; conversely, if two variables are normally distributed, their bivariate relationship is not necessarily linear. Returning to our data example, the scatterplot matrix in **Figure 1** shows that none of the bivariate relationships among these nine variables has any clear departure from linearity.

Scatterplots can also be effective for identifying unusual cases, although relying on scatterplots alone for this purpose is not

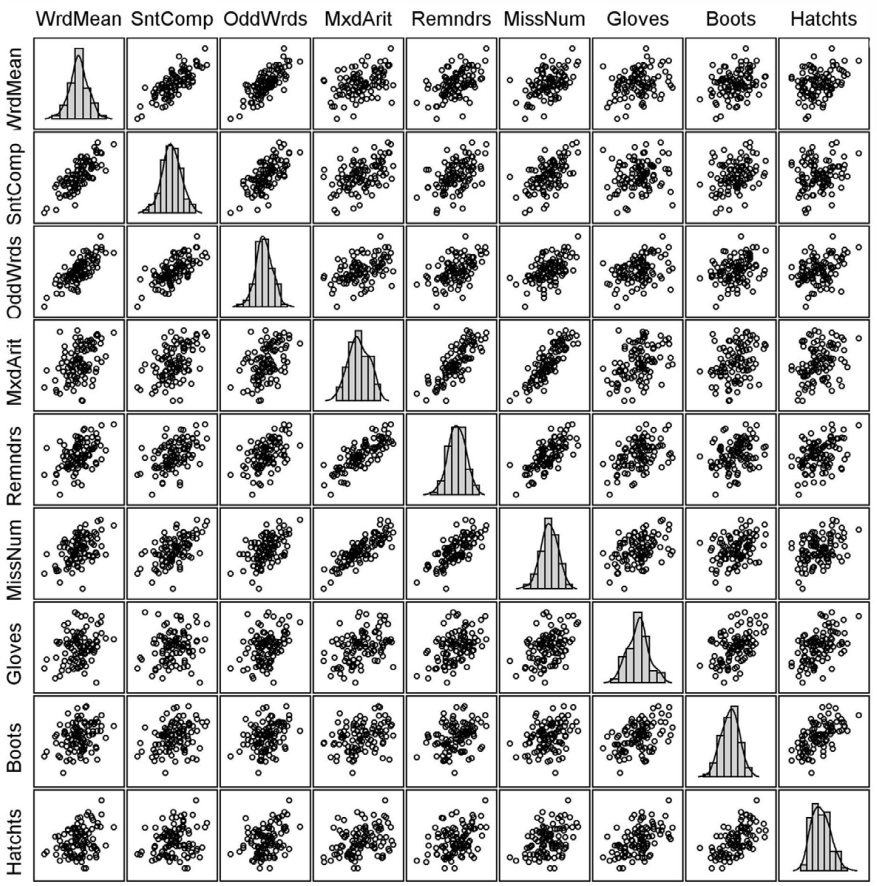


FIGURE 1 | Scatterplot matrix for multivariate normal random sample consistent with Holzinger data ($N = 100$; no unusual cases).

foolproof. Cases that appear to be outliers in a scatterplot might not actually be *influential* in that they produce distorted or otherwise misleading factor analysis results; conversely, certain influential cases might not easily reveal themselves in a basic bivariate scatterplot. Here, concepts from regression diagnostics come into play. Given that regression (and hence factor analysis) is a procedure for modeling a dependent variable conditional on one or more explanatory variables, a *regression outlier* is a case whose dependent variable value is unusual *relative to* its predicted, or modeled, value given its scores on the explanatory variables (Fox, 2008). In other words, regression outliers are cases with large residuals.

A regression outlier will have an impact on the estimated regression line (i.e., its slope) to the extent that it has high *leverage*. Leverage refers to the extent that a case has an unusual combination of values on the set of explanatory variables. Thus, if a regression outlier has low leverage (i.e., it is near the center of the multivariate distribution of explanatory variables), it should have relatively little *influence* on the estimated regression slope; that is, the estimated value of the regression slope should not change substantially if such a case is deleted. Conversely, a case with high leverage but a small residual also has little influence on the estimated slope value. Such a high leverage, small residual case is often called a "good leverage" case because its inclusion in the analysis leads to a more precise estimate of the regression slope (i.e., the SE of the slope is smaller). Visual inspection of a bivariate scatterplot might reveal such a case, and a naïve researcher might be tempted to call it an "outlier" and delete it. But doing so would be unwise because of the loss of statistical precision. Hence, although visual inspection of raw, observed data with univariate plots and bivariate scatterplots is always good practice, more sophisticated procedures are needed to gain a full understanding of whether unusual cases are likely to have detrimental impact on modeling results. Next, we describe how these concepts from regression diagnostics extend to factor analysis.

FACTOR MODEL OUTLIERS

The factor analysis analog to a regression outlier is a case whose value for a particular observed variable is extremely different from its predicted value given its scores on the factors. In other words, cases with large (absolute) values for one or more unique factors, that is, scores on residual terms in $\boldsymbol{\varepsilon}$, are factor model outliers. Eq. 1 obviously defines the residuals $\boldsymbol{\varepsilon}$ as

$$\boldsymbol{\varepsilon} = \mathbf{y} - \boldsymbol{\Lambda}\boldsymbol{\eta}. \tag{4}$$

But because the factor scores (i.e., scores on latent variables in $\boldsymbol{\eta}$) are unobserved and cannot be calculated precisely, so too the residuals cannot be calculated precisely, even with known population factor loadings, $\boldsymbol{\Lambda}$. Thus, to obtain estimates of the residuals, $\hat{\boldsymbol{\varepsilon}}$, it is necessary first to estimate the factor scores, $\hat{\boldsymbol{\eta}}$, which themselves must be based on sample estimates of $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Theta}}$. Bollen and Arminger (1991) show how two well-known approaches to estimating factor scores, the least-squares regression method and Bartlett's method, can be applied to obtain $\hat{\boldsymbol{\varepsilon}}$. As implied by Eq. 4, the estimated residuals $\hat{\boldsymbol{\varepsilon}}$ are unstandardized in that they are in the metric of the observed variables, $\mathbf{y}$. Bollen and Arminger thus present formulas to convert unstandardized residuals into standardized residuals. Simulation results by Bollen and Arminger show good performance of their standardized residuals for revealing outliers, with little difference according to whether they are estimated from regression-based factor scores or Bartlett-based factor scores.

Returning to our example sample data, we estimated the standardized residuals from the three-factor EFA model for each of the $N = 100$ cases; **Figure 2** illustrates these residuals for all nine observed variables (recall that in this example, $\mathbf{y}$ consists of nine variables and thus there are nine unique factors in $\hat{\boldsymbol{\varepsilon}}$). Because the data were drawn from a standard normal distribution conforming to a known population model, these residuals themselves should be approximately normally distributed with no extreme outliers. Any deviation from normality in **Figure 2** is thus only due to sampling error and error due to the approximation of factor scores. Later, we introduce an extreme outlier in the data set to illustrate its effects.

LEVERAGE

As mentioned above, in regression it is important to consider leverage in addition to outlying residuals. The same concept applies in factor analysis (see Yuan and Zhong, 2008, for an extensive discussion). Leverage is most commonly quantified using "hat values" in multiple regression analysis, but a related statistic, Mahalonobis distance (MD), also can be used to measure leverage (e.g., Fox, 2008). MD helps measure the extent to which an observation is a *multivariate* outlier with respect to the set of explanatory variables[6]. Here, our use of MD draws from Pek and MacCallum (2011), who recommend its use for uncovering multivariate outliers in the context of SEM. In a factor analysis, the MD for a given

[6]Weisberg (1985) showed that $(n-1)h^*$ is the MD for a given case from the centroid of the explanatory variables, where h^* is the regression hat value calculated using mean-deviated independent variables.

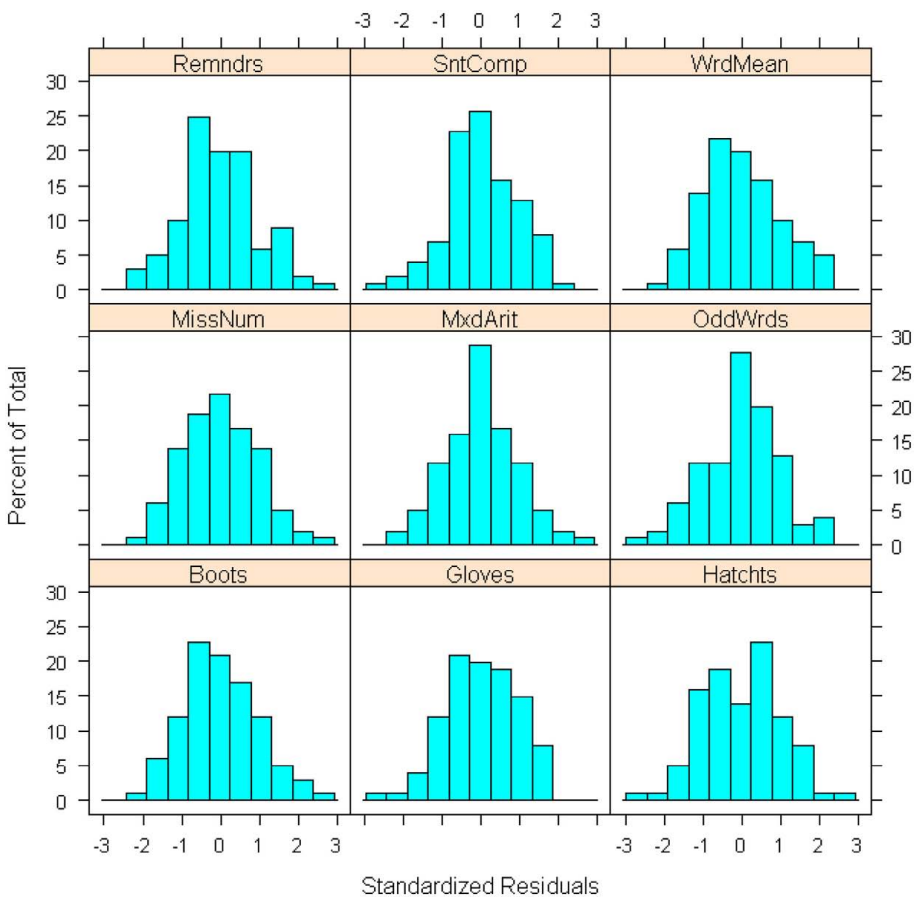


FIGURE 2 | Histograms of standardized residuals for each observed variable from three-factor model fitted to random sample data ($N = 100$; no unusual cases).

observation can be measured for the set of observed variables with

$$\mathrm{MD}_i = \left(\mathbf{y}_i - \bar{\mathbf{y}}\right)' \mathbf{S}^{-1} \left(\mathbf{y}_i - \bar{\mathbf{y}}\right)$$

where $\mathbf{y}_i$ is the vector of observed variable scores for case i, $\bar{\mathbf{y}}$ is the vector of means for the set of observed variables, and $\mathbf{S}$ is the sample covariance matrix. Conceptually, MD_i is the squared distance between the data for case i and the center of the observed multivariate "data cloud" (or data *centroid*), standardized with respect to the observed variables' variances and covariances. Although it is possible to determine critical cut-offs for extreme MDs under certain conditions, we instead advocate graphical methods (shown below) for inspecting MDs because distributional assumptions may be dubious and values just exceeding a cut-off may still be legitimate observations in the tail of the distribution.

A potential source of confusion is that in a regression analysis, MD is defined with respect to the explanatory variables, but for factor analysis MD is defined with respect to the observed variables, which are the dependent variables in the model. But for both multiple regression and factor analysis, MD is a model-free index in that its value is not a function of the estimated parameters (but see Yuan and Zhong, 2008, for model-based MD-type measures). Thus, analogous to multiple regression, we can apply MD to find cases that are far from the center of the observed data centroid to measure their potential impact on results. Additionally, an important property of both MD and model residuals is that they are based on the full multivariate distribution of observed variables, and as such can uncover outlying observations that may not easily appear as outliers in a univariate distribution or bivariate scatterplot.

The MD values for our simulated sample data are summarized in **Figure 3**. The histogram shows that the distribution has a minimum of zero and positive skewness, with no apparent outliers, but the boxplot does reveal potential outlying MDs. Here, these are not extreme outliers but instead represent legitimate values in the tail of the distribution. Because we generated the data from a multivariate normal distribution we do not expect any extreme MDs, but in practice the data generation process is unknown and subjective judgment is needed to determine whether a MD is extreme enough to warrant concern. Assessing *influence* (see below) can aid that judgment. This example shows that extreme MDs may occur even with simple random sampling from a well-behaved distribution, but such cases are perfectly legitimate and should not be removed from the data set used to estimate factor models.

INFLUENCE

Again, cases with large residuals are not necessarily influential and cases with high MD are not necessarily bad leverage points (Yuan and Zhong, 2008). A key heuristic in regression diagnostics is that case *influence* is a product of both leverage and the discrepancy of predicted values from observed values as measured by residuals. Influence statistics known as deletion statistics summarize the extent to which parameter estimates (e.g., regression slopes or factor loadings) change when an observation is deleted from a data set.

A common deletion statistic used in multiple regression is Cook's distance, which can be broadened to generalized Cook's distance (gCD) to measure the influence of a case on a set of parameter estimates from a factor analysis model (Pek and MacCallum, 2011) such that

$$\mathrm{gCD}_i = \left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}\right)' \left[\hat{\mathrm{VAR}}\left(\hat{\boldsymbol{\theta}}_{(i)}\right)\right]^{-1} \left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}\right),$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{(i)}$ are vectors of parameter estimates obtained from the original, full sample and from the sample with case i deleted and $\hat{\mathrm{VAR}}\left(\hat{\boldsymbol{\theta}}_{(i)}\right)$ consists of the estimated asymptotic variances (i.e., squared SEs) and covariances of the parameter estimates obtained with case i deleted. Like MD, gCD is in a squared metric with values close to zero indicating little case influence on parameter estimates and those far from zero indicating strong case influence on the estimates. **Figure 4** presents the distribution of gCD values calculated across the set of factor loading estimates from the three-factor EFA model fitted to the example data. Given the squared metric of gCD, the strong positive skewness is expected. The boxplot indicates some potential outlying gCDs, but again these are not extreme outliers and instead are legitimate observations in the long tail of the distribution.

Because gCD is calculated by deleting only a single case i from the complete data set, it is susceptible to *masking errors*, which

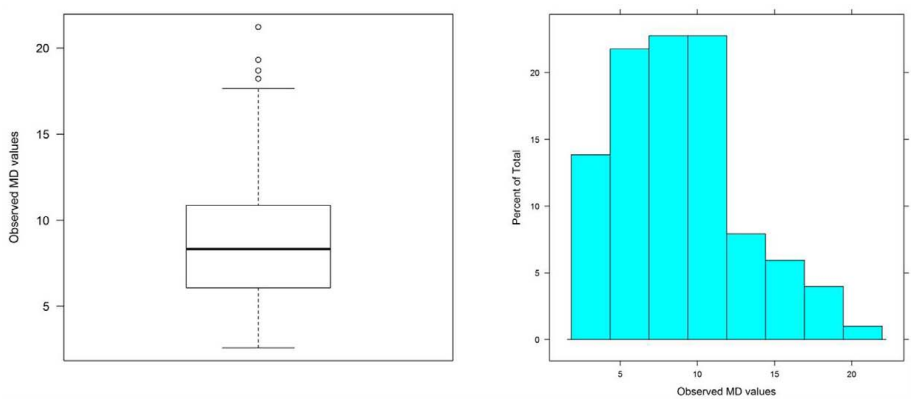


FIGURE 3 | Distribution of Mahalanobis Distance (MD) for multivariate normal random sample data ($N = 100$; no unusual cases).

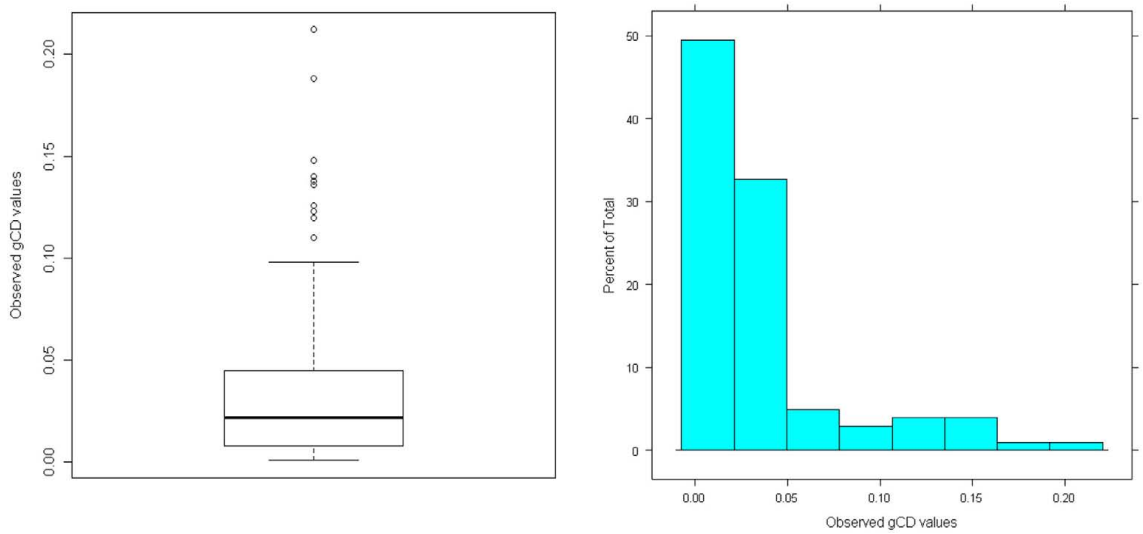


FIGURE 4 | Distribution of generalized Cook's distance (gCD) for multivariate normal random sample data (*N* = 100; no unusual cases).

occur when an influential case is not identified as such because it is located in the multivariate space close to one or more similarly influential cases. Instead, a local influence (Cadigan, 1995; Lee and Wang, 1996) or forward search (Poon and Wong, 2004; Mavridis and Moustaki, 2008) approach can be used to identify groups of influential cases. It is important to recognize that when a model is poorly specified (e.g., the wrong number of factors has been extracted), it is likely that many cases in a sample would be flagged as influential, but when there are only a few bad cases, the model may be consistent with the major regularities in the data except for these cases (Pek and MacCallum, 2011).

EXAMPLE DEMONSTRATION

To give one demonstration of the potential effects of an influential case, we replaced one of the randomly sampled cases from the example data with a non-random case that equaled the original case with $Z = 2$ added to its values for five of the observed variables (odd-numbered items) and $Z = 2$ subtracted from the other four observed variables (even-numbered items). **Figure 5** indicates the clear presence of this case in the scatterplot of Remainders by Mixed Arithmetic. When we conducted EFA with this perturbed data set, the scree plot was ambiguous as to the optimal number of factors, although other model fit information, such as the root mean square residual (RMSR) statistic (see MacCallum, 2009), more clearly suggested a three-factor solution. Importantly, the quartimin-rotated three-factor solution (see **Table 4**) has some strong differences from both the known population factor structure (**Table 2**) and the factor structure obtained with the original random sample (**Table 3**). In particular, while Remainders still has its strongest association with η_2, its loading on this factor has dropped to 0.49 from 0.75 with the original sample (0.80 in the population). Additionally, Remainders now has a noticeable cross-loading on η_3 equaling 0.32, whereas this loading had been 0.09 with the original sample (0.06 in the population). Finally, the loading for Boots on η_3 has dropped substantially to 0.44 from 0.68 with the original sample (0.74 in the population).

Having estimated a three-factor model with the perturbed data set, we then calculated the associated residuals $\hat{\boldsymbol{\varepsilon}}$, the sample MD values, and the gCD values. **Figure 6** gives histograms of the residuals, where it is clear that there is an outlying case for Mixed Arithmetic in particular. In **Figure 7**, the distribution of MD also indicates the presence of a case that falls particularly far from the centroid of the observed data; here the outlying observation has MD = 53.39, whereas the maximum MD value in the original data set was only 21.22. Given that the outlying case has both large residuals and large leverage, we expect it to have a strong influence on the set of model estimates. Hence, **Figure 8** reveals that in the perturbed data set, all observations have gCD values very close to 0, but the outlying case has a much larger gCD reflecting its strong influence on the parameter estimates.

The demonstration above shows how the presence of even one unusual case can have a drastic effect on a model's parameter estimates; one can imagine how such an effect can produce a radically different substantive interpretation for a given variable within a factor model or even for the entire set of observed variables, especially if the unusual case leads to a different conclusion regarding the number of common factors. Improper solutions (e.g., a model solution with at least one negative estimated residual variance term, or "Heywood case") are also likely to occur in the presence of one or more unusual cases (Bollen, 1987), which can lead to a researcher unwittingly revising a model or removing an observed variable from the analysis. Another potential effect of unusual cases is that they can make an otherwise approximately normal distribution appear non-normal by creating heavy tails in the distribution, that is, excess kurtosis (Yuan et al., 2002). As we discuss below, excess kurtosis can produce biased model fit statistics and SE estimates. Even if they do not introduce excess kurtosis, unusual cases can still impact overall model fit. The

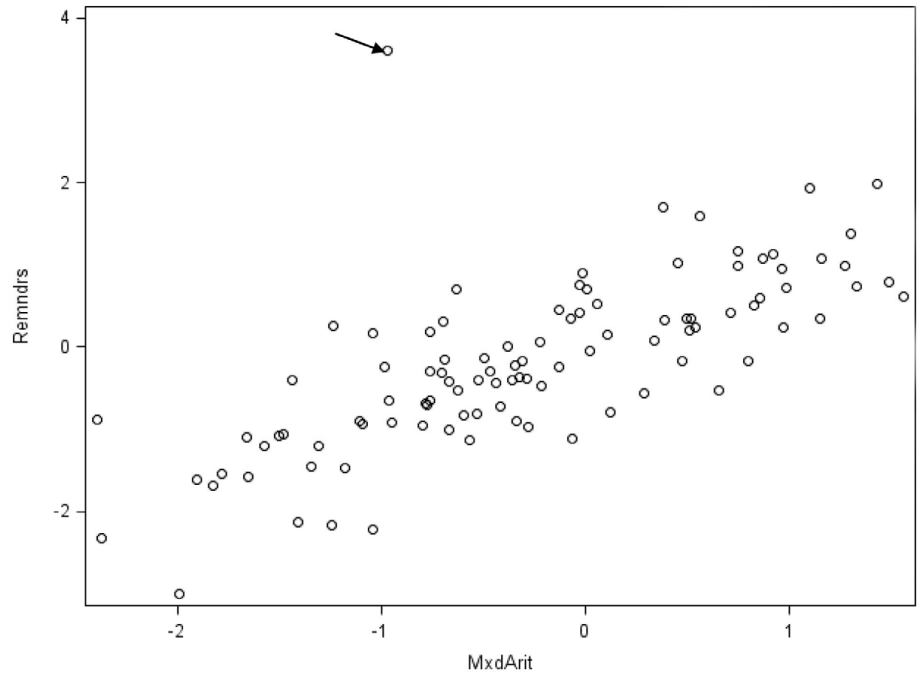


FIGURE 5 | Scatterplot of "Remainders" by "Mixed Arithmetic" for perturbed sample with influential case indicated.

Table 4 | Factor loading matrix obtained with perturbed sample data.

Variable	Factor		
	η_1	η_2	η_3
WrdMean	**0.97**	−0.13	0.09
SntComp	**0.70**	0.29	−0.23
OddWrds	**0.74**	−0.01	0.21
MxdArit	−0.07	**1.01**	0.03
Remndrs	0.17	**0.49**	0.32
MissNum	0.08	**0.81**	0.06
Gloves	0.01	0.09	**0.68**
Boots	0.05	0.12	**0.44**
Hatchts	0.02	−0.08	**0.89**

N = 100. WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

effect of an individual case on model fit with ML estimation can be formally measured with an influence statistic known as *likelihood distance*, which measures the difference in the likelihood of the model when a potentially influential case is deleted (Pek and MacCallum, 2011).

Upon discovering unusual cases, it is important to determine their likely source. Often, outliers and influential cases arise from either researcher error (e.g., data entry error or faulty administration of study procedures) or participant error (e.g., misunderstanding of study instructions or non-compliance with random responding) or they may be observations from a population other than the population of interest (e.g., a participant with no history of depression included in a study of depressed individuals). In these situations, it is best to remove such cases from the data set. Conversely, if unusual cases are simply extreme cases with otherwise legitimate values, most methodologists recommend that they *not* be deleted from the data set prior to model fitting (e.g., Bollen and Arminger, 1991; Yuan and Zhong, 2008; Pek and MacCallum, 2011). Instead, robust procedures that minimize the excessive influence of extreme cases are recommended; in particular, case-robust methods developed by Yuan and Bentler (1998) are implemented in the EQS software package (Bentler, 2004) or one can factor analyze a minimum covariance determinant (MCD) estimated covariance matrix (Pison et al., 2003), which can be calculated with SAS or the R package "MASS."

COLLINEARITY

Another potential concern for both multiple regression analysis and factor analysis is *collinearity*, which refers to perfect or near-perfect linear relationships among observed variables. With multiple regression, the focus is on collinearity among explanatory variables, but with factor analysis, the concern is collinearity among dependent variables, that is, the set of variables being factor analyzed. When collinear variables are included, the product-moment correlation matrix **R** will be *singular*, or *non-positive definite*. ML estimation cannot be used with a singular **R**, and although ULS is possible, collinearity is still indicative of conceptual issues with variable selection. Collinearity in factor analysis is relatively simple to diagnose: if any eigenvalues of a product-moment **R** equal zero or are negative, then **R** is non-positive definite and collinearity is present (and software will likely produce

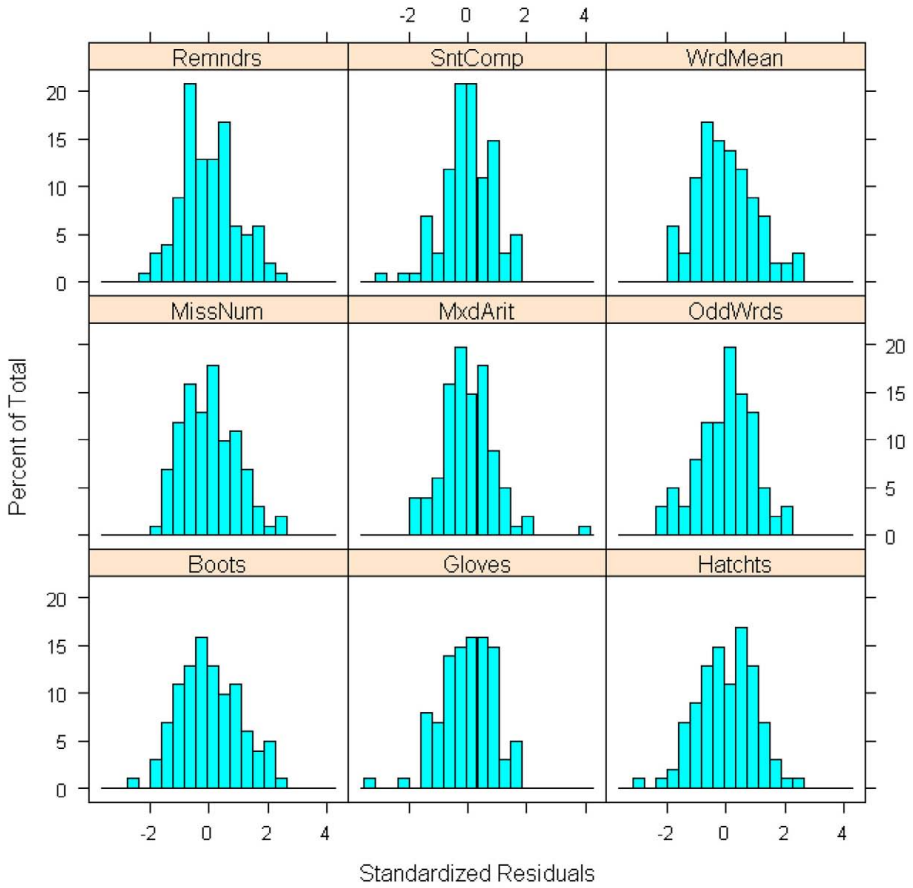


FIGURE 6 | Histograms of standardized residuals for each observed variable from three-factor model fitted to perturbed sample data ($N = 100$).

a warning message)[7]. Eigenvalues extremely close to zero may also be suggestive of near-collinearity[8]. A commonly used statistic for evaluating near-collinearity in ordinary regression is the *condition index*, which equals the square root of the ratio of the largest to smallest eigenvalue, with larger values more strongly indicative of near-collinearity. For the current example, the condition index equals 6.34, which is well below the value of 30 suggested by Belsley et al. (1980) as indicative of problematic near-collinearity.

In the context of factor analysis, collinearity often implies an ill-conceived selection of variables for the analysis. For example, if both total scores and sub-scale scores (or just one sub-scale) from the same instrument are included in a factor analysis, the total score is linearly dependent on, or collinear with, the sub-scale score. A more subtle example may be the inclusion of both a "positive mood" scale and a "negative mood" scale; although scores from these two scales may not be perfectly related, they will likely have a very strong (negative) correlation, which can be problematic for factor analysis. In these situations, researchers should carefully reconsider their choice of variables for the analysis and remove any that is collinear with one or more of the other observed variables, which in turn redefines the overall research question addressed by the analysis (see Fox, 2008, p. 342).

NON-NORMALITY

In terms of obtaining accurate parameter estimates, the ULS estimation method mentioned above makes no assumption regarding observed variable distributions whereas ML estimation is based on the multivariate normal distribution (MacCallum, 2009). Specifically, for both estimators, parameter estimates are trustworthy as long as the sample size is large and the model is properly specified (i.e., the estimator is *consistent*), even when the normality assumption for ML is violated (Bollen, 1989). However, SE estimates and certain model fit statistics (i.e., the χ^2 fit statistic and statistics based on χ^2 such as CFI and RMSEA) are adversely affected by non-normality, particularly excess multivariate kurtosis. Although they are less commonly used with EFA, SEs and model fit statistics are equally applicable with EFA as with CFA and are easily obtained with modern software (but see Cudeck and O'Dell, 1994, for cautions and recommendations regarding SEs with EFA). With both approaches, SEs convey information about the sampling variability

[7]Polychoric correlation matrices (defined below) are often non-positive definite. Unlike a product-moment **R**, a non-positive definite polychoric correlation matrix is not necessarily problematic.

[8]Most EFA software includes eigenvalues of **R** as default output. However, it is important not to confuse the eigenvalues of **R** with the eigenvalues of the "reduced" correlation matrix, $\mathbf{R} - \hat{\mathbf{\Theta}}$. The reduced correlation matrix often has negative eigenvalues when **R** does not.

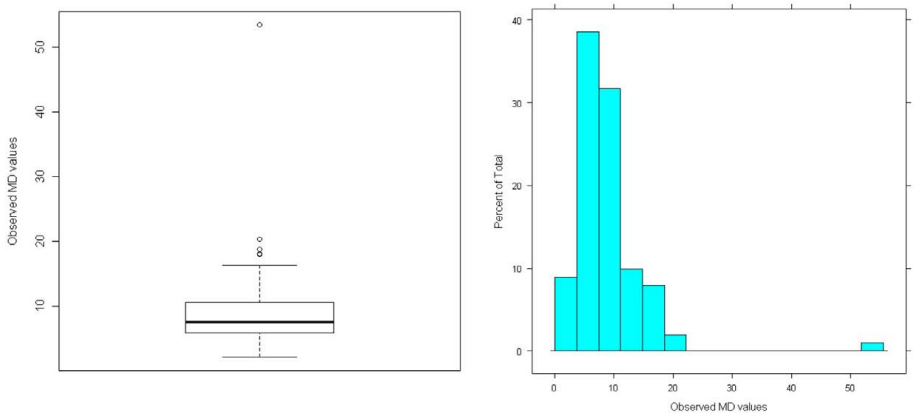


FIGURE 7 | Distribution of Mahalanobis distance (MD) for perturbed sample data (*N* = 100).

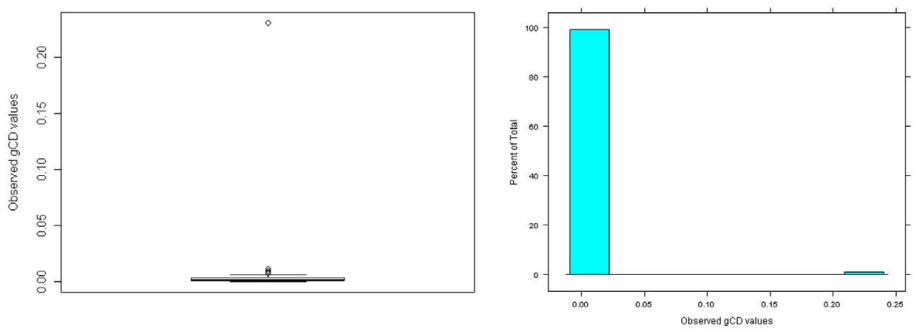


FIGURE 8 | Distribution of generalized Cook's distance (gCD) for perturbed sample data (*N* = 100).

of the parameter estimates (i.e., they are needed for significance tests and forming confidence intervals) while model fit statistics can aid decisions about the number of common factors in EFA or more subtle model misspecification in CFA.

There is a large literature on ramifications of non-normality for SEM (in which the common factor model is imbedded), and procedures for handling non-normal data (for reviews see Bollen, 1989; West et al., 1995; Finney and DiStefano, 2006). In particular, for CFA we recommend using the Satorra–Bentler scaled χ^2 and robust SEs with non-normal continuous variables (Satorra and Bentler, 1994), which is available in most SEM software. Although these Satorra–Bentler procedures for non-normal data have seen little application in EFA, they can be obtained for EFA models using Mplus software (Muthén and Muthén, 2010; i.e., using the SE estimation procedure outlined in Asparouhov and Muthén, 2009). Alternatively, one might factor analyze transformed observed variables that more closely approximate normal distributions (e.g., Gorsuch, 1983, pp. 297–309).

FACTOR ANALYSIS OF ITEM-LEVEL OBSERVED VARIABLES

Through its history in psychometrics, factor analysis developed primarily in the sub-field of cognitive ability testing, where researchers sought to refine theories of intelligence using factor analysis to understand patterns of covariation among different ability tests. Scores from these tests typically elicited continuously distributed observed variables, and thus it was natural for factor analysis to develop as a method for analyzing Pearson product-moment correlations and eventually to be recognized as a linear model for continuous observed variables (Bartholomew, 2007). However, modern applications of factor analysis usually use individual test items rather than sets of total test scores as observed variables. Yet, because the most common kinds of test items, such as Likert-type items, produce categorical (dichotomous or ordinal) rather than continuous distributions, a linear factor analysis model using product-moment **R** is suboptimal, as we illustrate below.

As early as Ferguson (1941), methodologists have shown that factor analysis of product-moment **R** among dichotomous variables can produce misleading results. Subsequent research has further established that treating categorical items as continuous variables by factor analyzing product-moment **R** can lead to incorrect decisions about the number of common factors or overall model fit, biased parameter estimates, and biased SE estimates (Muthén and Kaplan, 1985, 1992; Babakus et al., 1987; Bernstein and Teng, 1989; Dolan, 1994; Green et al., 1997). Despite these issues, item-level factor analysis using product-moment **R** persists in the substantive literature likely because of either naiveté about the categorical nature of items or misinformed belief that

linear factor analysis is "robust" to the analysis of categorical items.

EXAMPLE DEMONSTRATION

To illustrate these potential problems using our running example, we categorized random samples of continuous variables with $N = 100$ and $N = 200$ that conform to the three-factor population model presented in **Table 2** according to four separate cases of categorization. In Case 1, all observed variables were dichotomized so that each univariate population distribution had a proportion = 0.5 in both categories. In Case 2, five-category items were created so that each univariate population distribution was symmetric with proportions of 0.10, 0.20, 0.40, 0.20, and 0.10 for the lowest to highest categorical levels, or item-response categories. Next, Case 3 items were dichotomous with five variables (odd-numbered items) having univariate response proportions of 0.80 and 0.20 for the first and second categories and the other four variables (even-numbered items) having proportions of 0.20 and 0.80 for the first and second categories. Finally, five-category items were created for Case 4 with odd-numbered items having positive skewness (response proportions of 0.51, 0.30, 0.11, 0.05, and 0.03 for the lowest to highest categories) and even-numbered items having negative skewness (response proportions of 0.03, 0.05, 0.11, 0.30, and 0.51). For each of the four cases at both sample sizes, we conducted EFA using the product-moment **R** among the categorical variables and applied quartimin rotation after determining the optimal number of common factors suggested by the sample data. If product-moment correlations are adequate representations of the true relationships among these items, then three-factor models should be supported and the rotated factor pattern should approximate the population factor loadings in **Table 2**. We describe analyses for Cases 1 and 2 first, as both of these cases consisted of items with approximately symmetric univariate distributions.

First, **Figure 9** shows the simple bivariate scatterplot for the dichotomized versions of the Word Meaning and Sentence Completion items from Case 1 ($N = 100$). We readily admit that this figure is not a good display for these data; instead, its crudeness is intended to help illustrate that it is often not appropriate to pretend that categorical variables are continuous. When variables are continuous, bivariate scatterplots (such as those in **Figure 1**) are very useful, but **Figure 9** shows that they are not particularly useful for dichotomous variables, which in turn should cast doubt on the usefulness of a product-moment correlation for such data. More specifically, because these two items are dichotomized (i.e., 0, 1), there are only four possible observed data patterns, or response patterns, for their bivariate distribution (i.e., 0, 0; 0, 1; 1, 0; and 1, 1). These response patterns represent the only possible points in the scatterplot. Yet, depending on the strength of relationship between the two variables, there is some frequency of observations associated with each point, as each represents potentially

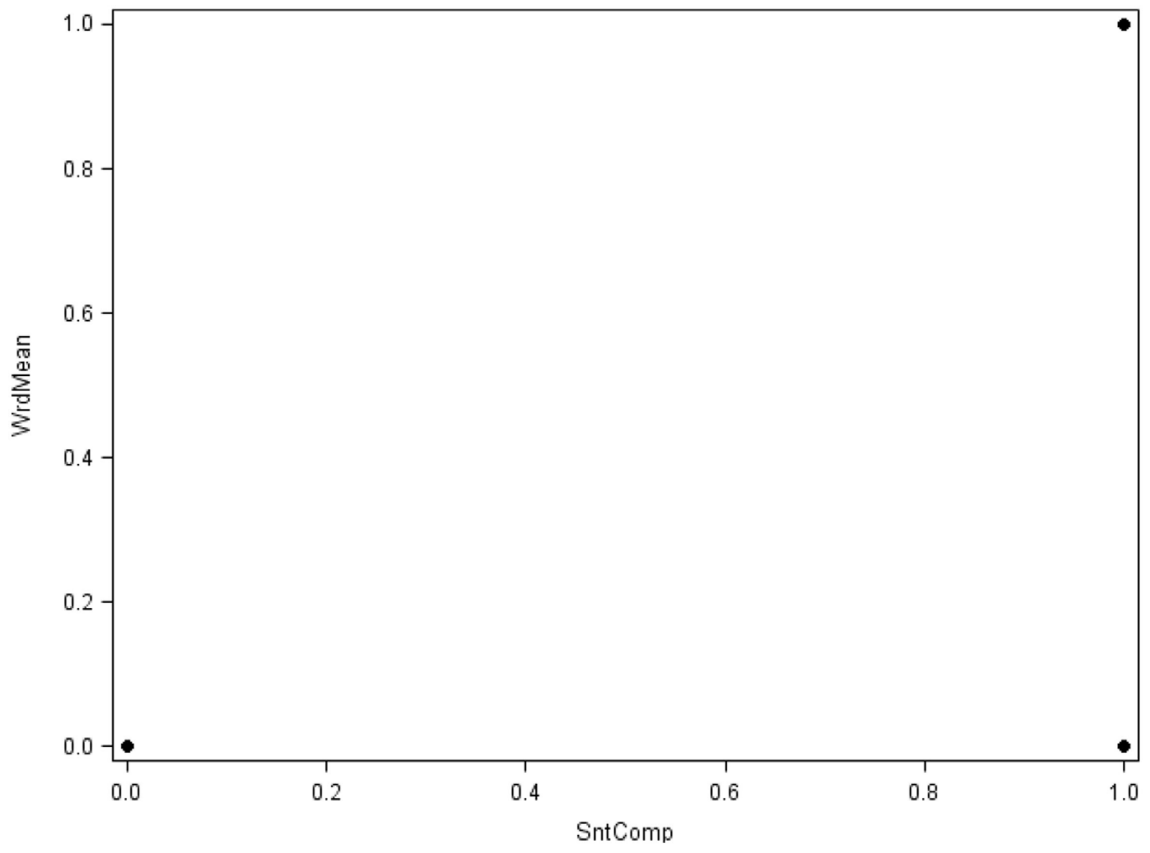


FIGURE 9 | Scatterplot of Case 1 items Word Meaning (WrdMean) by Sentence Completion (SntComp; *N* = 100).

many observations with the same response pattern. Conversely, the response pattern (0, 1) does not appear as a point in **Figure 9** because there were zero observations with a value of 0 on the Sentence Completion item and a value of 1 on Word Meaning. As emphasized above, a product-moment correlation measures the strength of *linear* association between two variables; given the appearance of the scatterplot, use and interpretation of such a correlation here is clearly dubious, which then has ramifications for a factor analysis of these variables based on product-moment **R** (or covariances).

Next, **Figure 10** shows the bivariate scatterplot for the five-category versions of the Word Meaning and Sentence Completion variables from Case 2 ($N = 100$). Now there are $5 \times 5 = 25$ potential response patterns, but again, not all appear as points in the plot because not all had a non-zero sample frequency. With more item-response categories, a linear model for the bivariate association between these variables may seem more reasonable but is still less than ideal, which again has implications for factor analysis of product-moment **R** among these items[9].

One consequence of categorization is that product-moment correlations are attenuated (see **Table 5** for correlations among Case 1 items). For example, the population correlation between Word Meaning and Sentence Completion is 0.75, but the sample product-moment correlation between these two Case 1 items is

[9]It is possible (and advisable) to make enhanced scatterplots in which each point has a different size (or color or symbol) according to the frequency or relative frequency of observations at each point.

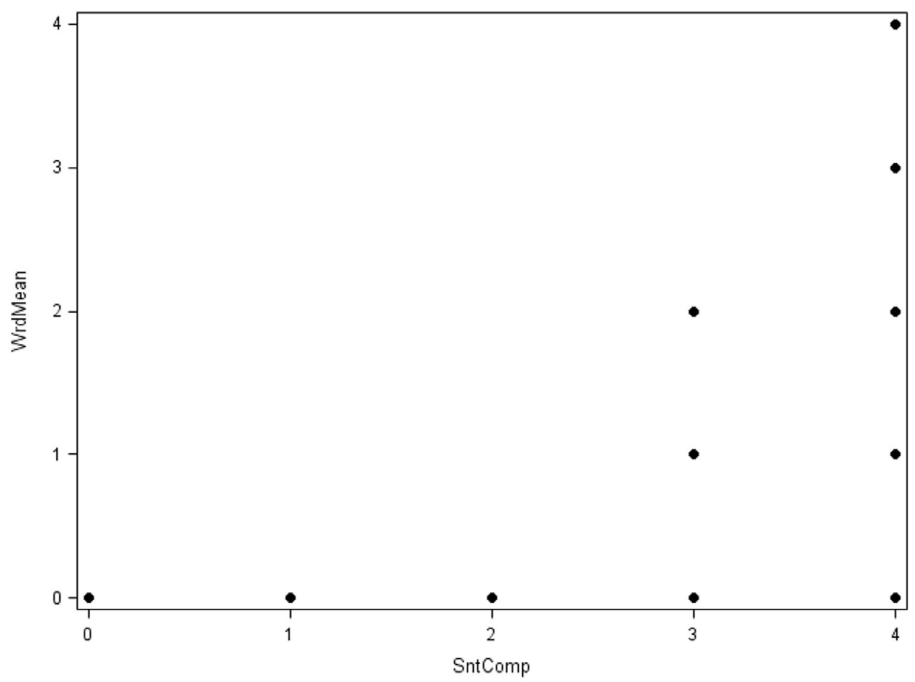


FIGURE 10 | Scatterplot of Case 2 items Word Meaning (WrdMean) by Sentence Completion (SntComp; *N* = 100).

Table 5 | Product-moment and polychoric correlations among Case 1 item-level variables.

Variable	1	2	3	4	5	6	7	8	9
WrdMean	1	0.81	0.78	0.42	0.35	0.35	0.12	0.17	0.11
SntComp	0.60	1	0.73	0.40	0.46	0.62	0.19	0.29	0.04
OddWrds	0.56	0.52	1	0.45	0.45	0.45	0.33	0.41	0.34
MxdArit	0.27	0.26	0.29	1	0.85	0.80	0.28	0.33	0.30
Remndrs	0.23	0.30	0.30	0.64	1	0.73	0.48	0.38	0.19
MissNum	0.23	0.42	0.29	0.58	0.51	1	0.20	0.27	0.12
Gloves	0.07	0.12	0.21	0.17	0.31	0.12	1	0.45	0.52
Boots	0.11	0.19	0.27	0.21	0.25	0.17	0.29	1	0.69
Hatchts	0.07	0.02	0.22	0.19	0.12	0.07	0.35	0.49	1

N = 100. Product-moment correlations are below the diagonal; polychoric correlations are above the diagonal. WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets.

0.60 (with $N = 100$). But if all correlations among items are attenuated to a similar degree, then the overall pattern of correlations in **R** should be very similar under categorization, and thus factor analysis results may not be strongly affected (although certain model fit indices may be affected, adversely impacting decisions about the number of factors). Indeed, in the EFA of the Case 1 items it was evident that a three-factor model was ideal (based on numerous criteria) and the set of rotated factor loadings (see **Table 6**) led to essentially the same interpretations of the three factors as for the population model. Nonetheless, the magnitude of each primary factor loading was considerably smaller compared to the population model, reflecting that attenuation of correlations due to categorization leads to biased parameter estimates. Note also that factor loading bias was no better with $N = 200$ compared to $N = 100$, because having a larger sample size does not make the observed bivariate relationships stronger or "more linear." However, attenuation of correlation is less severe with a larger number of response categories (see **Table 7**). In Case 2, the sample product-moment correlation between the Word Meaning and Sentence Completion items is 0.68, which is still attenuated relative to the population value 0.75, but less so than in Case 1. And when the EFA was conducted with the Case 2 variables, a similar three-factor model was obtained, but with factor loading estimates (see **Table 8**) that were less biased than those from Case 1.

Now consider Cases 3 and 4, which had a mix of positively and negatively skewed univariate item distributions. Any simple scatterplot of two items from Case 3 will look very similar to **Figure 9** for the dichotomized versions of Word Meaning and Sentence Completion from Case 1 because the Case 3 items are also dichotomized, again leading to only four possible bivariate response patterns[10]. Likewise, any simple scatterplot of two items from Case 4 will look similar to **Figure 10** for the five-category items from Case 2 because any bivariate distribution from Case 4 also has 25 possible response patterns. Yet, it is well-known that the product-moment correlation between two dichotomous or ordered categorical variables is strongly determined by the shape of their univariate distributions (e.g., Nunnally and Bernstein, 1994). For example, in Case 3, because the dichotomized Word Meaning and Sentence Completion items had opposite skewness, the sample correlation between them is only 0.24 (with $N = 100$) compared to the population correlation = 0.75 (see **Table 9** for correlations among Case 3 items). In Case 4, these two items also had opposite skewness, but the sample correlation = 0.52 (with $N = 100$) is less severely biased because Case 4 items have five categories rather than two (see **Table 10** for correlations among Case 4 items). Conversely, the Word Meaning and Hatchets items were both positively skewed; in Case 3, the correlation between these two items is 0.37 (with $N = 100$), which is greater than the population correlation = 0.31, whereas this correlation is 0.29 (with $N = 100$) for Case 4 items.

When we conducted EFA with the Case 3 and Case 4 items, scree plots suggested the estimation of two-, rather than three-factor models, and RMSR was sufficiently low to support the adequacy

Table 6 | Factor loading matrix obtained with EFA of product-moment R among Case 1 item-level variables.

Variable	Factor					
	$N = 100$			$N = 200$		
	η_1	η_2	η_3	η_1	η_2	η_3
WrdMean	**0.82**	−0.06	−0.05	**0.79**	−0.09	−0.02
SntComp	**0.74**	0.10	−0.07	**0.72**	0.14	−0.05
OddWrds	**0.66**	0.00	0.20	**0.69**	0.02	0.11
MxdArit	−0.04	**0.81**	0.07	−0.03	**0.82**	0.11
Remndrs	−0.03	**0.77**	−0.08	−0.02	**0.70**	−0.08
MissNum	0.10	**0.68**	−0.08	0.09	**0.76**	−0.08
Gloves	0.00	0.12	**0.43**	0.03	0.01	**0.47**
Boots	0.07	0.03	**0.60**	0.00	0.01	**0.61**
Hatchts	−0.03	−0.10	**0.81**	−0.02	−0.02	**0.80**

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

[10] An enhanced scatterplot that gives information about the frequency of each cell for Case 3 would look quite different than that for Case 1 because the categorizations we applied lead to different frequency tables for each data set.

Table 7 | Product-moment and polychoric correlations among Case 2 item-level variables.

Variable	1	2	3	4	5	6	7	8	9
WrdMean	1	0.76	0.80	0.47	0.54	0.53	0.38	0.33	0.34
SntComp	0.68	1	0.66	0.51	0.52	0.60	0.21	0.33	0.16
OddWrds	0.72	0.60	1	0.49	0.59	0.57	0.38	0.34	0.34
MxdArit	0.43	0.46	0.45	1	0.81	0.82	0.39	0.33	0.30
Remndrs	0.48	0.47	0.53	0.75	1	0.73	0.44	0.34	0.40
MissNum	0.48	0.54	0.52	0.76	0.67	1	0.50	0.35	0.35
Gloves	0.34	0.19	0.35	0.36	0.40	0.45	1	0.58	0.51
Boots	0.30	0.30	0.31	0.31	0.31	0.32	0.53	1	0.66
Hatchts	0.32	0.16	0.33	0.27	0.36	0.32	0.45	0.60	1

$N = 100$. Product-moment correlations are below the diagonal; polychoric correlations are above the diagonal. WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets.

of a two-factor solution (RMSR = 0.056 for Case 3 and = 0.077 for Case 4). In practice, it is wise to compare results for models with varying numbers of factors, and here we know that the population model has three factors. Thus, we estimated both two- and three-factor models for the Case 3 and Case 4 items. In Case 3, the estimated three-factor models obtained with both $N = 100$ and $N = 200$ were improper in that there were variables with negative estimated residual variance (Heywood cases); thus, in practice a researcher would typically reject the three-factor model and interpret the two-factor model. **Table 11** gives the rotated factor pattern for the two-factor models estimated with Case 3 items. Here, the two factors are essentially defined by the skewness direction of the observed variables: odd-numbered items, which are all positively skewed, are predominately determined by η_1 while the negatively skewed even-numbered items are determined by η_2 (with the exception of the Boots variable). A similar factor pattern emerges for the two-factor model estimated with $N = 100$ for the Case 4 items, but not with $N = 200$ (see **Table 12**). Finally, the factor pattern for the three-factor model estimated with Case 4 items is in **Table 13**. Here, the factors have the same basic interpretation as with the population model, but some of the individual factor loadings are quite different. For example, at both sample sizes, the estimated primary factor loadings for Sentence Completion (0.43 with $N = 100$) and Remainders (0.40 with $N = 100$) are much smaller than their population values of 0.77 and 0.80.

In general, the problems with factoring product-moment **R** among item-level variables are less severe when there are more response categories (e.g., more than five) and the observed univariate item distributions are symmetric (Finney and DiStefano, 2006). Our demonstration above is consistent with this statement. Yet, although there are situations where ordinary linear regression of a categorical dependent variable can potentially produce useful results, the field still universally accepts non-linear modeling methods such as logistic regression as standard for limited dependent variables. Similarly, although there may be situations where factoring product-moment **R** (and hence adapting a linear factor model) produces reasonably accurate results, the field should accept alternative, non-linear factor models as standard for categorical, item-level observed variables.

Table 8 | Factor loading matrix obtained with EFA of product-moment R among Case 2 item-level variables.

Variable	Factor					
	$N = 100$			$N = 200$		
	η_1	η_2	η_3	η_1	η_2	η_3
WrdMean	**0.94**	−0.10	0.05	**0.94**	−0.07	0.02
SntComp	**0.73**	0.13	−0.10	**0.82**	0.09	−0.08
OddWrds	**0.72**	0.07	0.09	**0.74**	0.04	0.10
MxdArit	0.11	**0.98**	−0.03	0.04	**0.96**	−0.04
Remndrs	0.11	**0.70**	0.09	0.04	**0.76**	0.12
MissNum	0.13	**0.73**	0.06	0.14	**0.78**	−0.01
Gloves	−0.01	0.17	**0.59**	−0.02	0.10	**0.60**
Boots	0.03	−0.05	**0.81**	0.04	0.07	**0.67**
Hatchts	0.00	−0.02	**0.75**	0.01	−0.13	**0.86**

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

ALTERNATIVE METHODS FOR ITEM-LEVEL OBSERVED VARIABLES

Wirth and Edwards (2007) give a comprehensive review of methods for factor analyzing categorical item-level variables. In general, these methods can be classified as either *limited-information* or *full-information*. The complete data for N participants on p categorical, item-level variables form a multi-way frequency table with C_j^P cells (i.e., $C_1 \times C_2 \times \ldots \times C_p$), or potential response patterns, where C_j is the number of categories for item j. Full-information factor models draw from multidimensional item-response theory (IRT) to predict directly the probability that a given individual's response pattern falls into a particular cell of this multi-way frequency table (Bock et al., 1988). Limited-information methods instead fit the factor model to a set of intermediate summary statistics which are calculated from the observed frequency table. These summary statistics include the univariate response proportions for each item and the bivariate

Table 9 | Product-moment and polychoric correlations among Case 3 item-level variables.

Variable	1	2	3	4	5	6	7	8	9
WrdMean	1	0.57	0.71	0.35	0.42	0.54	0.30	0.17	0.61
SntComp	0.24	1	0.54	0.57	0.50	0.47	0.26	0.45	0.08
OddWrds	0.46	0.22	1	0.60	0.57	0.61	0.64	0.46	0.66
MxdArit	0.16	0.35	0.26	1	0.57	0.83	0.52	0.03	0.19
Remndrs	0.23	0.21	0.33	0.24	1	0.58	0.30	0.22	0.64
MissNum	0.23	0.28	0.26	0.61	0.25	1	0.54	0.29	0.52
Gloves	0.16	0.11	0.39	0.22	0.16	0.23	1	0.32	0.61
Boots	0.08	0.26	0.18	0.02	0.09	0.17	0.14	1	0.47
Hatchts	0.37	0.04	0.41	0.09	0.39	0.22	0.37	0.19	1

$N = 100$. Product-moment correlations are below the diagonal; polychoric correlations are above the diagonal. WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets.

Table 10 | Product-moment and polychoric correlations among Case 4 item-level variables.

Variable	1	2	3	4	5	6	7	8	9
WrdMean	1	0.81	0.75	0.42	0.47	0.41	0.21	0.21	0.29
SntComp	0.52	1	0.62	0.49	0.44	0.55	0.22	0.34	0.13
OddWrds	0.66	0.42	1	0.54	0.47	0.53	0.36	0.42	0.50
MxdArit	0.31	0.40	0.38	1	0.82	0.86	0.32	0.25	0.27
Remndrs	0.43	0.31	0.41	0.54	1	0.76	0.41	0.36	0.42
MissNum	0.32	0.46	0.38	0.79	0.51	1	0.35	0.30	0.33
Gloves	0.20	0.14	0.33	0.25	0.31	0.28	1	0.35	0.51
Boots	0.15	0.29	0.29	0.18	0.24	0.25	0.27	1	0.68
Hatchts	0.29	0.12	0.46	0.19	0.38	0.27	0.42	0.45	1

N = 100. Product-moment correlations are below the diagonal; polychoric correlations are above the diagonal. WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets.

Table 11 | Two-factor model loading matrix obtained with EFA of product-moment R among Case 3 item-level variables.

Variable	Factor			
	N = 100		*N* = 200	
	η_1	η_2	η_1	η_2
WrdMean	**0.53**	0.02	**0.60**	0.00
SntComp	0.12	**0.37**	0.24	0.22
OddWrds	**0.69**	0.04	**0.75**	−0.06
MxdArit	−0.13	**0.95**	−0.01	**0.76**
Remndrs	**0.42**	0.11	**0.37**	0.13
MissNum	0.12	**0.62**	−0.01	**0.87**
Gloves	**0.44**	0.06	**0.35**	0.07
Boots	**0.25**	0.03	**0.32**	−0.04
Hatchts	**0.78**	−0.19	**0.43**	−0.03

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

Table 12 | Two-factor model loading matrix obtained with EFA of product-moment R among Case 4 item-level variables.

Variable	Factor			
	N = 100		*N* = 200	
	η_1	η_2	η_1	η_2
WrdMean	**0.49**	0.21	**0.46**	0.27
SntComp	0.23	**0.40**	**0.55**	0.12
OddWrds	**0.68**	0.14	**0.46**	0.36
MxdArit	−0.11	**0.96**	**0.87**	−0.17
Remndrs	0.34	**0.42**	**0.58**	0.14
MissNum	0.02	**0.84**	**0.88**	−0.13
Gloves	**0.47**	0.04	0.06	**0.45**
Boots	**0.49**	−0.01	0.00	**0.56**
Hatchts	**0.77**	−0.15	−0.06	**0.79**

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

polychoric correlations among items. Hence, these methods are referred to as "limited-information" because they collapse the complete multi-way frequency data into univariate and bivariate marginal information. Full-information factor analysis is an active area of methodological research, but the limited-information method has become quite popular in applied settings and simulation studies indicate that it performs well across a range of situations (e.g., Flora and Curran, 2004; Forero et al., 2009; also see Forero and Maydeu-Olivares, 2009, for a study comparing full- and limited-information modeling). Thus, our remaining presentation focuses on the limited-information, polychoric correlation approach.

The key idea to this approach is the assumption that an unobserved, normally distributed continuous variable, y^*, underlies each categorical, ordinally scaled observed variable, y with response categories $c = 0, 1, \ldots, C$. The latent y^* links to the observed y according to

$$y = c \text{ if } \tau_{c-1} < y^* < \tau_c,$$

where each τ is a threshold parameter (i.e., a Z-score) determined from the univariate proportions of y with $\tau_0 = -\infty$ and $\tau_C = \infty$. Adopting Eq. 1, the common factor model is then a model for the y^* variables themselves[11],

$$\mathbf{y}^* = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}.$$

Like factor analysis of continuous variables, the model parameters are actually estimated from the correlation structure, where the correlations among the y^* variables are polychoric correlations[12]. Thus, the polychoric correlation between two observed, ordinal y

[11]The factor model can be equivalently conceptualized as a *probit* regression for the categorical dependent variables. Probit regression is nearly identical to logistic regression, but the normal cumulative distribution function is used in place of the logistic distribution (see Fox, 2008).

[12]Analogous to the incorporation of mean structure among continuous variables, advanced models such as multiple-group measurement models are fitted to the set of both estimated thresholds and polychoric correlations.

variables is an estimate of the correlation between two unobserved, continuous y^* variables (Olsson, 1979)[13]. Adopting Eq. 3 leads to

$$\mathbf{P}^* = \mathbf{\Lambda\Psi\Lambda}' + \mathbf{\Theta},$$

where $\mathbf{P}^*$ is the population correlation matrix among $\mathbf{y}^*$, which is estimated with the polychoric correlations. Next, because ML estimation provides model estimates that most likely would have produced *observed* multivariate normal data, methodologists do not recommend simply substituting polychoric **R** for product-moment **R** and proceed with ML estimation (e.g., Yang-Wallentin et al., 2010). Instead, we recommend ULS estimation, where the purpose is to minimize squared residual polychoric correlations (Muthén, 1978)[14]. The normality assumption for each unobserved y^* is a mathematical convenience that allows estimation of $\mathbf{P}^*$; Flora and Curran (2004) showed that polychoric correlations remain reasonably accurate under moderate violation of this assumption, which then has only a very small effect on factor analysis results. Finally, this polychoric approach is implemented in several popular SEM software packages (as well as R, SAS, and Stata), each of which is capable of both EFA and CFA.

EXAMPLE DEMONSTRATION CONTINUED

To demonstrate limited-information item factor analysis, we conducted EFA of polychoric correlation matrices among the same categorized variables for Cases 1–4 presented above, again using ULS estimation and quartimin rotation. We begin with analyses of the approximately symmetric Case 1 and Case 2 items. First, the sample polychoric correlations are closer to the known population correlations than were the product-moment correlations among these categorical variables (see **Tables 5** and **7**). For example, the population correlation between Word Meaning and Sentence Completion is 0.75 and the sample polychoric correlation between these two Case 1 items is 0.81, but the product-moment correlation was only 0.60 (both with $N = 100$). As with product-moment **R**, the EFA of polychoric **R** among Case 1 variables also strongly suggested retaining a three-factor model at both $N = 100$ and $N = 200$; the rotated factor-loading matrices are in **Table 14**. Here, the loading of each observed variable on its primary factor is consistently larger than that obtained with the EFA of product-moment **R** (compare to **Table 6**) and much closer to the population factor loadings in **Table 2**. Next, with Case 2 variables, the EFA of polychoric **R** again led to a three-factor model at both $N = 100$ and $N = 200$. The primary factor loadings in **Table 15** were only slightly larger than those obtained with product-moment **R** (in **Table 8**), which were themselves reasonably close to the population factor loadings. Thus, in a situation where strong attenuation of product-moment correlations led to strong underestimation of factor loadings obtained from EFA of product-moment **R** (i.e., Case 1), an alternate EFA of polychoric **R** produced relatively accurate factor loadings. Yet, when attenuation of product-moment correlations is less severe (i.e., Case 2), EFA of polychoric **R** was still just as accurate.

Recall that in Case 3 and 4, some items were positively skewed and some were negatively skewed. First, the polychoric correlations among Case 3 and 4 items are generally closer to the population correlations than were the product-moment correlations, although many of the polychoric correlations among Case 3 dichotomous items are quite inaccurate (see **Tables 9** and **10**). For example, the population correlation between Word Meaning

[13]The *tetrachoric* correlation is a specific type of polychoric correlation that obtains when both observed variables are dichotomous.

[14]Alternatively, "robust weighted least-squares," also known as "diagonally weighted least-squares" (DWLS or WLSMV), is also commonly recommended. Simulation studies suggest that ULS and DWLS produce very similar results, with slightly more accurate estimates obtained with ULS (Forero et al., 2009; Yang-Wallentin et al., 2010).

Table 13 | Three-factor model loading matrix obtained with EFA of product-moment R among Case 4 item-level variables.

Variable	Factor					
	$N = 100$			$N = 200$		
	η_1	η_2	η_3	η_1	η_2	η_3
WrdMean	**1.05**	−0.10	−0.04	**0.95**	−0.10	−0.05
SntComp	**0.43**	0.30	−0.05	**0.50**	0.23	−0.03
OddWrds	**0.54**	0.05	0.32	**0.68**	0.04	0.14
MxdArit	0.17	**0.92**	−0.04	0.16	**0.94**	−0.03
Remndrs	0.17	**0.40**	0.25	0.16	**0.46**	0.17
MissNum	−0.02	**0.87**	0.06	0.03	**0.85**	0.00
Gloves	0.00	0.10	**0.47**	0.00	0.06	**0.46**
Boots	0.01	0.05	**0.48**	0.02	0.00	**0.54**
Hatchts	−0.01	−0.14	**0.95**	−0.03	−0.08	**0.92**

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

Table 14 | Three-factor model loading matrix obtained with EFA of polychoric R among Case 1 item-level variables.

Variable	Factor					
	$N = 100$			$N = 200$		
	η_1	η_2	η_3	η_1	η_2	η_3
WrdMean	**0.95**	−0.07	−0.05	**0.94**	−0.11	−0.02
SntComp	**0.85**	0.14	−0.09	**0.82**	0.19	−0.05
OddWrds	**0.80**	0.02	0.24	**0.81**	0.03	0.14
MxdArit	−0.02	**0.91**	0.04	−0.02	**0.93**	−0.02
Remndrs	−0.04	**0.93**	0.06	−0.05	**0.86**	0.16
MissNum	0.12	**0.82**	−0.12	0.12	**0.87**	−0.10
Gloves	−0.01	0.22	**0.51**	0.04	0.03	**0.59**
Boots	0.10	0.12	**0.68**	0.02	0.03	**0.72**
Hatchts	−0.01	−0.04	**0.96**	0.02	−0.01	**0.92**

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

Table 15 | Three-factor model loading matrix obtained with EFA of polychoric R among Case 2 item-level variables.

Variable	Factor					
	$N = 100$			$N = 200$		
	η_1	η_2	η_3	η_1	η_2	η_3
WrdMean	**0.99**	−0.08	0.04	**0.97**	−0.07	0.02
SntComp	**0.75**	0.17	−0.12	**0.84**	0.10	−0.08
OddWrds	**0.74**	0.10	0.07	**0.77**	0.06	0.10
MxdArit	−0.06	**1.00**	−0.03	−0.06	**1.00**	−0.04
Remndrs	0.11	**0.73**	0.10	0.04	**0.80**	0.12
MissNum	0.13	**0.77**	0.06	0.15	**0.81**	0.00
Gloves	−0.01	0.20	**0.61**	−0.03	0.14	**0.61**
Boots	0.04	−0.04	**0.82**	0.04	0.12	**0.68**
Hatchts	−0.02	0.00	**0.80**	0.01	−0.08	**0.89**

WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

Table 16 | Three-factor model loading matrix obtained with EFA of polychoric R among Case 3 item-level variables.

Variable	Factor		
	η_1	η_2	η_3
WrdMean	**0.52**	0.27	0.18
SntComp	**0.96**	0.03	−0.03
OddWrds	**0.44**	0.34	0.32
MxdArit	0.05	**0.92**	−0.07
Remndrs	0.36	**0.44**	0.07
MissNum	0.06	**0.95**	0.01
Gloves	−0.22	**0.53**	0.42
Boots	0.22	−0.22	**0.74**
Hatchts	−0.09	0.11	**0.82**

$N = 200$. WrdMean, word meaning; SntComp, sentence completion; OddWrds, dd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

Table 17 | Three-factor model loading matrix obtained with EFA of polychoric R among Case 4 item-level variables.

Variable	Factor		
	η_1	η_2	η_3
WrdMean	**0.97**	−0.08	−0.02
SntComp	**0.85**	0.09	−0.04
OddWrds	**0.70**	0.11	0.17
MxdArit	−0.03	**1.01**	−0.05
Remndrs	−0.02	**0.83**	0.12
MissNum	0.12	**0.81**	−0.02
Gloves	0.00	0.11	**0.49**
Boots	0.07	−0.02	**0.69**
Hatchts	−0.02	−0.01	**0.99**

$N = 200$. WrdMean, word meaning; SntComp, sentence completion; OddWrds, odd words; MxdArit, mixed arithmetic; Remndrs, remainders; MissNum, missing numbers; Hatchts, hatchets. Primary loadings for each observed variable are in bold.

and Sentence Completion was 0.75 while the sample polychoric correlation between these oppositely skewed Case 3 items was 0.57 with $N = 100$ and improved to 0.67 with $N = 200$; but the product-moment correlation between these two items was only 0.24 with $N = 100$ and 0.25 with $N = 200$. With five-category Case 4 items, the polychoric correlation between Word Meaning and Sentence Completion is 0.81 with both sample sizes, but the product-moment correlations were only 0.52 with $N = 100$ and 0.51 with $N = 200$.

Given that the polychoric correlations were generally more resistant to skewness in observed items than product-moment correlations, factor analyses of Case 3 and 4 items should also be improved. With the dichotomous Case 3 variables and $N = 100$, the scree plot suggested retaining a one-factor model, but other fit statistics did not support a one-factor model (e.g., RMSR = 0.14). Yet, the estimated two- and three-factor models were improper with negative residual variance estimates. This outcome likely occurred because many of the bivariate frequency tables for item pairs had cells with zero frequency. This outcome highlights the general concern of *sparseness*, or the tendency of highly skewed items to produce observed bivariate distributions with cell frequencies equaling zero or close to zero, especially with relatively small overall sample size. Sparseness can cause biased polychoric correlation estimates, which in turn leads to inaccurate factor analysis results (Olsson, 1979; Savalei, 2011). With $N = 200$, a three-factor model for Case 3 variables was strongly supported; the rotated factor loading matrix is in **Table 16**. Excluding Gloves, each item has its strongest loading on the same factor as indicated in the population (see **Table 2**), but many of these primary factor loadings are strongly biased and many items have moderate cross-loadings. Thus, we see an example of the tendency for EFA of polychoric **R** to produce inaccurate results with skewed dichotomous items. Nonetheless, recall that EFA of product-moment **R** for Case 3 items did not even lead to a model with the correct number of factors.

With the five-category Case 4 items, retention of a three-factor model was supported at both sample sizes. However, with $N = 100$, we again obtained improper estimates. But with $N = 200$, the rotated factor loading matrix (**Table 17**) is quite accurate relative to the population factor loadings, and certainly improved over that obtained with EFA of product-moment **R** among these items. Thus, even though there was a mix of positively and negatively skewed items, having a larger sample size and more item-response categories mitigated the potential for sparseness to produce inaccurate results from the polychoric EFA.

In sum, our demonstration illustrated that factor analyses of polychoric **R** among categorized variables consistently outperformed analyses of product-moment **R** for the same variables. In particular, with symmetric items (Cases 1 and 2),

product-moment correlations were attenuated which led to negatively biased factor loading estimates (especially with fewer response categories), whereas polychoric correlations remained accurate and produced accurate factor loadings. When the observed variable set contained a mix of positively and negatively skewed items (Cases 3 and 4), product-moment correlations were strongly affected by the direction of skewness (especially with fewer response categories), which can lead to dramatically misleading factor analysis results. Unfortunately, strongly skewed items can also be problematic for factor analyses of polychoric **R** in part because they produce sparse observed frequency tables, which leads to higher rates of improper solutions and inaccurate results (e.g., Flora and Curran, 2004; Forero et al., 2009). Yet this difficulty is alleviated with larger sample size and more item-response categories; if a proper solution is obtained, then the results of a factor analysis of polychoric **R** are more trustworthy than those obtained with product-moment **R**.

GENERAL DISCUSSION AND CONCLUSION

Factor analysis is traditionally and most commonly an analysis of the correlations (or covariances) among a set of observed variables. A unifying theme of this paper is that if the correlations being analyzed are misrepresentative or inappropriate summaries of the relationships among the variables, then the factor analysis is compromised. Thus, the process of data screening and assumption testing for factor analysis should begin with a focus on the adequacy of the correlation matrix among the observed variables. In particular, analysis of product-moment correlations or covariances implies that the bivariate association between two observed variables can be adequately modeled with a straight line, which in turn leads to the expression of the common factor model as a linear regression model. This crucial linearity assumption takes precedence over any concerns about having normally distributed variables, although normality is important for certain model fit statistics and estimating parameter SEs. Other concerns about the appropriateness of the correlation matrix involve collinearity and potential influence of unusual cases. Once one accepts the sufficiency of a given correlation matrix as a representation of the observed data, the actual formal assumptions of the common factor model are relatively mild. These assumptions are that the unique factors (i.e., the residuals, $\boldsymbol{\varepsilon}$, in Eq. 1) are uncorrelated with each other and are uncorrelated with the common factors (i.e., $\boldsymbol{\eta}$ in Eq. 1). Substantial violation of these assumptions typically manifests as poor model-data fit, and is otherwise difficult to assess with *a priori* data-screening procedures based on descriptive statistics or graphs.

After reviewing the common factor model, we gave separate presentations of issues concerning factor analysis of continuous observed variables and issues concerning factor analysis of categorical, item-level observed variables. For the former, we showed how concepts from regression diagnostics apply to factor analysis, given that the common factor model is itself a multivariate multiple regression model with unobserved explanatory variables. An important point was that cases that appear as outliers in univariate or bivariate plots are not necessarily influential and conversely that influential cases may not appear as outliers in univariate or bivariate plots (though they often do). If one can determine that unusual observations are not a result of researcher or participant error, then we recommend the use of robust estimation procedures instead of deleting the unusual observation. Likewise, we also recommend the use of robust procedures for calculating model fit statistics and SEs when observed, continuous variables are non-normal.

Next, a crucial message was that the linearity assumption is necessarily violated when the common factor model is fitted to product-moment correlations among categorical, ordinally scaled items, including the ubiquitous Likert-type items. At worst (e.g., with a mix of positively and negatively skewed dichotomous items), this assumption violation has severe consequences for factor analysis results. At best (e.g., with symmetric items with five response categories), this assumption violation still produces biased factor-loading estimates. Alternatively, factor analysis of polychoric **R** among item-level variables explicitly specifies a nonlinear link between the common factors and the observed variables, and as such is theoretically well-suited to the analysis of item-level variables. However, this method is also vulnerable to certain data characteristics, particularly sparseness in the bivariate frequency tables for item pairs, which occurs when strongly skewed items are analyzed with a relatively small sample. Yet, factor analysis of polychoric **R** among items generally produces superior results compared to those obtained with product-moment **R**, especially if there are five or fewer item-response categories.

We have not yet directly addressed the role of sample size. In short, no simple rule-of-thumb regarding sample size is reasonably generalizable across factor analysis applications. Instead, adequate sample size depends on many features of the research, such as the major substantive goals of the analysis, the number of observed variables per factor, closeness to simple structure, and the strength of the factor loadings (MacCallum et al., 1999, 2001). Beyond these considerations, having a larger sample size can guard against some of the harmful consequences of unusual cases and assumption violation. For example, unusual cases are less likely to exert strong influence on model estimates as overall sample size increases. Conversely, removing unusual cases decreases the sample size, which reduces the precision of parameter estimation and statistical power for hypothesis tests about model fit or parameter estimates. Yet, having a larger sample size does not protect against the negative consequences of treating categorical item-level variables as continuous by factor analyzing product-moment **R**. But we did illustrate that larger sample size produces better results for factor analysis of polychoric **R** among strongly skewed items, in part because larger sample size reduces the occurrence of sparseness.

In closing, we emphasize that factor analysis, whether EFA or CFA, is a method for *modeling* relationships among observed variables. It is important for researchers to recognize that it is impossible for a statistical model to be perfect; assumptions will always be violated to some extent in that no model can ever exactly capture the intricacies of nature. Instead, researchers should strive to find models that have an approximate fit to data such that the inevitable assumption violations are trivial, but the models can still provide useful results that help answer important substantive research questions (see MacCallum, 2003, and Rodgers, 2010, for discussions of this principle). We recommend extensive use of *sensitivity analyses* and *cross-validation* to aid in this endeavor. For example, researchers should compare

results obtained from the same data using different estimation procedures, such as comparing traditional ULS or ML estimation with robust procedures with continuous variables or comparing full-information factor analysis results with limited-information results with item-level variables. Additionally, as even CFA analyses may become exploratory through model modification, it is important to cross-validate models across independent data sets. Because different modeling procedures place different demands on data, comparing results obtained with different methods and samples can help researchers gain a fuller, richer understanding of the usefulness of their statistical models given the natural complexity of real data.

REFERENCES

Asparouhov, T., and Muthén, B. (2009). Exploratory structural equation modeling. *Struct. Equ. Modeling* 16, 397–438.

Babakus, E., Ferguson, C. E., and Joreskög, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *J. Mark. Res.* 37, 72–141.

Bartholomew, D. J. (2007). "Three faces of factor analysis," in *Factor Analysis at 100: Historical Developments and Future Directions*, eds R. Cudeck and R. C. MacCallum (Mahwah, NJ: Lawrence Erlbaum Associates), 9–21.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Bentler, P. M. (2004). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M. (2007). Can scientifically useful hypotheses be tested with correlations? *Am. Psychol.* 62, 772–782.

Bentler, P. M., and Savalei, V. (2010). "Analysis of correlation structures: current status and open problems," in *Statistics in the Social Sciences: Current Methodological Developments*, eds S. Kolenikov, D. Steinley, and L. Thombs, (Hoboken, NJ: Wiley), 1–36.

Bernstein, I. H., and Teng, G. (1989). Factoring items and factoring scales are different: spurious evidence for multidimensionality due to item categorization. *Psychol. Bull.* 105, 467–477.

Bock, R. D., Gibbons, R., and Muraki, E. (1988). Full-information item factor analysis. *Appl. Psychol. Meas.* 12, 261–280.

Bollen, K. A. (1987). Outliers and improper solutions: a confirmatory factor analysis example. *Sociol. Methods Res.* 15, 375–384.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K. A., and Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociol. Methodol.* 21, 235–262.

Briggs, N. E., and MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behav. Res.* 38, 25–56.

Browne, M. W., Cudeck, R., Tateneni, K., and Mels, G. (2010). *CEFA: Comprehensive Exploratory Factor Analysis, Version 3. 03 [Computer Software, and Manual]*. Available at: http://faculty.psy.ohio-state.edu/browne/

Cadigan, N. G. (1995). Local influence in structural equation models. *Struct. Equ. Modeling* 2, 13–30.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychol. Bull.* 105, 317–327.

Cudeck, R., and O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: significance tests for factor loadings and correlations. *Psychol. Bull.* 115, 475–487.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: a comparison of categorical variable estimators using simulated data. *Br. J. Math. Stat. Psychol.* 47, 309–326.

Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika* 6, 323–329.

Finney, S. J., and DiStefano, C. (2006). "Non-normal and categorical data in structural equation modeling," in *Structural Equation Modeling: A Second Course*, eds G. R. Hancock and R. O. Mueller (Greenwich, CT: Information Age Publishing), 269–314.

Flora, D. B., and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol. Methods* 9, 466–491.

Forero, C. G., and Maydeu-Olivares, A. (2009). Estimation of IRT graded models for rating data: limited vs. full information methods. *Psychol. Methods* 14, 275–299.

Forero, C. G., Maydeu-Olivares, A., and Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: a Monte Carlo study comparing DWLS and ULS estimation. *Struct. Equ. Modeling* 16, 625–641.

Fox, J. (1991). *Regression Diagnostics: An Introduction*. Thousand Oaks, CA: SAGE Publications.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*, 2nd Edn. Thousand Oaks, CA: SAGE Publications.

Gorsuch, R. L. (1983). *Factor Analysis*, 2nd Edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., and Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Struct. Equ. Modeling* 4, 108–120.

Harman, H. H. (1960). *Modern Factor Analysis*. Chicago, IL: University of Chicago Press.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34, 183–202.

Lawley, D. N., and Maxwell, A. E. (1963). *Factor Analysis as a Statistical Method*. London: Butterworth.

Lee, S.-Y., and Wang, S. J. (1996). Sensitivity analysis of structural equation models. *Psychometrika* 61, 93–108.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behav. Res.* 38, 113–139.

MacCallum, R. C. (2009). "Factor analysis," in *The SAGE Handbook of Quantitative Methods in Psychology*, eds R. E. Millsap and A. Maydeu-Olivares (Thousand Oaks, CA: SAGE Publications), 123–147.

MacCallum, R. C., Widaman, K. F., Preacher, K., and Hong, S. (2001). Sample size in factor analysis: the role of model error. *Multivariate Behav. Res.* 36, 611–637.

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychol. Methods* 4, 84–99.

Mavridis, D., and Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behav. Res.* 43, 453–475.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika* 43, 551–560.

Muthén, B., and Kaplan, D. (1985). A comparison of some methodologies for the factor-analysis of non-normal Likert variables. *Br. J. Math. Stat. Psychol.* 38, 171–180.

Muthén, B., and Kaplan, D. (1992). A comparison of some methodologies for the factor-analysis of non-normal Likert variables: a note on the size of the model. *Br. J. Math. Stat. Psychol.* 45, 19–30.

Muthén, L., and Muthén, B. (2010). *Mplus User's Guide*, 6th Edn. Los Angeles, CA: Muthén & Muthén.

Nunnally, J.C., and Bernstein, I. H. (1994). *Psychometric Theory*, 3rd Edn. New York: McGraw-Hill.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* 44, 443–460.

Pek, J., and MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: cases and their influence. *Multivariate Behav. Res.* 46, 202–228.

Pison, G., Rousseeuw, P. J., Filzmoser, P., and Croux, C. (2003). Robust factor analysis. *J. Multivar. Anal.* 84, 145–172.

Poon, W.-Y., and Wong, Y.-K. (2004). A forward search procedure for identifying influential observations in the estimation of a covariance matrix. *Struct. Equ. Modeling* 11, 357–374.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *J. R. Stat. Soc. Series B* 31, 350–371.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *Am. Psychol.* 65, 1–12.

Satorra, A., and Bentler, P. M. (1994). "Corrections to test statistics and standard errors in covariance structure analysis," in *Latent Variables Analysis: Applications for Developmental Research*, eds A. von Eye and C. C. Clogg (Thousand Oaks, CA: SAGE Publications), 399–419.

Savalei, V. (2011). What to do about zero frequency cells when estimating polychoric correlations. *Struct. Equ. Modeling* 18, 253–273.

Spearman, C. (1904). General intelligence objectively determined and measured. *Am. J. Psychol.* 5, 201–293.

Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: University of Chicago Press.

Weisberg, S. (1985). *Applied Linear Regression*, 2nd Edn. New York: Wiley.

West, S. G., Finch, J. F., and Curran, P. J. (1995). "Structural equation models with non-normal variables: problems and remedies," in *Structural Equation Modeling: Concepts, Issues, and Applications*, ed. R. H. Hoyle (Thousand Oaks, CA: SAGE Publications), 56–75.

Wirth, R. J., and Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychol. Methods* 12, 58–79.

Yang-Wallentin, F., Jöreskog, K. G., and Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Struct. Equ. Modeling* 17, 392–423.

Yuan, K.-H., and Bentler, P. M. (1998). Structural equation modeling with robust covariances. *Sociol. Methodol.* 28, 363–396.

Yuan, K.-H., Marshall, L. L., and Bentler, P. M. (2002). A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika* 67, 95–122.

Yuan, K.-H., and Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: using robust procedures to minimize their effect. *Sociol. Methodol.* 38, 329–368.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 October 2011; paper pending published: 21 November 2011; accepted: 13 February 2012; published online: 01 March 2012.

*Citation: Flora DB, LaBrish C and Chalmers RP (2012) Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. Front. Psychology **3**:55. doi: 10.3389/fpsyg.2012.00055*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 27 March 2013
doi: 10.3389/fpsyg.2013.00109

On the relevance of assumptions associated with classical factor analytic approaches†

Daniel Kasper and Ali Ünlü*

Chair for Methods in Empirical Educational Research, TUM School of Education and Centre for International Student Assessment, Technische Universität München, Munich, Germany

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Andrew Jones, American Board of Surgery, USA
Evgueni Borokhovski, Concordia University, Canada

***Correspondence:**
Ali Ünlü, Chair for Methods in Empirical Educational Research, TUM School of Education and Centre for International Student Assessment, Technische Universität München, Lothstrasse 17, 80335 Munich, Germany.
e-mail: ali.uenlue@tum.de

†*The research reported in this paper is based on the dissertation thesis by Kasper (2012).*

A personal trait, for example a person's cognitive ability, represents a theoretical concept postulated to explain behavior. Interesting constructs are latent, that is, they cannot be observed. Latent variable modeling constitutes a methodology to deal with hypothetical constructs. Constructs are modeled as random variables and become components of a statistical model. As random variables, they possess a probability distribution in the population of reference. In applications, this distribution is typically assumed to be the normal distribution. The normality assumption may be reasonable in many cases, but there are situations where it cannot be justified. For example, this is true for criterion-referenced tests or for background characteristics of students in large scale assessment studies. Nevertheless, the normal procedures in combination with the classical factor analytic methods are frequently pursued, despite the effects of violating this "implicit" assumption are not clear in general. In a simulation study, we investigate whether classical factor analytic approaches can be instrumental in estimating the factorial structure and properties of the population distribution of a latent personal trait from educational test data, when violations of classical assumptions as the aforementioned are present. The results indicate that having a latent non-normal distribution clearly affects the estimation of the distribution of the factor scores and properties thereof. Thus, when the population distribution of a personal trait is assumed to be non-symmetric, we recommend avoiding those factor analytic approaches for estimation of a person's factor score, even though the number of extracted factors and the estimated loading matrix may not be strongly affected. An application to the Progress in International Reading Literacy Study (PIRLS) is given. Comments on possible implications for the Programme for International Student Assessment (PISA) complete the presentation.

Keywords: factor analysis, latent variable model, normality assumption, factorial structure, criterion-referenced test, large scale educational assessment, Programme for International Student Assessment, Progress in International Reading Literacy Study

1. INTRODUCTION

Educational research is concerned with the study of processes of learning and teaching. Typically, the investigated processes are not observable, and to unveil these, manifest human behavior in test situations is recorded. According to Lienert and Raatz (1998, p. 1) "a test [. . .] is a routine procedure for the investigation of one or more empirically definable personality traits" (translated by the authors), and to satisfy a minimum of quality criteria, a test is required to be objective, reliable, and valid.

In this paper we deal with factor analytic methods for assessing construct validity of a test, in the sense of its factorial validity (e.g., Cronbach and Meehl, 1955; Lienert and Raatz, 1998). Factorial validity refers to the factorial structure of the test, that is, to the number (and interpretation) of underlying factors, the correlation structure among the factors, and the correlations of each test item with the factors. There are a number of latent variable models that may be used to analyze the factorial structure of a test – for generalized latent variable modeling covering a plethora of models as special cases of a much broader framework, see Bartholomew et al. (2011) and Skrondal and Rabe-Hesketh (2004). This paper focuses on classical factor analytic approaches, and it examines how accurately different methods of classical factor analysis can estimate the factorial structure of test data, if assumptions associated with the classical approaches are not satisfied. The methods of classical factor analysis will include principal component analysis (PCA; Pearson, 1901; Hotelling, 1933a,b; Kelley, 1935), exploratory factor analysis (EFA; Spearman, 1904; Burt, 1909; Thurstone, 1931, 1965), and principal axis analysis (PAA; Thurstone, 1931, 1965). More recent works on factor analysis and related methods are Harman (1976), McDonald (1985), Cudeck and MacCallum (2007), and Mulaik (2009). Further references, to more specific topics in factor analysis, are given below, later in the text.[1]

[1]For the sake of simplicity and for the purpose and analysis of this paper, we want to refer to all of these approaches (PCA, EFA, PAA) collectively as classical factor analysis/analytic methods. Albeit it is known that PCA differs from factor analysis in important aspects, and that PAA rather represents an alternative estimation procedure for EFA. PCA and EFA are different technically and conceptually. PCA

A second objective of this paper is to examine the scope of these classical methods for estimating the probability distribution of latent ability values or properties thereof postulated in a population under investigation, especially when this distribution is skewed (and not normal). In applied educational contexts, for instance, that is not seldom the practice. Therefore a critical evaluation of this usage of classical factor analytic methods for estimating distributional properties of ability is important, as we do present with our simulation study in this paper, in which metric scale (i.e., at least interval scale; not dichotomous) items are used.

The results of the simulation study indicate that having a non-normal distribution for latent variables does not strongly affect the number of extracted factors and the estimation of the loading matrix. However, as shown in this paper, it clearly affects the estimation of the latent factor score distribution and properties thereof (e.g., skewness).

More precisely, the "estimation accuracy" for factorial structure of these models is shown to be worse when the assumption of interval-scaled data is not met or item statistics are skewed. This corroborates related findings published in other works, which we briefly review later in this paper. More importantly, the empirical distribution of estimated latent ability values is biased compared to the true distribution (i.e., estimates deviate from the true values) when population abilities are skewly distributed. It seems therefore that classical factor analytic procedures, even though they are performed with metric (instead of non-metric) scale indicator variables, are not appropriate approaches to ability estimation when skewly distributed population ability values are to be estimated.

Why should that be of interest? In large scale assessment studies such as the Programme for International Student Assessment (PISA)[2] latent person-related background (conditioning) variables such as sex or socioeconomic status are obtained as well by principal component analysis, and that "covariate" information is part of the PISA procedure that assigns to students their literacy or plausible values (OECD, 2012; see also Section 3.1 in the present paper). Now, if it is assumed that the distribution of latent background information conducted through questionnaires at the students, schools, or parents levels (the true latent variable distribution) is skewed, based on the simulation study of this paper we can expect that the empirical distribution of estimated background information (the "empirical" distribution of the calculated component scores) is biased compared to the true distribution (and is most likely skewed as well). In other words, estimated background values do deviate from their corresponding true values they ought to approximate, and so the inferred students' plausible values may be biased. Further research is necessary in order to investigate the effects and possible implications of potentially biased estimates of latent background information on students' assigned literacy values and competence levels, based on which the PISA rankings of OECD countries are reported. For an analysis of empirical large scale assessment (Progress in International Reading Literacy Study; PIRLS) data, see Section 6.

The paper is structured as follows. We introduce the considered classical factor analysis models in Section 2 and discuss the relevance of the assumptions associated with these models in Section 3. We describe the simulation study in Section 4 and present the results of it in Section 5. We give an empirical data analysis example in Section 6. In Section 7, we conclude with a summary of the main findings and an outlook on possible implications and further research.

2. CLASSICAL FACTOR ANALYSIS METHODS

We consider the method of principal component analysis on the one hand, and the method of exploratory factor and principal axis analysis on the other. At this point recall Footnote 1, where we clarified that, strictly speaking, principal component analysis is not factor analysis and that principal axis analysis is a specific method for estimating the exploratory factor analysis model. Despite this, for the sake of simplicity and for our purposes and analyses, we call these approaches collectively factor analysis/analytic methods or even models. For a more detailed discussion of these methods, see Bartholomew et al. (2011).

Our study shows, amongst others, that the purely computational dimensionality reduction method PCA performs surprisingly well, as compared to the results obtained based on the latent variable models EFA and PAA. This is important, because applied researchers often use PCA in situations where factor analysis more closely matches their purpose of analysis. In general, such computational procedures as PCA are easy to use. Moreover, the comparison of EFA (based on ML) with PAA (eigenstructure of the reduced correlation matrix based on communality estimates) in this paper represents an evaluation of different estimation procedures for the classical factor analysis model. This comparison of the two estimation procedures seems to be justified and interesting, as the (manifest) normality assumption in the observed indicators for the ML procedure is violated, both in the simulation study and empirical large scale assessment PIRLS application. At this point, see also Footnote 1.

seeks to create composite scores of observed variables while EFA assumes latent variables. There is no latent variable in PCA. PCA is not a model and instead is simply a re-expression of variables based on the eigenstructure of their correlation matrix. A statistical model, as is for EFA, is a simplification of observed data that necessarily does not perfectly reproduce the data, leading to the inclusion of an error term. This point is well-established in the methodological literature (e.g., Velicer and Jackson, 1990; Widaman, 2007). Correlation matrix is usually used in EFA, and the models for EFA and PAA are the same. There are several methods to fit EFA such as unweighted least squares (ULS), generalized least squares (GLS), or maximum likelihood (ML). PAA is just one of the various methods to fit EFA. PAA is a method of estimating the model of EFA that does not rely on a discrepancy function such as for ULS, GLS, or ML. This point is made clear, for instance, in MacCallum (2009). In fact, PAA with iterative communality estimation is asymptotically equivalent to ULS estimation. Applied researchers often use PCA in situations where factor analysis more closely matches the purpose of their analysis. This is why we want to include PCA in our present study with latent variables, to examine how well PCA results may approximate a factor analysis model. Such practice is frequently pursued, for example in empirical educational research, as we tried to criticize for the large scale assessment PISA study (e.g., OECD, 2005, 2012). Moreover, the comparison of EFA (based on ML) with PAA in this paper seems to be justified and interesting, as the (manifest) normality assumption in the observed indicator variables for the ML procedure is violated in the simulation study and empirical large scale assessment PIRLS application.

[2]PISA is an international large scale assessment study funded by the Organisation for Economic Co-operation and Development (OECD), which aims to evaluate education systems worldwide by assessing 15-year-old students' competencies in reading, mathematics, and science. For comprehensive and detailed information, see www.pisa.oecd.org.

2.1. PRINCIPAL COMPONENT ANALYSIS

The model of principal component analysis (PCA) is

$$\boldsymbol{Z} = \boldsymbol{FL'},$$

where $\boldsymbol{Z}$ is a $n \times p$ matrix of standardized test results of n persons on p items, $\boldsymbol{F}$ is a $n \times p$ matrix of p principal components ("factors"), and $\boldsymbol{L}$ is a $p \times p$ loading matrix.[3] In the estimation (computation) procedure $\boldsymbol{F}$ and $\boldsymbol{L}$ are determined as $\boldsymbol{F} = \boldsymbol{ZC\Lambda}^{-1/2}$ and $\boldsymbol{L} = \boldsymbol{C\Lambda}^{1/2}$ with a $p \times p$ matrix $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_p\}$, where λ_l are the eigenvalues of the empirical correlation matrix $\boldsymbol{R} = \boldsymbol{Z'Z}$, and with a $p \times p$ matrix $\boldsymbol{C} = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_p})$ of corresponding eigenvectors $\boldsymbol{c_l}$.

In principal component analysis we assume that $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$, $\boldsymbol{F} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{L} \in \mathbb{R}^{n \times p}$ and that empirical moments of the manifest variables exist such that, for any manifest variable $j = 1, \ldots, p$, its empirical variance is not zero ($s_j^2 \neq 0$). Moreover we assume that $\text{rk}(\boldsymbol{Z}) = \text{rk}(\boldsymbol{R}) = p$ (rk, the matrix rank) and that $\boldsymbol{Z}$, $\boldsymbol{F}$, and $\boldsymbol{L}$ are interval-scaled (at the least).

The relevance of the assumption of interval-scaled variables for classical factor analytic approaches is the subject matter of various research works, which we briefly discuss later in this paper.

2.2. EXPLORATORY FACTOR ANALYSIS

The model of exploratory factor analysis (EFA) is

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Lf} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is a $p \times 1$ vector of responses on p items, $\boldsymbol{\mu}$ is the $p \times 1$ vector of means of the p items, $\boldsymbol{L}$ is a $p \times k$ matrix of factor loadings, $\boldsymbol{f}$ is a $k \times 1$ vector of ability values (of factor scores) on k latent continua (on factors), and $\boldsymbol{e}$ is a $p \times 1$ vector subsuming remaining item specific effects or measurement errors.

In exploratory factor analysis, we assume that

$$\begin{aligned}
&\boldsymbol{y} \in \mathbb{R}^{p\times 1}, \boldsymbol{\mu} \in \mathbb{R}^{p\times 1}, \boldsymbol{L} \in \mathbb{R}^{p\times k}, \boldsymbol{f} \in \mathbb{R}^{k\times 1}, \quad \text{and} \quad \boldsymbol{e} \in \mathbb{R}^{p\times 1},\\
&\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{L}, \boldsymbol{f}, \text{and } \boldsymbol{e} \text{ are interval-scaled (at the least)},\\
&\text{E}(\boldsymbol{f}) = \boldsymbol{0},\\
&\text{E}(\boldsymbol{e}) = \boldsymbol{0},\\
&\text{cov}(\boldsymbol{e},\boldsymbol{e}) = \text{E}(\boldsymbol{ee'}) = \boldsymbol{D} = \text{diag}\{v_1, \ldots, v_p\},\\
&\text{cov}(\boldsymbol{f},\boldsymbol{e}) = \text{E}(\boldsymbol{fe'}) = \boldsymbol{0},
\end{aligned}$$

where v_i are the variances of e_i ($i = 1, \ldots, p$). If the factors are not correlated, we call this the orthogonal factor model; otherwise it is called the oblique factor model. In this paper, we investigate the sensitivity of the classical factor analysis model against violated assumptions only for the orthogonal case (with $\text{cov}(\boldsymbol{f}, \boldsymbol{f}) = \text{E}(\boldsymbol{ff'}) = \boldsymbol{I} = \text{diag}\{1, \ldots, 1\}$).

Under this orthogonal factor model, $\boldsymbol{\Sigma}$ can be decomposed as follows:

$$\boldsymbol{\Sigma} = \text{E}\left[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'\right] = \text{E}\left[(\boldsymbol{Lf} + \boldsymbol{e})(\boldsymbol{Lf} + \boldsymbol{e})'\right] = \boldsymbol{LL'} + \boldsymbol{D}.$$

This decomposition is utilized by the methods of unweighted least squares (ULS), generalized least squares (GLS), or maximum likelihood (ML) for the estimation of $\boldsymbol{L}$ and $\boldsymbol{D}$. For ULS and GLS, the corresponding discrepancy function is minimized with respect to $\boldsymbol{L}$ and $\boldsymbol{D}$ (Browne, 1974). ML estimation is performed based on the partial derivatives of the logarithm of the Wishart (W) density function of the empirical covariance matrix $\boldsymbol{S}$, with $(n-1)\,\boldsymbol{S} \sim \text{W}\,(\boldsymbol{\Sigma}, n-1)$ (Jöreskog, 1967). After estimates for $\boldsymbol{\mu}$, k, $\boldsymbol{L}$, and $\boldsymbol{D}$ are obtained, the vector $\boldsymbol{f}$ can be estimated by $\hat{\boldsymbol{f}} = (\boldsymbol{L'D}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L'D}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$.

When applying this exploratory factor analysis, $\boldsymbol{y}$ is typically assumed to be normally distributed, and hence $\text{rk}(\boldsymbol{\Sigma}) = p$, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{y}$. For instance, one condition required for ULS or GLS estimation is that the fourth cumulants of $\boldsymbol{y}$ must be zero, which is the case, for example, if $\boldsymbol{y}$ follows a multivariate normal distribution (for this and other conditions, see Browne, 1974). For ML estimation note that $(n-1)\boldsymbol{S} \sim \text{W}(\boldsymbol{\Sigma}, n-1)$ if $\boldsymbol{y} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Another possibility of estimation for the EFA model is principal axis analysis (PAA). The model of PAA is

$$\boldsymbol{Z} = \boldsymbol{FL'} + \boldsymbol{E},$$

where $\boldsymbol{Z}$ is a $n \times p$ matrix of standardized test results, $\boldsymbol{F}$ is a $n \times p$ matrix of factor scores, $\boldsymbol{L}$ is a $p \times p$ matrix of factor loadings, and $\boldsymbol{E}$ is a $n \times p$ matrix of error terms. For estimation of $\boldsymbol{F}$ and $\boldsymbol{L}$ based on the representation $\boldsymbol{Z'Z} = \boldsymbol{R} = \boldsymbol{LL'} + \boldsymbol{D}$ the principal components transformation is applied. However, the eigenvalue decomposition is not based on $\boldsymbol{R}$, but is based on the reduced correlation matrix $\boldsymbol{R}_h = \boldsymbol{R} - \hat{\boldsymbol{D}}$, where $\hat{\boldsymbol{D}}$ is an estimate for $\boldsymbol{D}$. An estimate $\hat{\boldsymbol{D}}$ is derived using $h_j^2 = 1 - v_j$ and estimating the communalities h_j (for methods for estimating the communalities, see Harman, 1976).

The assumptions of principal axis analysis are

$$\begin{aligned}
&\boldsymbol{Z} \in \mathbb{R}^{n\times p}, \boldsymbol{L} \in \mathbb{R}^{p\times p}, \boldsymbol{F} \in \mathbb{R}^{n\times p}, \quad \text{and} \quad \boldsymbol{E} \in \mathbb{R}^{n\times p},\\
&\text{E}(\boldsymbol{f}) = \boldsymbol{0},\\
&\text{E}(\boldsymbol{e}) = \boldsymbol{0},\\
&\text{cov}(\boldsymbol{e},\boldsymbol{e}) = \text{E}(\boldsymbol{ee'}) = \boldsymbol{D} = \text{diag}\{v_1, \ldots, v_p\},\\
&\text{cov}(\boldsymbol{f},\boldsymbol{e}) = \text{E}(\boldsymbol{fe'}) = \boldsymbol{0},\\
&\text{cov}(\boldsymbol{f},\boldsymbol{f}) = \text{E}(\boldsymbol{ff'}) = \boldsymbol{I},
\end{aligned}$$

and empirical moments of the manifest variables are assumed to exist such that, for any manifest variable $j = 1, \ldots, p$, its empirical variance is not zero ($s_j^2 \neq 0$). Moreover, we assume that $\text{rk}(\boldsymbol{Z}) = \text{rk}(\boldsymbol{R}) = p$ and that the matrices $\boldsymbol{Z}$, $\boldsymbol{F}$, $\boldsymbol{L}$, and $\boldsymbol{E}$ are interval-scaled (at the least).

2.3. GENERAL REMARKS

Two remarks are important before we discuss the assumptions associated with the classical factor models in the next section.

First, it can be shown that $\boldsymbol{L}$ is unique up to an orthogonal transformation. As different orthogonal transformations may yield different correlation patterns, a specific orthogonal transformation must be taken into account (and fixed) before the estimation

[3] For the sake of simplicity and without ambiguity, in this paper we want to refer to component scores from PCA as "factor scores" or "ability values," albeit components conceptually may not be viewed as latent variables or factors. See also Footnote 1.

accuracies of the factor models can be compared. This is known as "rotational indeterminacy" in the factor analysis approach (e.g., see Maraun, 1996). For more information, the reader is also referred to Footnote 8 and Section 7.

Second, the criterion used to determine the number of factors extracted from the data must be distinguished as well. In practice, not all k or p but instead $\hat{k} < k$ or p factors with the $\hat{k}$ largest eigenvalues are extracted. Various procedures are available to determine $\hat{k}$. Commonly used criteria in educational research are the Kaiser-Guttman criterion (Guttman, 1954; Kaiser and Dickman, 1959), the scree test (Cattell, 1966), and the method of parallel analysis (Horn, 1965).

3. ASSUMPTIONS ASSOCIATED WITH THE CLASSICAL FACTOR MODELS

The three models described in the previous section in particular assume interval-scaled data and full rank covariance or correlation matrices for the manifest variables. Typically in the exploratory factor analysis model, the manifest variables $\boldsymbol{y}$ or the standardized variables $\boldsymbol{z}$ are assumed to be normally distributed. For the PCA and PAA models, we additionally want to presuppose – for computational reasons – that the variances of the manifest variables are substantially large. The EFA and PAA models assume uncorrelated factor terms and uncorrelated error terms (which can be relaxed in the framework of structural equation models; e.g., Jöreskog, 1966), uncorrelatedness between the error and latent ability variables, and expected values of zero for the errors as well as latent ability variables.

The question now arises whether the assumptions are critical when it comes to educational tests or survey data?[4]

3.1. CRITERION-REFERENCED TESTS AND PISA QUESTIONNAIRE DATA

From the perspective of applying these models to data of criterion-referenced tests, the last three of the above mentioned assumptions are less problematic. For a criterion-referenced test, it is important that all items of the test are valid for the investigated content. As such, the usual way of excluding items from the analysis when the covariance or correlation matrices are not of full rank does not work for criterion-referenced tests, because this can reduce content validity of a test. A similar argument applies to the assumption of substantially large variances of the manifest variables. As Klauer (1987) suggested and Sturzbecher et al. (2008) have shown for the driving license test in Germany, the variances of the manifest variables of criterion-referenced tests are seldom high, and in general the data obtained from those tests may lead to extracting too few dimensions. However, for the analysis of criterion-referenced tests, the assumption of interval-scaled data and the assumption of normality of the manifest test and latent ability scores are even more problematic. Data from criterion-referenced tests are rarely interval-scaled – instead the items of criterion-referenced tests are often dichotomous (Klauer, 1987). For criterion-referenced tests, it is plausible to have skewed (non-symmetric) test and ability score distributions, because criterion-referenced tests are constructed to assess whether a desired and excessive teaching goal has been achieved or not. In other words, the tested population is explicitly and intensively trained regarding the evaluated ability, and so it is rather likely that most people will have high values on the manifest test score as well as latent ability (e.g., see the German driving license test; Sturzbecher et al., 2008).

The assumption of interval-scaled data and the normality assumption for the manifest test and latent ability scores may also be crucial for the scaling of cognitive data in PISA (OECD, 2012; Chap. 9 therein). In PISA, the generated students' scores are plausible values. These are randomly drawn realizations basically from a multivariate normal distribution (as the prior) of latent ability values (person ability is modeled as a random effect, a latent variable), in correspondence to a fitted item response theory model (Adams et al., 1997) giving the estimated parameters of the normal distribution. The mean of the multivariate normal distribution is expressed as linear regression of various direct manifest regressors (e.g., administered test booklet, gender) and indirect "latent" or complex regressors obtained by aggregating over manifest and latent context or background variables (e.g., indicators for economic, social, and cultural status) in a principal component analysis. The component scores used in the scaling model as the indirect "latent" regressors are extracted, in the purely computational sense, to account for approximately 95% of the total variance in all the original variables. The background variables may be categorical or dummy-coded and may not be measured at an interval scale (nor be normally distributed). So as we said before, if one can assume that the distribution of latent background information revealed through questionnaires is skewed, we can expect that the empirical distribution of background information computed by principal component analysis is likely to be biased compared to the true distribution. This is suggested by the results of our simulation study. The bias of the empirical distribution in turn may result in biasing the regression expression for the mean. Therefore, special caution has to be taken regarding possible violations of those assumptions, and a minimum of related sensitivity analyses are required and necessary in order to control for their potential effects.

3.2. HISTORICAL REMARKS

The primary aim is to review results of previous studies focusing on the impact of violations of model assumptions. As to our knowledge, such studies did not systematically vary the distributions of the factors (in the case of continuous data as well) and

[4]Note that, at the latent level, there is no formal assumption that the latent factors (what we synonymously also want to call "person abilities") are normally distributed. At the manifest level, maximum likelihood estimation (EFA) assumes that the observed variables are normal; ULS and GLS (EFA), PAA (EFA), and PCA do not. The latter two methods only require a non-singular correlation matrix (e.g., see MacCallum, 2009). However, in applications, for example in empirical educational research, one often assumes that the latent ability values follow a normal distribution in the population of reference. Moreover, Mattson (1997)'s method described in Section 4.1 states that there is a connection between the manifest and latent distributions in factor analysis. Hence the question is what implications one can expect if this "implicit assumption" may not be justified. Related to the study and evaluation of the underlying assumptions associated with these classical factor models, this paper, amongst others, shows that the data re-expression method PCA performs surprisingly well if compared to the results obtained based on the latent variable approaches EFA and PAA. Moreover, ML and PAA estimation procedures for EFA are compared with one another, for different degrees of violating the normality assumption at the manifest or latent levels.

primarily investigated the impact of categorical data (however, not varying the latent distributions for the factors). Reviewing results of previous simulation studies based on continuous indicator variables that have compared different estimation methods (including PCA) and have compared different methods for determining the number of factors, as to our knowledge, would have not constituted reviewing relevant literature focusing primarily on the violations of the assumptions associated with those models.

Literature on classical factor models has in particular investigated violations of the assumption of interval-scaled data. In classical factor analysis, Green (1983) simulated dichotomous data based on the 3PL (three parameter logistic) model (Birnbaum, 1968) and applied PCA and PAA to the data, whereat Cattell's scree test and Horn's parallel analysis were used as extraction criteria. Although both methods were applied to the same data, the results regarding the extracted factors obtained from the analyses differed, and the true dimensionality was not detected. In general, the models extracted too many factors. These findings are in line with expectations. Green (1983) used the phi-coefficient ϕ as the input data, and according to Ferguson (1941), the maximum value of ϕ depends on the difficulty parameters of the items. Dependence of ϕ on item difficulty can in extreme cases lead to factors being extracted solely due to the difficulties of the items. Roznowski et al. (1991) referred to such factors as difficulty factors.

Carroll (1945) recommended to use the tetrachoric correlation ρ_{tet} for factor analysis of dichotomous data. The coefficient ρ_{tet} is an estimate of the dependency between two dichotomous items based on the assumption that the items measure a latent continuous ability – an assumption that corresponds to the factor analysis approach. Although one would expect that ρ_{tet} leads to less biased results as compared to ϕ, Collins et al. (1986) were able to show that ϕ was much better suited to capture the true dimensionality than ρ_{tet}. In simulations, they compared the two correlation coefficients within the principal component analysis, using a version of the scree test as extraction criterion. The simulated data followed the 2PL model with three latent dimensions, and in addition to item discrimination (moderate, high, very high), the item difficulty and its distribution were varied (easy, moderate, difficult, and extreme difficult item parameters; distributed normal, low frequency, rectangular, and bimodal). The coefficient ρ_{tet} led to better results when the distribution of item difficulty was rectangular. In all other cases, ϕ was superior to ρ_{tet}. But with neither of the two methods it was possible to detect the true number of factors in more than 45% of the simulated data sets. See Roznowski et al. (1991) for another study illustrating the superiority of the coefficient ϕ to the coefficient ρ_{tet}.

Clarification for findings in Green (1983), Collins et al. (1986), and Roznowski et al. (1991) was provided by Weng and Cheng (2005). Weng and Cheng varied the number of items, the factor loadings and difficulties of the items, and sample size. The authors used the parallel analysis extraction method to determine the number of factors. However, the eigenvalues of the correlation matrices were computed using a different algorithm, which in a comparative study proved to be more reliable (Wang, 2001). With this algorithm, ϕ and ρ_{tet} performed equally well and misjudged true unidimensionality only when the factor loadings or sample sizes were small, or when the items were easy. This means that it was not the correlation coefficient *per se* that led to inadequate estimation of the number of factors but the extraction method that was used.

Muthén (1978, 1983, 1984), Muthén and Christoffersson (1981), Dolan (1994), Gorsuch (1997), Bolt (2005), Maydeu-Olivares (2005), and Wirth and Edwards (2007) present alternative or more sophisticated ways for dealing with categorical variables in factor analysis or structural equation modeling. Muthén (1989), Muthén and Kaplan (1992), and Ferguson and Cox (1993) compared the performances of factor analytic methods under conditions of (manifest) non-normality for the observed indicator variables.

We will add to and extend this literature and investigate in this paper whether the classical factor analysis models can reasonably unveil the factorial structure or properties of the population latent ability distribution in educational test data (e.g., obtained from criterion-referenced tests) when the assumption of normality in the latency may not be justified. None of the studies mentioned above has investigated the "true distribution impact" in these problems.

4. SIMULATION STUDY

A simulation study is used to evaluate the performances of the classical factor analytic approaches when the latent variables are not normally distributed.

True factorial structures under the exploratory factor analysis model are simulated, that is, the values of n, k, $\boldsymbol{L}$, $\boldsymbol{f}$, and $\boldsymbol{e}$ are varied.[5] On the basis of the constructed factorial structures, the matrices of the manifest variables are computed. These matrices are used as input data and analyzed with classical factor analytic methods. The estimates (or computed values) $\hat{k}$, $\hat{\boldsymbol{L}}$, and $\hat{\boldsymbol{f}}$ (or $\hat{\boldsymbol{F}}$) are then compared to the underlying population values. As criteria for "estimation accuracy" we use the number of extracted factors (as compared to true dimensionality), the skewness of the estimated latent ability distribution, and the discrepancy between estimated and true loading matrix. Shapiro-Wilk tests for normality of the ability estimates are presented and distributions of the estimated and true factor scores are compared as well.

Note that in the simulation study metric scale, not dichotomous, items are analyzed. This can be viewed as a baseline informative for the dichotomous indicator case as well (cf. Section 6). The results of the simulation study can serve as a reference also for situations where violations of normality for latent and manifest variables and metric scale data are present. One may expect the reported results to become worse when, in addition to (latent) non-normality of person ability, data are discretized or item statistics are skewed (manifest non-normality).

4.1. MOTIVATION AND PRELIMINARIES

The present simulation study particularly aims at analyzing and answering such questions as:

[5]Obviously, PCA as introduced in this paper cannot be used as a data generating probability model underlying the population. However, the simulation study shows that PCA results can approximate a factor analysis (cf. also Footnote 1).

- To what extent does the estimation accuracy for factorial structure of the classical factor analysis models depend on the skewness of the population latent ability distribution?
- Are there specific aspects of the factorial structure or latent ability distribution with respect to which the classical factor analysis models are more or less robust in estimation when true ability values are skewed?
- Given a skewed population ability distribution does the estimation accuracy for factorial structure of the classical factor analysis models depend on the extraction criterion applied for determining the number of factors from the data?
- Can person ability scores estimated under classical factor analytic approaches be representative of the true ability distribution or properties thereof when this distribution is skewed?

Mattson (1997)'s method can be used for specifying the parameter settings for the simulation study (cf. Section 4.2). We briefly describe this method (for details, see Mattson, 1997). Assume the standardized manifest variables are expressed as $\boldsymbol{z}=\boldsymbol{A}\boldsymbol{v}$, where $\boldsymbol{v}$ is the vector of latent variables and $\boldsymbol{A}$ is the matrix of model parameters. Moreover, assume that $\boldsymbol{v}=\boldsymbol{T}\boldsymbol{\omega}$, where $\boldsymbol{T}$ is a lower triangular square matrix such that each component of $\boldsymbol{v}$ is a linear combination of at most two components of $\boldsymbol{\omega}$, $E(\boldsymbol{v}\boldsymbol{v}')=\boldsymbol{\Sigma}_{\boldsymbol{v}}=\boldsymbol{T}\boldsymbol{T}'$, and $\boldsymbol{\omega}$ is a vector of mutually independent standardized random variables ω_i with finite central moments μ_{1i}, μ_{2i}, μ_{3i}, and μ_{4i}, of order up to four. Then

$$E(z)=\boldsymbol{A}\boldsymbol{T}E(\boldsymbol{\omega})=\boldsymbol{0}$$

and

$$E(\boldsymbol{z}\boldsymbol{z}')(=\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{A}')=\boldsymbol{A}\boldsymbol{T}E(\boldsymbol{\omega}\boldsymbol{\omega}')\boldsymbol{T}'\boldsymbol{A}'=\boldsymbol{A}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{A}'.$$

Or equivalently, $E(z_iz_j)=\boldsymbol{\gamma}_{\boldsymbol{i}}'\boldsymbol{\gamma}_{\boldsymbol{j}}$, where $\boldsymbol{\gamma}_{\boldsymbol{i}}=(\boldsymbol{a}_{\boldsymbol{i}}'\boldsymbol{T})'$ and $\boldsymbol{a}_{\boldsymbol{i}}'$ is the i-th row of $\boldsymbol{A}$. Under these conditions the third and fourth order central moments of z_i are given by

$$E(z_i^3)=\sum_m \gamma_{im}^3\mu_{3m} \qquad \text{and}$$

$$E(z_i^4)=\sum_m \gamma_{im}^4\mu_{4m}+6\sum_{m\geqslant 2}\sum_{o=1}^{m-1}\gamma_{im}^2\gamma_{io}^2.$$

Hence the univariate skewness $\sqrt{\beta_{1i}}$ and kurtosis β_{2i} of any z_i can be calculated by

$$\sqrt{\beta_{1i}}=\frac{E\left(z_i^3\right)}{\left[E\left(z_i^2\right)\right]^{3/2}} \qquad \text{and} \qquad \beta_{2i}=\frac{E\left(z_i^4\right)}{\left[E\left(z_i^2\right)\right]^2}.$$

In the simulation study, the exploratory factor analysis model with orthogonal factors ($\text{cov}(\boldsymbol{f},\boldsymbol{f})=\boldsymbol{I}$) and error variables assumed to be uncorrelated and unit normal (with standardized manifest variables) is used as the data generating model. Let $\boldsymbol{A}:=(\boldsymbol{L},\boldsymbol{I_p})$ be the concatenated matrix of dimension $p\times(k+p)$, where $\boldsymbol{I_p}$ is the unit matrix of order $p\times p$, and let $\boldsymbol{v}:=(\boldsymbol{f}',\boldsymbol{e}')'$ be the concatenated vector of length $k+p$. Then we have $\boldsymbol{z}=\boldsymbol{A}\boldsymbol{v}$ for the simulation factor model. Let $\boldsymbol{T}:=\boldsymbol{I_{(k+p)\times(k+p)}}$ and $\boldsymbol{\omega}:=\boldsymbol{v}$, then $\boldsymbol{T}$ and $\boldsymbol{\omega}$ satisfy the required assumptions afore mentioned. Hence the skewness and kurtosis of any z_i are given by, respectively,

$$\sqrt{\beta_{1i}}=\frac{\sum_{m=1}^{k+p}a_{im}^3\mu_{3m}}{[\boldsymbol{a}_{\boldsymbol{i}}'\boldsymbol{a}_{\boldsymbol{i}}]^{3/2}} \qquad \text{and}$$

$$\beta_{2i}=\frac{\sum_{m=1}^{k+p}a_{im}^4\mu_{4m}+6\sum_{m=2}^{k+p}\sum_{o=1}^{m-1}a_{im}^2a_{io}^2}{[\boldsymbol{a}_{\boldsymbol{i}}'\boldsymbol{a}_{\boldsymbol{i}}]^2}.$$

Mattson's method is used to specify such settings for the simulation study as they may be observed in large scale assessment data. The next section describes this in detail.

4.2. DESIGN OF THE SIMULATION STUDY

The number of manifest variables was fixed to $p=24$ throughout the simulation study. For the number of factors, we used numbers typically found in large scale assessment studies such as the Progress in International Reading Literacy Study (PIRLS, e.g., Mullis et al., 2006) or PISA (e.g., OECD, 2005). According to the assessment framework of PIRLS 2006 the number of dimensions for reading literacy was four, in PISA 2003 the scaling model had seven dimensions. We decided to use a simple loading structure for $\boldsymbol{L}$, in the sense that every manifest variable was assumed to load on only one factor (within-item unidimensionality) and that each factor was measured by the same number of manifest variables. In reliance on PIRLS and PISA in our simulation study, the numbers of factors were assumed to be four or eight. We assumed that some of the factors were well explained by their indicators while others were not, with upper rows (variables) of the loading matrix generally having higher factor loadings than lower rows (variables). Thus, the loading matrices employed in our study for the four and eight dimensional simulation models were, respectively,

$$\boldsymbol{L}=\begin{pmatrix}
0.9&0&0&0\\0.8&0&0&0\\0.7&0&0&0\\0.6&0&0&0\\0.5&0&0&0\\0.4&0&0&0\\
0&0.8&0&0\\0&0.7&0&0\\0&0.6&0&0\\0&0.5&0&0\\0&0.4&0&0\\0&0.3&0&0\\
0&0&0.6&0\\0&0&0.6&0\\0&0&0.5&0\\0&0&0.4&0\\0&0&0.4&0\\0&0&0.3&0\\
0&0&0&0.6\\0&0&0&0.5\\0&0&0&0.5\\0&0&0&0.4\\0&0&0&0.3\\0&0&0&0.3
\end{pmatrix} \quad \text{and} \quad \boldsymbol{L}=\begin{pmatrix}
0.9&0&0&0&0&0&0&0\\0.8&0&0&0&0&0&0&0\\0.7&0&0&0&0&0&0&0\\
0&0.8&0&0&0&0&0&0\\0&0.8&0&0&0&0&0&0\\0&0.7&0&0&0&0&0&0\\
0&0&0.8&0&0&0&0&0\\0&0&0.7&0&0&0&0&0\\0&0&0.6&0&0&0&0&0\\
0&0&0&0.7&0&0&0&0\\0&0&0&0.7&0&0&0&0\\0&0&0&0.7&0&0&0&0\\
0&0&0&0&0.7&0&0&0\\0&0&0&0&0.6&0&0&0\\0&0&0&0&0.6&0&0&0\\
0&0&0&0&0&0.6&0&0\\0&0&0&0&0&0.6&0&0\\0&0&0&0&0&0.5&0&0\\
0&0&0&0&0&0&0.5&0\\0&0&0&0&0&0&0.4&0\\0&0&0&0&0&0&0.4&0\\
0&0&0&0&0&0&0&0.4\\0&0&0&0&0&0&0&0.4\\0&0&0&0&0&0&0&0.3
\end{pmatrix}.$$

We decided to analyze released items of the PIRLS 2006 study (IEA, 2007) to have an empirical basis for the selection of skewness values for $\boldsymbol{\omega}(=\boldsymbol{\nu})$. We used a data set of dichotomously scored responses of 7,899 German students to 125 test items. **Figure 1** displays the distribution of the PIRLS items' (empirical) skewness values.[6]

We decided to simulate under three conditions for the distributions of $\boldsymbol{\omega}$. Under the first condition, ω_m $(m=1, \ldots, k)$ are normal with $\mu_{1m}=0$, $\mu_{2m}=1$, $\mu_{3m}=0$, and $\mu_{4m}=3$. Under the second condition, ω_m $(m=1, \ldots, k)$ are slightly skewed with $\mu_{1m}=0$, $\mu_{2m}=1$, $\mu_{3m}=-0.20$, and $\mu_{4m}=3$. Under the third condition, ω_m $(m=1, \ldots, k)$ are strongly skewed with $\mu_{1m}=0$, $\mu_{2m}=1$, $\mu_{3m}=-2$, and $\mu_{4m}=9$. The error terms were assumed to be unit normal, that is, we specified $\mu_{1h}=0$, $\mu_{2h}=1$, $\mu_{3h}=0$, and $\mu_{4m}=3$ for ω_h $(h=k+1, \ldots, k+p)$. Skewness and kurtosis of any z_i under each of the three conditions were computed using Mattson's method (Section 4.1). The values are reported in **Tables 1** and **2** for the four and eight dimensional factor spaces, respectively.

Under the slightly skewed distribution condition, the theoretical values of skewness for the manifest variables range between -0.060 and -0.005, a condition that captured approximately 20% of the considered PIRLS test items. Under the strongly skewed distribution condition, the theoretical values of skewness lie between -0.599 and -0.047, a condition that covered circa 30% of the PIRLS items (cf. **Figure 1**). Based on these theoretical skewness and kurtosis statistics, we can see to what extent under these model specifications the distributions of the manifest variables deviate from the normal distribution.

How to generate variates ω_i $(i=1, \ldots, k+p)$ such that they possess predetermined moments μ_{1i}, μ_{2i}, μ_{3i}, and μ_{4i}? To simulate values for ω_i with predetermined moments, we used the generalized lambda distribution (Ramberg et al., 1979)

$$\omega_i = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2},$$

where u is uniform (0, 1), λ_1 is a location parameter, λ_2 a scale parameter, and λ_3 and λ_4 are shape parameters. To realize the desired distribution conditions for the simulation study (normal, slightly skewed, strongly skewed) using this general distribution its parameters λ_1, λ_2, λ_3, and λ_4 had to be specified accordingly. Ramberg et al. (1979) tabulate the required values for the λ parameters for different values of μ. In particular, for a (more or less) normal distribution with $\mu_1=0$, $\mu_2=1$, $\mu_3=0$, and $\mu_4=3$ the corresponding values are $\lambda_1=0$, $\lambda_2=0.197$, $\lambda_3=0.135$, and $\lambda_4=0.135$. For a slightly skewed distribution with $\mu_1=0$, $\mu_2=1$, $\mu_3=-0.20$, and $\mu_4=3$, the values are $\lambda_1=0.237$, $\lambda_2=0.193$, $\lambda_3=0.167$, and $\lambda_4=0.107$. For a strongly skewed distribution with $\mu_1=0$, $\mu_2=1$, $\mu_3=-2$, and $\mu_4=9$, the parameter values are given by $\lambda_1=0.993, \lambda_2=-0.108\cdot10^{-2}, \lambda_3=-0.108\cdot10^{-2}$, and $\lambda_4=-0.041\cdot10^{-3}$.

Remark. Indeed, various distributions are possible (see Mattson, 1997); however, the generalized lambda distribution proves to be special. It performs very well in comparison to other distributions, when theoretical moments calculated according to the Mattson formulae are compared to their corresponding empirical moments computed from data simulated under a factor model (based on that distribution). For details, see Reinartz et al. (2002). These authors have also studied the effects of the use of different (pseudo) random number generators for realizing the uniform distribution in such a comparison study. Out of three compared random number generators – RANUNI from SAS, URAND from PRELIS, and RANDOM from Mathematica – the generator RANUNI performed relatively well or better. In this paper, we used the SAS program for our simulation study.[7]

Besides the number of factors and the distributions of the latent variables, sample size was varied. In the small sample case, every z_i consisted of $n=200$ observations, and in the large sample case z_i contained $n=600$ observations. **Table 3** summarizes the design of the simulation study. Overall there

[6]All figures of this paper were produced using the R statistical computing environment (R Development Core Team, 2011; www.r-project.org). The source files are freely available from the authors.

[7]For the factor analyses in this paper, we used the SAS program and its PROC FACTOR implementation of the methods PCA, EFA, and PAA. More precisely, variation of the PROC FACTOR statements, run in their default settings, yields the performed procedures PCA, EFA, and PAA (e.g., EFA if METHOD = ML).

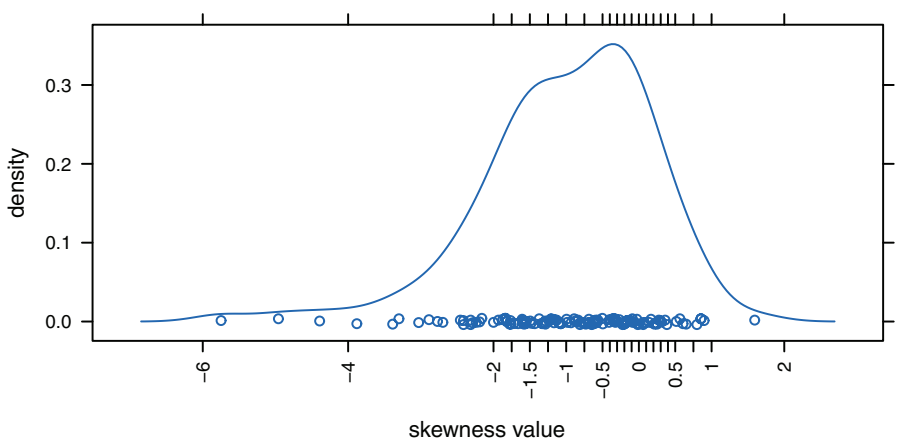


FIGURE 1 | Distribution of the skewness values for the 125 PIRLS test items.

Table 1 | Theoretical values of skewness and kurtosis for z_i (four factors).

z_i	Latent variable					
	Normal[a]		Slightly skewed[b]		Strongly skewed[c]	
	$\sqrt{\beta_{1i}}$	β_{2i}	$\sqrt{\beta_{1i}}$	β_{2i}	$\sqrt{\beta_{1i}}$	β_{2i}
z_1	0	3	−0.060	3	−0.599	4.202
z_2	0	3	−0.049	3	−0.488	3.914
z_3	0	3	−0.038	3	−0.377	3.649
z_4	0	3	−0.027	3	−0.272	3.420
z_5	0	3	−0.018	3	−0.179	3.240
z_6	0	3	−0.010	3	−0.102	3.114
z_7	0	3	−0.049	3	−0.488	3.914
z_8	0	3	−0.038	3	−0.377	3.649
z_9	0	3	−0.027	3	−0.272	3.420
z_{10}	0	3	−0.018	3	−0.179	3.240
z_{11}	0	3	−0.010	3	−0.102	3.114
z_{12}	0	3	−0.005	3	−0.047	3.041
z_{13}	0	3	−0.027	3	−0.272	3.420
z_{14}	0	3	−0.027	3	−0.272	3.420
z_{15}	0	3	−0.018	3	−0.179	3.240
z_{16}	0	3	−0.010	3	−0.102	3.114
z_{17}	0	3	−0.010	3	−0.102	3.114
z_{18}	0	3	−0.005	3	−0.047	3.041
z_{19}	0	3	−0.027	3	−0.272	3.420
z_{20}	0	3	−0.018	3	−0.179	3.240
z_{21}	0	3	−0.018	3	−0.179	3.240
z_{22}	0	3	−0.010	3	−0.102	3.114
z_{23}	0	3	−0.005	3	−0.047	3.041
z_{24}	0	3	−0.005	3	−0.047	3.041

[a] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = 0$, and $\mu_{4m} = 3$.
[b] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = -0.20$, and $\mu_{4m} = 3$.
[c] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = -2$, and $\mu_{4m} = 9$.

Table 2 | Theoretical values of skewness and kurtosis for z_i (eight factors).

z_i	Latent variable					
	Normal[a]		Slightly skewed[b]		Strongly skewed[c]	
	$\sqrt{\beta_{1i}}$	β_{2i}	$\sqrt{\beta_{1i}}$	β_{2i}	$\sqrt{\beta_{1i}}$	β_{2i}
z_1	0	3	−0.060	3	−0.599	4.202
z_2	0	3	−0.049	3	−0.488	3.914
z_3	0	3	−0.038	3	−0.377	3.649
z_4	0	3	−0.049	3	−0.488	3.914
z_5	0	3	−0.049	3	−0.488	3.914
z_6	0	3	−0.038	3	−0.377	3.649
z_7	0	3	−0.049	3	−0.488	3.914
z_8	0	3	−0.038	3	−0.377	3.649
z_9	0	3	−0.027	3	−0.272	3.420
z_{10}	0	3	−0.038	3	−0.377	3.649
z_{11}	0	3	−0.038	3	−0.377	3.649
z_{12}	0	3	−0.038	3	−0.377	3.649
z_{13}	0	3	−0.038	3	−0.377	3.649
z_{14}	0	3	−0.027	3	−0.272	3.420
z_{15}	0	3	−0.027	3	−0.272	3.420
z_{16}	0	3	−0.027	3	−0.272	3.420
z_{17}	0	3	−0.027	3	−0.272	3.420
z_{18}	0	3	−0.018	3	−0.179	3.240
z_{19}	0	3	−0.018	3	−0.179	3.240
z_{20}	0	3	−0.010	3	−0.102	3.114
z_{21}	0	3	−0.010	3	−0.102	3.114
z_{22}	0	3	−0.010	3	−0.102	3.114
z_{23}	0	3	−0.010	3	−0.102	3.114
z_{24}	0	3	−0.005	3	−0.047	3.041

[a] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = 0$, and $\mu_{4m} = 3$.
[b] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = -0.20$, and $\mu_{4m} = 3$.
[c] $\mu_{1m} = 0$, $\mu_{2m} = 1$, $\mu_{3m} = -2$, and $\mu_{4m} = 9$.

are 12 conditions and for every condition 100 data sets were simulated.

Each of the generated 1,200 data sets were analyzed using all of the models of principal component analysis, exploratory factor analysis (ML estimation), and principal axis analysis altogether with a varimax rotation (Kaiser, 1958).[8] For any data set under each model, the factors, and hence, the numbers of retained factors were determined by applying the following three extraction criteria or approaches: the Kaiser-Guttman criterion, the scree test, and the parallel analysis procedure.[9]

Table 3 | Summary of the simulation design and number of generated data sets.

Sample size	Number of factors	Latent variable distribution		
		Normal	Slightly skewed	Strongly skewed
200	4	100	100	100
	8	100	100	100
600	4	100	100	100
	8	100	100	100

4.3. EVALUATION CRITERIA

The criteria for evaluating the performance of the classical factor models are the number of extracted factors (as compared to true dimensionality), the skewness of the estimated latent ability

[8] Because of rotational indeterminacy in the factor analysis approach (e.g., Maraun, 1996), the results are as much an evaluation of varimax rotation as they are an evaluation of the manipulated variables in the study. For more information, see Section 7.

[9] The Kaiser-Guttman criterion is a poor way to determine the number of factors. However, due to the fact that none of the existing studies has investigated the estimation accuracy of this criterion when the latent ability distribution is skewed, we have decided to include the Kaiser-Guttman criterion in our study. This criterion may also be viewed as a "worst performing" baseline criterion, which other extraction methods need to outperform, as best as possible.

distribution, and the discrepancy between the estimated and the true loading matrix. The latter two criteria are computed using the true number of factors. Furthermore, Shapiro-Wilk tests for assessing normality of the ability estimates are presented and distributions of the estimated and true factor scores are compared.

For the skewness criterion, under a factor model and a simulation condition, for any data set the factor scores on a factor were computed and their empirical skewness was the value for this data set that was used and plotted. For the discrepancy criterion, under a factor model and a simulation condition, for any data set $i = 1, \ldots, 100$ a discrepancy measure D_i was calculated,

$$D_i = \frac{\sum_{x=1}^{p} \sum_{y=1}^{k} |\hat{l}_{i;xy} - l_{xy}|}{kp},$$

where $\hat{l}_{i;xy}$ and l_{xy} represent the entries of the estimated (varimax rotated, for data set $i = 1, \ldots, 100$) and true loading matrices, respectively. It gives the averaged sum of the absolute differences between the estimated and true factor loadings. We also report the average and variance (or standard deviation) of these discrepancy measures, over all simulated data sets,

$$\overline{D} = \frac{1}{100} \sum_{i=1}^{100} D_i \qquad \text{and} \qquad s^2 = \frac{1}{100 - 1} \sum_{i=1}^{100} (D_i - \overline{D})^2.$$

In addition to calculating estimated factor score skewness values, we also tested for univariate normality of the estimated factor scores. We used the Shapiro-Wilk test statistic W (Shapiro and Wilk, 1965). In comparison to other univariate normality tests, the Shapiro-Wilk test seems to have relatively high power (Seier, 2002). In our study, under a factor model and a simulation condition, for any data set the Shapiro-Wilk test statistic's p-value was calculated for the estimated factor scores on a factor and the distribution of the p-values obtained from 100 simulated data sets was plotted.

5. RESULTS

We present the results of our simulation study.

5.1. NUMBER OF EXTRACTED FACTORS

Figure 2 shows the relative frequencies of the numbers of extracted factors for sample size $n = 200$ and $k = 4$ as true number of factors. If the Kaiser-Guttman criterion is used, the number of extracted factors is overestimated (for PCA) or tends to be underestimated (for EFA and PAA). With the scree test, four dimensions where extracted in the majority of cases, but variation of the numbers of extracted factors over the different data sets is high. High variation in this case can be explained by the ambiguous and hence difficult to interpret eigenvalue graphics that one needs to visually inspect for the scree test. Applying the parallel analysis method, variation of the numbers of extracted factors can be reduced and the true number of factors is estimated very well (e.g., for PCA). There does not seem to be a relationship between the number of extracted factors and the underlying distribution (normal, slightly skewed, strongly skewed) of the latent ability values.

When sample size is increased to $n = 600$, variation of the estimated numbers of factors decreases substantially under many conditions (see **Figure 3**). Compared to small sample sizes, the scree test and the parallel analysis method perform very well. The Kaiser-Guttman criterion still leads to a biased estimation of the true number of factors. Once again, there seems to be no relationship between the distribution of the latent ability values and the number of extracted factors.

Figure 4, for a sample size of $n = 200$, shows the case when there are $k = 8$ factors underlying the data. The Kaiser-Guttman criterion again leads to overestimation or underestimation of the true number of factors. The extraction results for the scree test have very high variation, and estimation of the true number of factors is least biased when the parallel analysis method is used.

Increasing sample size from $n = 200$ to 600 results in a significant reduction of variation (**Figure 5**). However, the true number of factors can be estimated without bias only when the parallel analysis method is used as extraction criterion. A possible relationship between the distribution of the latent ability values and the number of extracted factors once again does not seem to be apparent.

To sum up, we suppose that the "number of factors extracted" is relatively robust against the extent the latent ability values may be skewed. Another observation is that the parallel analysis method seems to outperform the scree test and the Kaiser-Guttman criterion when it comes to detecting the number of underlying factors.

5.2. SKEWNESS OF THE ESTIMATED LATENT ABILITY DISTRIBUTION

Figure 6A shows the distributions of the estimated factor score skewness values, for $n = 200$, $k = 4$, and $\mu_{3m} = 0$. The majority of the skewness values lies in close vicinity of 0. In other words, for a true normal latent ability distribution with skewness $\mu_3 = 0$, under the classical factor models the estimated latent ability scores most likely seem to have skewness values of approximately 0. An impact of the factor model used for the analysis of the data on the skewness of the estimated latent ability values cannot be seen under this simulation condition. However, the standard deviations of the skewness values clearly decrease from the first to the fourth factor. In other words, the true skewness of the latent ability distribution may be more precisely estimated for the fourth factor than for the first.

When true latent ability values are slightly negative skewed, $\mu_3 = -0.20$, in our simulation study this skewness may only be properly estimated for the first and second extracted factors (**Figure 6B**). The estimated latent ability values of the third and fourth extracted factors more give skewness values of approximately 0. The true value of skewness for these factors hence may likely to be overestimated.

If true latent ability values are strongly negative skewed, $\mu_3 = -2$, unbiased estimation of true skewness may not be possible (**Figure 6C**). Even in the case of the first and second factors, the estimation is biased now. True skewness of the latent ability distribution may be overestimated regardless of the used factor model or factor position.

To sum up, under the classical factor models, the concept of "skewness of the estimated latent ability distribution" seems to be sensitive with respect to the extent the latent ability values may be skewed. It seems that, the more the true latent ability values are skewed, the greater is overestimation of true skewness. In other

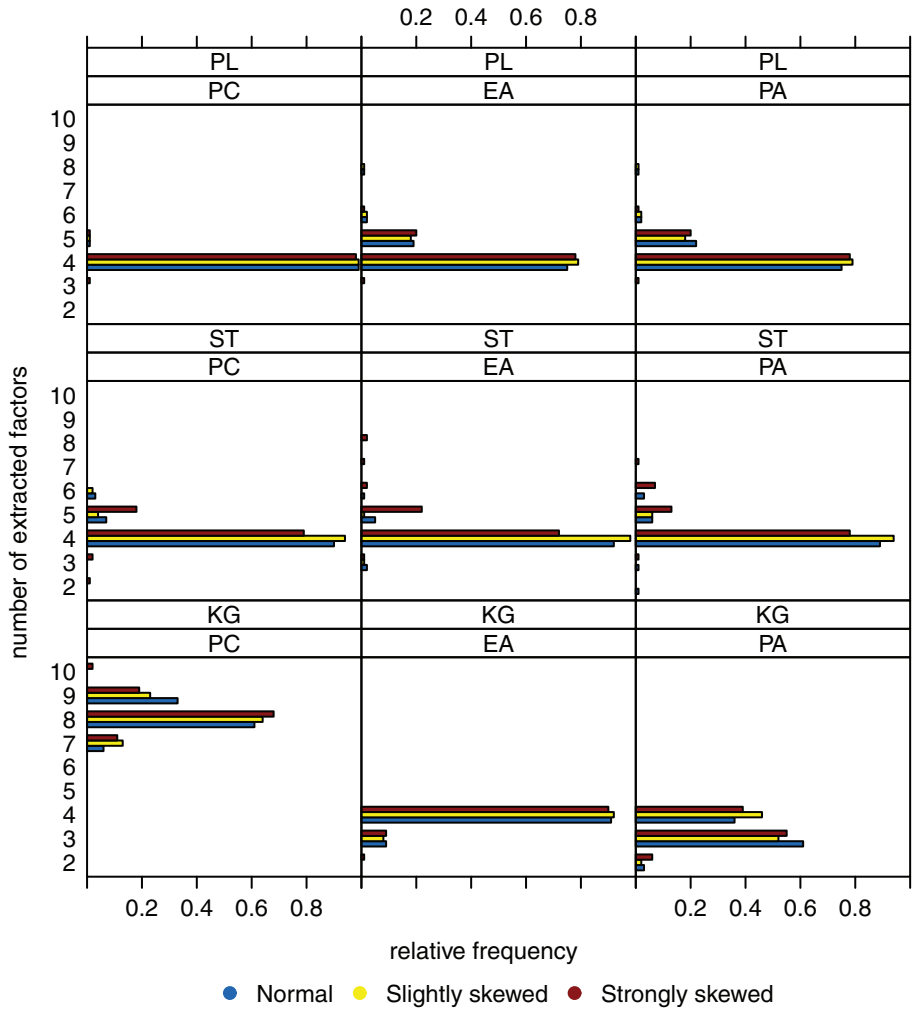


FIGURE 2 | Relative frequencies of the numbers of extracted factors, for $n = 200$ and $k = 4$. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Kaiser-Guttman criterion (KG), scree test (ST), and parallel analysis (PL) serve as factor extraction criteria.

words, strongly negative skewed distributions may not be estimated without bias based on the classical factor models. Increasing sample size, for example from $n = 200$ to 600, or changing the number of underlying factors, say from $k = 4$ to 8, did not alter this observation considerably. For that reason, the corresponding plots at this point of the paper are omitted and can be found in Kasper (2012).

We performed Shapiro-Wilk tests for univariate normality of the estimated factor scores. As can be seen from **Figure 7A**, under normally distributed true latent ability scores nearly all values of W are statistically non-significant. In these cases, the null hypothesis cannot be rejected.

A similar conclusion can be drawn when the true latent ability values are not normally distributed but instead follow a slightly skewed distribution (**Figure 7B**). Nearly all Shapiro-Wilk test statistic values are statistically non-significant. In other words, the null hypothesis stating normally distributed latent ability values is seldom rejected although the true latent distribution is skewed and not normal. No relationship between the p-values and the used factor model or factor position may be apparent (disregarding the observation that the p-values for the fourth factor are generally lower than for the other factors).

The case of a strongly skewed factor score distribution is depicted in **Figure 7C**. Virtually all values of W are statistically significant and the null hypothesis of normality of factor scores is rejected. Similar conclusions or observations may be drawn for increased sample size or factor space dimension and we do omit presenting plots thereof.

Finally, **Figure 8** shows the distribution of the estimated factor scores on the fourth factor (for $k = 4$) in comparison to the true strongly skewed ability distribution under the exploratory factor analysis model for a sample size of $n = 1{,}000$. The unit normal distribution is plotted as a reference. The estimated factor scores have a skewness value of -0.47 compared to true skewness -2. The estimated distribution deviates from the true distribution and does not approximate it acceptably well.

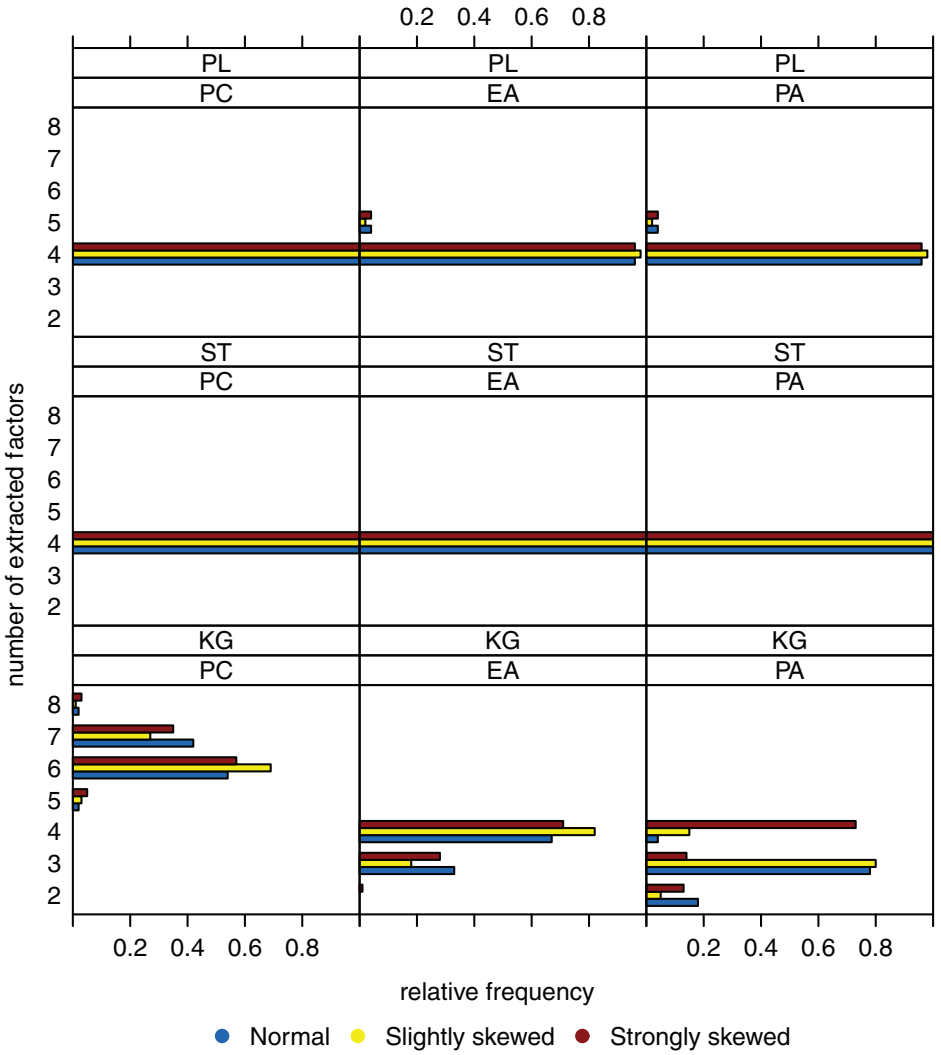


FIGURE 3 | Relative frequencies of the numbers of extracted factors, for $n = 600$ and $k = 4$. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Kaiser-Guttman criterion (KG), scree test (ST), and parallel analysis (PL) serve as factor extraction criteria.

5.3. DISCREPANCY BETWEEN THE ESTIMATED AND THE TRUE LOADING MATRIX

In **Table 4**, the average and standard deviation coefficients $\bar{D}$ and s for the discrepancies are reported. The largest average discrepancy values are obtained for the condition $n = 200$, $k = 8$, and the strongly skewed latent ability distribution: 0.173, 0.157, and 0.143 for PCA, EFA, and PAA, respectively. Under this condition, the true factor loadings are, mostly or clearly, overestimated or underestimated. Minor differences between the estimated and true factor loadings are obtained for $n = 600$, $k = 4$, and the normal latent ability distribution: with average discrepancies 0.076, 0.063, and 0.066 for PCA, EFA, and PAA, respectively.

Deviations of the estimated loading matrix from the true loading matrix can also be quantified and visualized at the level of individual absolute differences $|\hat{l}_{i;xy} - l_{xy}|$. In this way not only overall discrepancy averages can be studied but also the distribution of absolute differences at the individual entry level. **Figure 9** shows the distributions of the absolute differences $|\hat{l}_{i;xy} - l_{xy}|$ for the different sample sizes and numbers of underlying factors. In each panel, $100pk$ absolute differences are plotted.

The majority of the absolute differences lies in the range from 0 to circa 0.20. Larger absolute differences between the estimated and true factor loadings occurred rather rarely. It is also apparent that the 36 distributions hardly differ. This observation suggests that the effects or impacts of sample size, true number of factors, and the latent ability distribution on the accuracy of the classical factor models for estimating the factor loadings are rather weak. In that sense, estimation of the loading matrix seems to be robust overall. In our simulation study, we were not able to see a clear relationship between the distribution of the latent ability values and the discrepancy between the estimated and the true loading matrix.

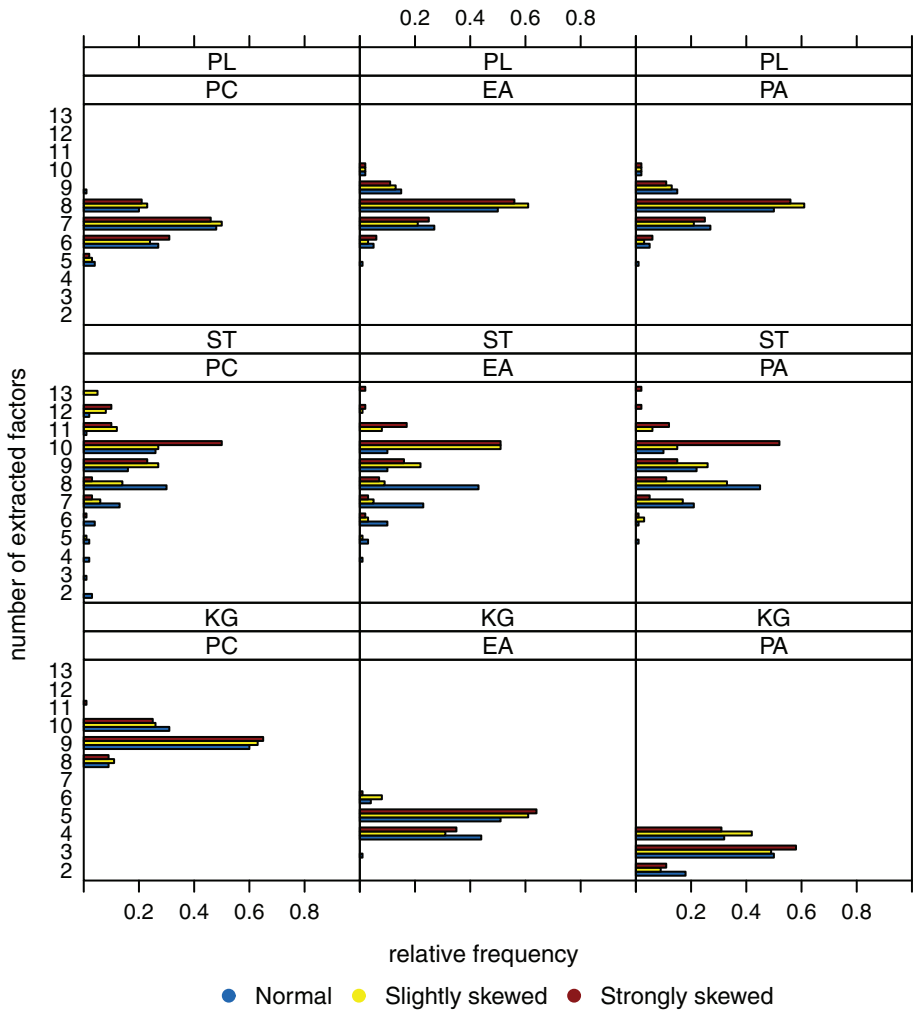


FIGURE 4 | Relative frequencies of the numbers of extracted factors, for $n = 200$ and $k = 8$. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Kaiser-Guttman criterion (KG), scree test (ST), and parallel analysis (PL) serve as factor extraction criteria.

6. ANALYSIS OF PIRLS 2006 DATA

In addition to the simulation study, the classical factor analytic approaches are also compared on the part of PIRLS 2006 data that we presented in Section 4.2. The booklet design in PIRLS implies that only a selection of the items has been administered to each student, depending on booklet approximately 23–26 test items per student (Mullis et al., 2006). As a consequence, the covariance or correlation matrices required for the factor models can only be computed for the items of a particular test booklet. Since analysis of all thirteen booklets of the PIRLS 2006 study is out of the scope of this paper, we decided to analyze booklet number 4. This booklet contains 23 items, and nine of these items (circa 40% of all items) have skewness values in the range of −0.6 to 0. This skewness range corresponds to the values considered in the simulation study, and no other test booklet had a comparably high percentage of items with skewness values in this range.

Note that in the empirical application dichotomized multicategory items are analyzed. In practice, large scale assessment data are discrete and not continuous. Yet, the metric scale indicator case considered in the simulation study can serve as an informative baseline; for instance (issue of polychoric approximation) to the extent that a product-moment correlation is a valid representation of bivariate relationships among interval-scaled variables (e.g., Flora et al., 2012). In our paper, the simulation results and the results obtained for the empirical large scale assessment application are, more or less, comparable.

In PIRLS 2006, four sorts of items were constructed and used for assigning "plausible values" to students (for details, see Martin et al., 2007). Any item loads on exactly one of the two dimensions "Literacy Experience" (L) and "Acquire and Use Information" (A) and also measures either the dimension "Retrieving and Straightforward Inferencing" (R) or the dimension "Interpreting, Integrating, and Evaluating" (I). Moreover, all of these items are assumed to be indicators for the postulated higher dimension "Overall Reading." In other words, PIRLS 2006 items may be assumed to be one-dimensional if the "uncorrelated" factor "Overall Reading" is

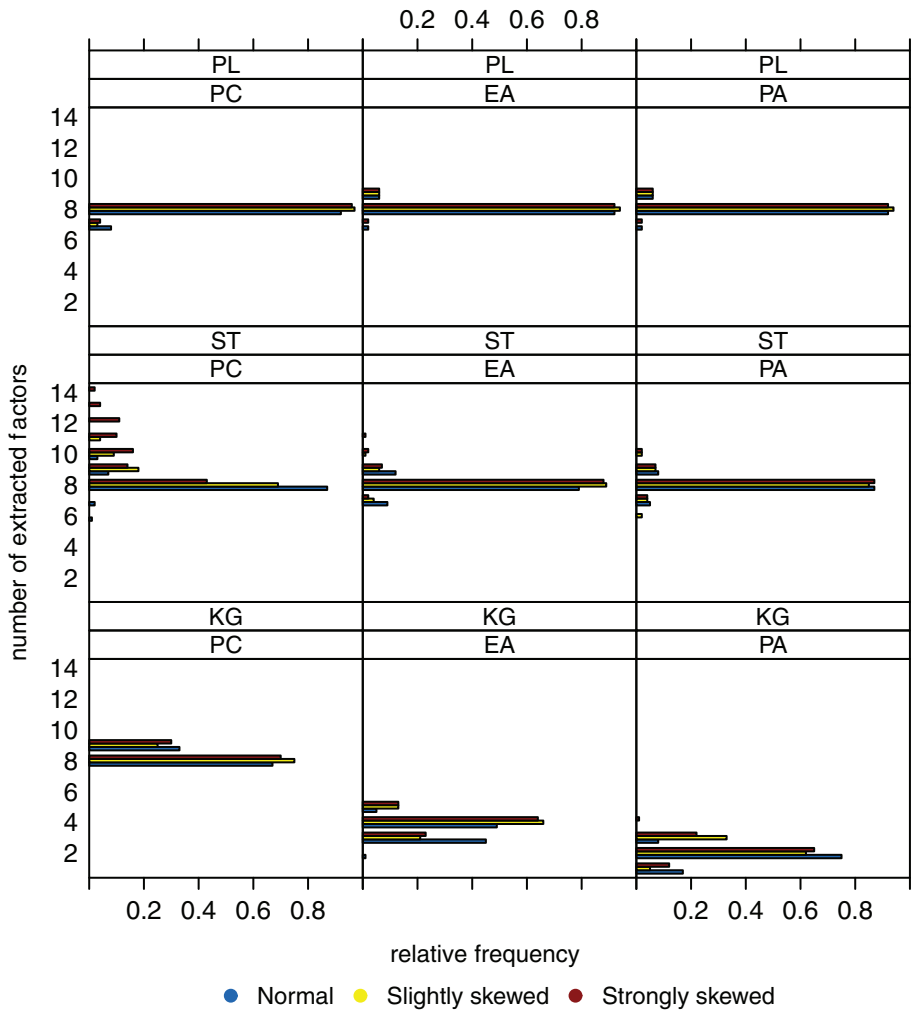


FIGURE 5 | Relative frequencies of the numbers of extracted factors, for $n = 600$ and $k = 8$. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Kaiser-Guttman criterion (KG), scree test (ST), and parallel analysis (PL) serve as factor extraction criteria.

considered ("orthogonal" case), or two-dimensional if any of the four combinations of correlated factors {A, L} × {I, R} is postulated (oblique case). In the latter case, "Overall Reading" may be assumed a higher order dimension common to the four factors. Booklet number 4 covers all these four sorts of PIRLS items.

A total of $n = 526$ students worked on booklet number 4. We investigated these data using principal component analysis, exploratory factor analysis, and principal axis analysis. For determining the number of underlying dimensions, the Kaiser-Guttman criterion, the scree test, and the method of parallel analysis were used. The results of the analyses can be found in **Table 5**.

The situation at this point is comparable to what we have reported in simulation in **Figure 3**. The scree test unveils unidimensionality of the test data independent of factor model. The numbers of factors extracted by the parallel analysis method depend on the factor model that was used. For PCA, again as for the scree test, unidimensionality is detected, however for the error component models EFA and PAA, four dimensions are uncovered (see also below). It seems that these "inferential" or "distributional" factor models, to some degree, are sensitive to dependencies among factors. According to the Kaiser-Guttman criterion, which performs worst, there are six dimensions underlying the data for any of the three factor models.

The varimax rotated loading matrices for the exploratory factor analysis and principal axis analysis models with four factors are reported in **Tables 6** and **7**. Once again, the situation is comparable to what we have obtained in simulation in **Table 4** or **Figure 9**. The estimated loading matrices under EFA and PAA are very similar. Highlighted factor loadings $\hat{l}_{xy} \geqslant 0.30$, for instance, are identically located in the matrices. As can be seen from **Tables 6** and **7**, substantially different items in regard to their PIRLS contents load on the same factors, and moreover, there are items of same PIRLS contents that show substantial loadings on different factors. We suppose that this may be a consequence of the factors, in this example, most likely being correlated with a postulated common single dimension underlying the factors.

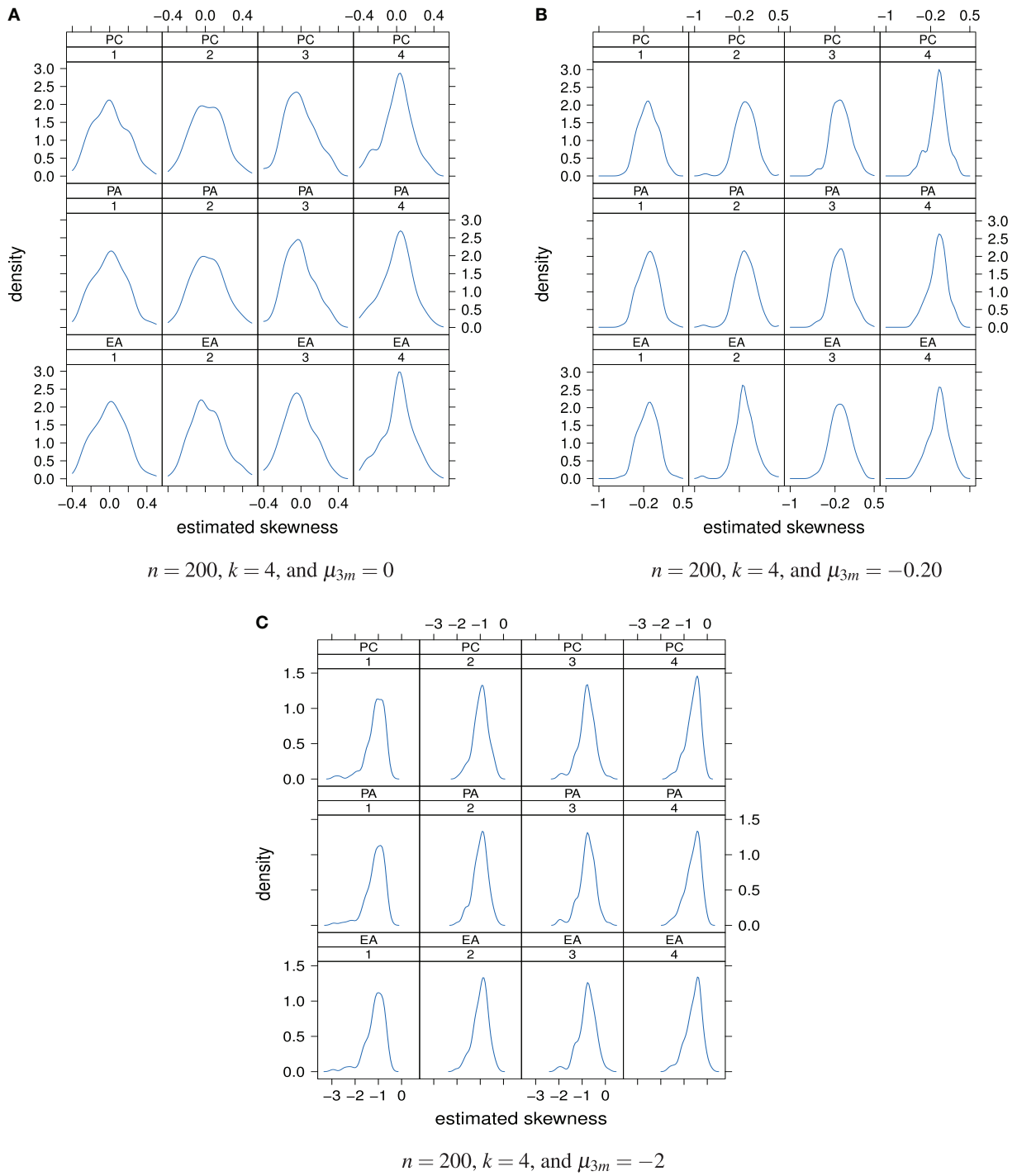


FIGURE 6 | Distributions of the estimated factor score skewness values as a "function" of factor model and factor position. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Numbers 1, 2, 3, and 4 stand for 1st, 2nd, 3rd, and 4th factors, respectively. The normal, slightly skewed, and strongly skewed distribution conditions are depicted in the panels **A**, **B**, and **C**, respectively.

7. CONCLUSION

7.1. SUMMARY

Assessing construct validity of a test in the sense of its factorial structure is important. For example, we have addressed possible implications for the analysis of criterion-referenced tests or for such large scale assessment studies as the PISA or PIRLS. There are a number of latent variable models that may be used to analyze the factorial structure of a test. This paper has focused

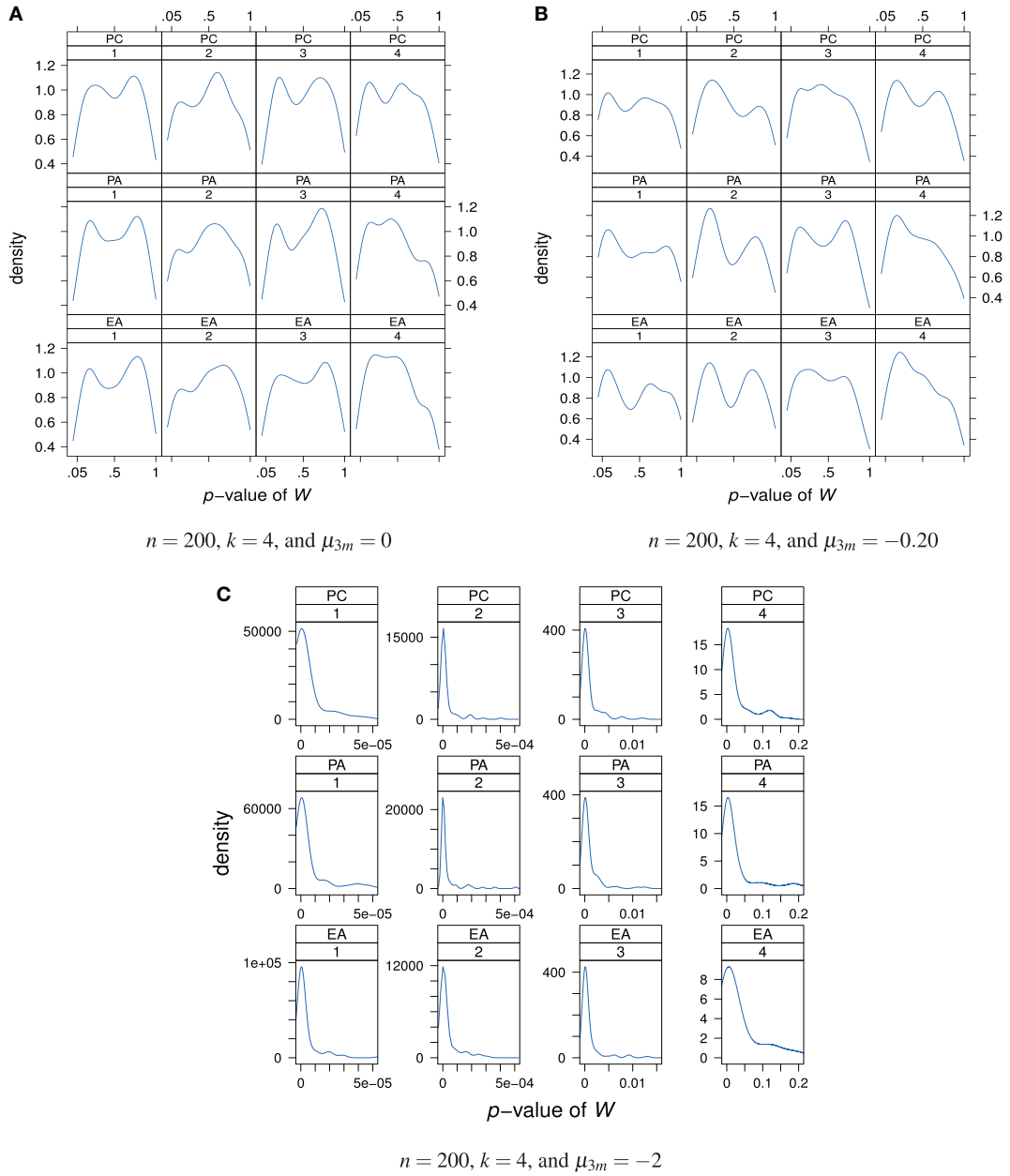


FIGURE 7 | Distributions of the *p*-values of the Shapiro-Wilk test statistic *W* as a "function" of factor model and factor position. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Numbers 1, 2, 3, and 4 stand for 1st, 2nd, 3rd, and 4th factors, respectively. The normal, slightly skewed, and strongly skewed distribution conditions are depicted in the panels **A**, **B**, and **C**, respectively.

on the following classical factor analytic approaches: principal component analysis, exploratory factor analysis, and principal axis analysis. We have investigated how accurately the factorial structure of test data can be estimated with these approaches, when assumptions associated with the procedures are not satisfied. We have examined the scope of those methods for estimating properties of the population latent ability distribution, especially when that distribution is slightly or strongly skewed (and not normal).

The estimation accuracy of the classical factor analytic approaches has been investigated in a simulation study. The study has in particular shown that the estimation of the true number of factors and of the underlying factor loadings seems to be relatively robust against a skewed population ability or factor score distribution (see Sections 5.1 and 5.3, respectively). Skewness and distribution of the estimated factor scores, on the other hand, have been seen to be sensitive concerning the properties of the true ability distribution (see Section 5.2). Therefore, the classical

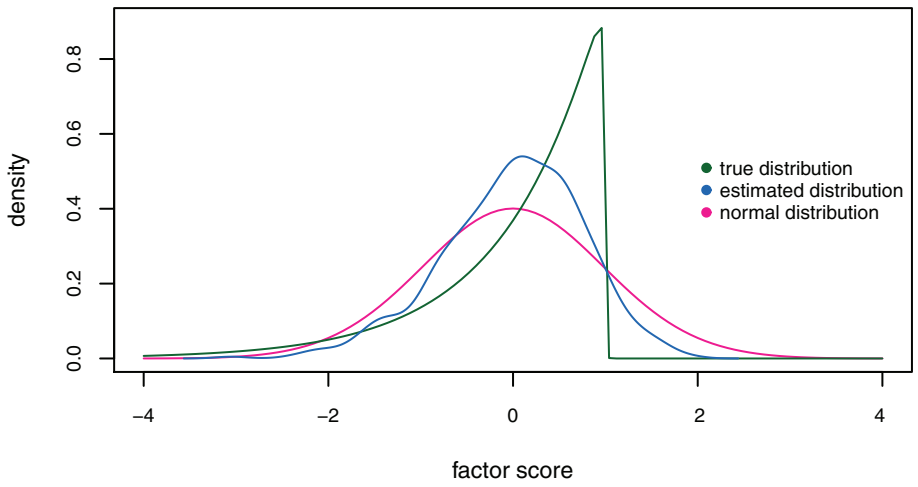


FIGURE 8 | Distributions of the estimated (blue curve) and true (green curve) factor scores on the fourth factor under the exploratory factor analysis model for sample size $n = 1{,}000$, factor space dimension $k = 4$, and true skewness $\mu_3 = -2$. The unit normal distribution is plotted as a reference (red curve).

Table 4 | Discrepancy averages and standard deviations $\bar{D}$ and *s*, respectively.

n	*k*	Model	Latent variable distribution					
			Normal		Slightly skewed		Strongly skewed	
			$\bar{D}$	*s*	$\bar{D}$	*s*	$\bar{D}$	*s*
200	4	PCA[a]	0.143	0.156	0.143	0.156	0.158	0.173
		EFA[b]	0.129	0.142	0.124	0.136	0.141	0.154
		PAA[c]	0.128	0.139	0.124	0.136	0.137	0.150
600	4	PCA	0.076	0.087	0.075	0.086	0.091	0.106
		EFA	0.063	0.072	0.062	0.072	0.080	0.095
		PAA	0.066	0.075	0.064	0.074	0.082	0.096
200	8	PCA	0.165	0.169	0.162	0.166	0.172	0.176
		EFA	0.154	0.157	0.152	0.155	0.156	0.159
		PAA	0.135	0.138	0.134	0.138	0.143	0.146
600	8	PCA	0.119	0.123	0.118	0.123	0.125	0.130
		EFA	0.106	0.112	0.107	0.112	0.115	0.120
		PAA	0.097	0.101	0.095	0.099	0.102	0.105

[a] *PCA, principal component analysis;* [b] *EFA, exploratory factor analysis;* [c] *PAA, principal axis analysis.*

factor analytic procedures, even though they are performed with metric scale indicator variables, seem not to be appropriate for estimating properties of ability in the "non-normal case." Significance of this result on sensitivity of factor score estimation to the nature of the latent distribution has been discussed for the PISA study, which is an international survey with impact on education policy making and the education system in Germany (see Sections 1 and 3.1). In addition to that discussion, the classical factor analytic approaches have been examined in more detail on PIRLS large scale assessment data, corroborating the results that we have obtained from the simulation study (see Section 6).

A primary aim of our work is to develop some basic understanding for how and to what extent the results of classical factor analyses (in the present paper, PCA, EFA, and PAA) may be affected by a non-normal latent factor score distribution. This has to be distinguished from non-normality in the manifest variables, which has been largely studied in the literature on the factor analysis of items (cf. Section 3.2). In this respect, regarding the investigation of non-normal factors, the present paper is novel. However, this is important, since it is not difficult to conceive of the possibility that latent variables may be skewed. Interestingly, moreover we have seen that a purely computational dimensionality reduction method can perform surprisingly well, as compared to the

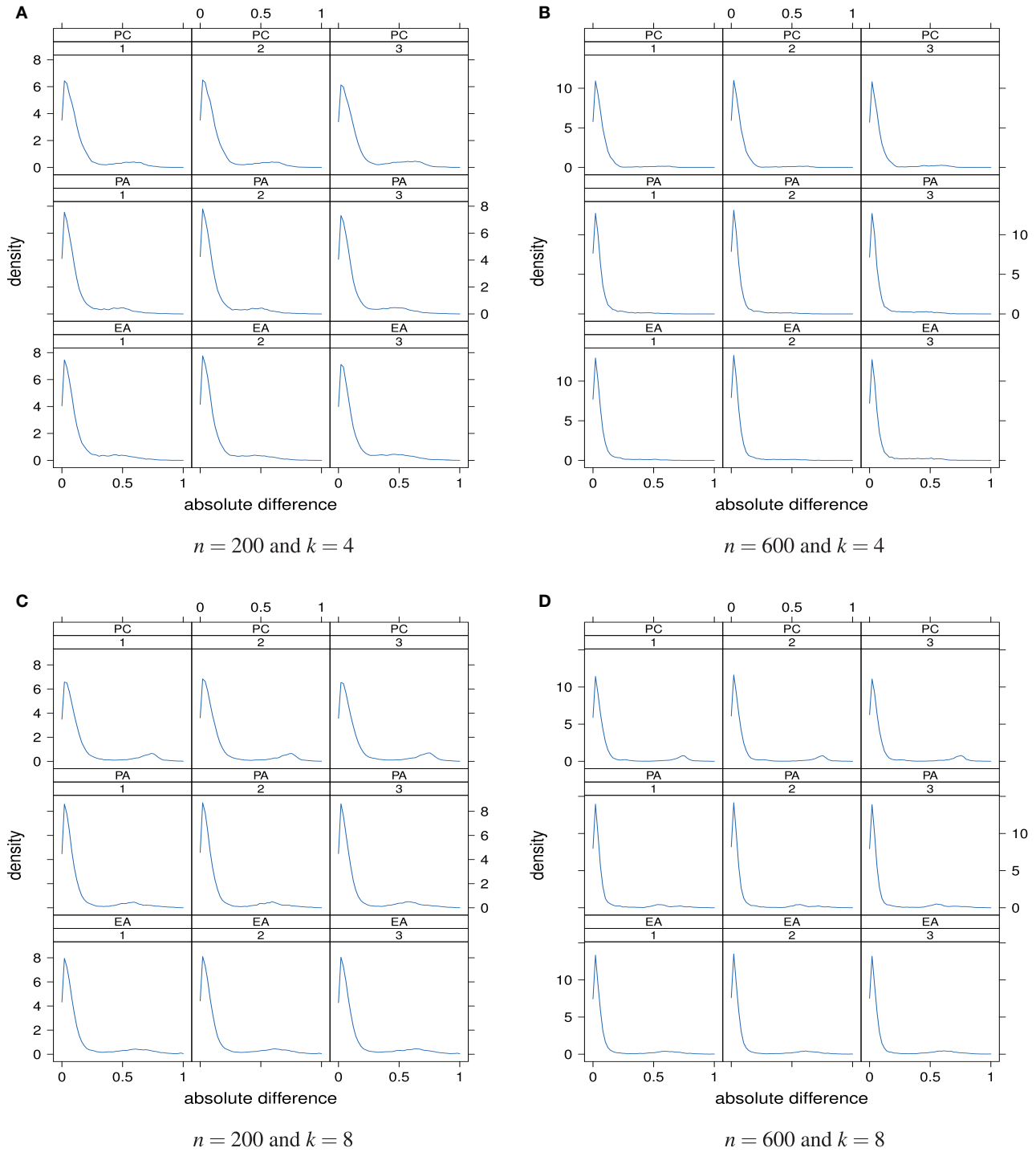


FIGURE 9 | Distributions of the absolute differences $\left|\hat{l}_{i;xy} - l_{xy}\right|$ as a "function" of factor model and skewness of the latent ability distribution. Factor models are principal component analysis (PCA, or PC), exploratory factor analysis (EFA, or EA), and principal axis analysis (PAA, or PA). Numbers 1, 2, and 3 stand for normal, slightly skewed, and strongly skewed population latent ability values, respectively. The panels are for the different sample sizes and numbers of underlying factors.

results obtained based on latent variable models. This observation may possibly be coined a general research program: whether genuine statistical approaches (originally based on variables without a measurement error) can work well, perhaps under specific restrictions to be explored, when latent variables are basically postulated, seemingly more closely matching the purpose of analysis.

Table 5 | Number of extracted dimensions for the PIRLS 2006 test booklet number 4, German sample.

Extraction method	Factor model		
	PCA[a]	EFA[b]	PAA[c]
Kaiser–Guttman criterion	6	6	6
Scree test	1	1	1
Parallel analysis method	1	4	4

[a]*PCA, principal component analysis;* [b]*EFA, exploratory factor analysis;* [c]*PAA, principal axis analysis.*

Table 6 | Loading matrix for four factors exploratory factor analysis of the PIRLS 2006 data for test booklet number 4, German sample.

Item	Factor			
	1	2	3	4
R011A01C^{A,R}	0.15	0.26	**0.39**	−0.05
R011A02M^{A,R}	0.14	0.28	**0.34**	0.19
R011A03C^{A,R}	0.16	0.24	0.09	0.03
R011A04C^{A,I}	**0.39**	0.19	0.10	0.06
R011A05M^{A,R}	0.22	0.08	0.19	0.21
R011A06M^{A,R}	0.20	0.03	0.14	0.06
R011A07C^{A,R}	**0.50**	0.20	0.22	0.15
R011A08C^{A,R}	**0.35**	0.04	**0.38**	−0.09
R011A09C^{A,I}	**0.55**	0.18	0.11	0.00
R011A10M^{A,I}	0.28	0.27	0.22	0.11
R011A11C^{A,I}	**0.37**	0.06	0.14	0.02
R021E01M^{L,R}	0.08	**0.45**	0.19	−0.06
R021E02M^{L,R}	0.02	**0.49**	0.09	0.24
R021E03M^{L,R}	0.14	**0.34**	−0.02	0.02
R021E04M^{L,R}	0.17	0.28	0.15	0.02
R021E05C^{L,R}	0.22	0.23	**0.32**	0.12
R021E06M^{L,R}	0.17	**0.44**	0.09	0.28
R021E07C^{L,I}	0.13	0.06	**0.48**	0.22
R021E08M^{L,I}	**0.32**	0.23	0.04	**0.48**
R021E09C^{L,I}	**0.45**	0.24	0.02	0.20
R021E10C^{L,I}	0.27	0.23	0.17	0.07
R021E11M^{L,I}	0.00	0.01	0.06	**0.40**
R021E12C^{L,I}	**0.38**	0.17	**0.31**	0.22

Factor loadings greater or equal 0.30 are highlighted.

A, Acquire and Use Information; L, Literary Experience; R, Retrieving and Straightforward Inferencing; I, Interpreting, Integrating, and Evaluating.

Table 7 | Loading matrix for four factors principal axis analysis of the PIRLS 2006 data for test booklet number 4, German sample.

Item	Factor			
	1	2	3	4
R011A01C^{A,R}	0.15	0.26	**0.40**	−0.06
R011A02M^{A,R}	0.14	0.29	**0.33**	0.18
R011A03C^{A,R}	0.16	0.24	0.09	0.02
R011A04C^{A,I}	**0.38**	0.20	0.10	0.06
R011A05M^{A,R}	0.22	0.07	0.19	0.24
R011A06M^{A,R}	0.19	0.02	0.14	0.07
R011A07C^{A,R}	**0.50**	0.20	0.22	0.16
R011A08C^{A,R}	**0.36**	0.03	**0.38**	−0.08
R011A09C^{A,I}	**0.54**	0.19	0.12	0.00
R011A10M^{A,I}	0.28	0.27	0.22	0.11
R011A11C^{A,I}	**0.38**	0.07	0.13	0.02
R021E01M^{L,R}	0.07	**0.45**	0.19	−0.06
R021E02M^{L,R}	0.03	**0.49**	0.09	0.24
R021E03M^{L,R}	0.14	**0.33**	−0.02	0.02
R021E04M^{L,R}	0.17	0.26	0.16	0.04
R021E05C^{L,R}	0.21	0.23	**0.32**	0.12
R021E06M^{L,R}	0.17	**0.44**	0.08	0.27
R021E07C^{L,I}	0.13	0.06	**0.47**	0.23
R021E08M^{L,I}	**0.32**	0.24	0.05	**0.46**
R021E09C^{L,I}	**0.45**	0.24	0.02	0.19
R021E10C^{L,I}	0.27	0.24	0.17	0.06
R021E11M^{L,I}	0.00	0.02	0.05	**0.40**
R021E12C^{L,I}	**0.38**	0.17	**0.30**	0.22

Factor loadings greater or equal 0.30 are highlighted.

A, Acquire and Use Information; L, Literary Experience; R, Retrieving and Straightforward Inferencing; I: Interpreting, Integrating, and Evaluating.

7.2. OUTLOOK

We have discussed possible implications of the findings for criterion-referenced tests and large scale educational assessment. The assumptions of the classical factor models have been seen to be crucial in these application fields. We suggest, for instance, that the presented classical procedures should not be used, unless with special caution if at all, to examine the factorial structure of dichotomously scored criterion-referenced tests. Instead, if model violations of the "sensitive" type are present, better suited or more sophisticated latent variable models can be used (see Skrondal and Rabe-Hesketh, 2004). Examples are item response theory parametric or non-parametric models for categorical response data (e.g., van der Linden and Hambleton, 1997). Furthermore, we would like to mention item response based factor analysis approaches by Bock and Lieberman (1970) or Christoffersson (1975, 1977). We may also pay attention to tetrachoric or polychoric based structural equation models by Muthén (1978, 1983, 1984) and Muthén and Christoffersson (1981).

As with factor analysis a general problem (e.g., Maraun, 1996), we had to deal with the issue of rotational indeterminacy and of selecting a specific rotation. We have decided to use varimax rotation, due to the fact that this rotation is most frequently used in empirical educational studies (for better interpretability of the factors). Future research may cover other rotations (e.g., quartimax or equimax) or the evaluation of parameter estimation by examining the communality estimates for each item (which are not dependent on rotation, but are a function of the factor loadings). Moreover, the orthogonal factor model may not be realistic, as factors are correlated in general. However, in the current study, it may be unlikely that having non-zero population factor loadings for correlated dimensions would substantially affect the findings. In further research, we will have to study the case of the oblique (non-orthogonal) factor model.

The results of this paper provide implications for popular research practices in the empirical educational research field. The methods that we have utilized are traditional and often applied in practice (e.g., by educational scientists), for instance to determine the factorial validity of criterion-referenced tests or to study large scale assessment measurement instruments. In addition, to consider other, more sophisticated fit statistics can be interesting and valuable. For example, such model fit statistics as the root mean square residual, comparative fit index, or the root mean squared error of approximation may be investigated. Albeit these fit statistics are well-known and applied in the confirmatory factor analysis (CFA) context, they could be produced for exploratory factor analysis (given that CFA and EFA are based on the same common factor model).

We conclude with important research questions related to the PISA study. In the context of PISA, principal component analysis is used, in the purely computational sense. Other distributional, inferential, or confirmatory factor models, especially those for the verification of the factorial validity of the PISA context questionnaires, have not been considered. Interesting questions arise: are there other approaches to dimensionality reduction that can perform at least as well as the principal component analysis method in PISA data (e.g., multidimensional scaling; Borg and Groenen, 2005)? Is the 95% extraction rule in principal component analysis of PISA data an "optimal" criterion? How sensitive are PISA results if, for example, the parallel analysis method is used as the extraction criterion? Answering these and other related questions is out of the scope of the present paper and can be pursued in more in-depth future analyses. Nonetheless, the important role of these problems in the PISA context is worth mentioning. The PISA procedure uses not only manifest background information but also principal component scores on complex constructs in order to assign literacy or plausible values to students. Future research is necessary to investigate the effects and possible implications of potentially biased estimates of latent or complex background information on students' assigned literacy values, and especially, their competence levels, based on which the PISA rankings are reported.

ACKNOWLEDGMENTS

The authors wish to thank Sabine Felbinger for her critical reading and helpful comments. In particular, we are deeply indebted to Jason W. Osborne, Chief Editor, and four reviewers. Their critical and valuable comments and suggestions have improved the manuscript greatly.

REFERENCES

Adams, R., Wilson, M., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Appl. Psychol. Meas.* 21, 1–23.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach.* Chichester: John Wiley & Sons.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley), 397–479.

Bock, D., and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika* 35, 179–197.

Bolt, D. M. (2005). "Limited- and full-information estimation of item response theory models," in *Contemporary Psychometrics*, eds A. Maydeu-Olivares and J. J. McArdle (Mahwah, NJ: Lawrence Erlbaum Associates), 27–72.

Borg, I., and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications.* Berlin: Springer.

Browne, M. (1974). Generalized least squares estimators in the analysis of covariance structures. *South Afr. Stat. J.* 8, 1–24.

Burt, C. (1909). Experimental tests of general intelligence. *Br. J. Psychol.* 3, 94–177.

Carroll, J. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika* 10, 1–19.

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behav. Res.* 1, 245–276.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika* 40, 5–32.

Christoffersson, A. (1977). Two-step weighted least squares factor analysis of dichotomized variables. *Psychometrika* 42, 433–438.

Collins, L., Cliff, N., McCormick, D., and Zatkin, J. (1986). Factor recovery in binary data sets: a simulation. *Multivariate Behav. Res.* 21, 377–391.

Cronbach, L., and Meehl, P. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302.

Cudeck, R., and MacCallum, R. (eds). (2007). *Factor Analysis at 100: Historical Developments and Future Directions.* Mahwah, NJ: Lawrence Erlbaum Associates.

Dolan, C. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: a comparison of categorical variable estimators using simulated data. *Br. J. Math. Stat. Psychol.* 47, 309–326.

Ferguson, E., and Cox, T. (1993). Exploratory factor analysis: a users' guide. *Int. J. Sel. Assess.* 1, 84–94.

Ferguson, G. (1941). The factorial interpretation of test difficulty. *Psychometrika* 6, 323–329.

Flora, D., LaBrish, C., and Palmers, R. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Front. Psychol.* 3:55. doi:10.3389/fpsyg.2012.00055

Gorsuch, R. (1997). Exploratory factor analysis: its role in item analysis. *J. Pers. Assess.* 68, 532–560.

Green, S. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Appl. Psychol. Meas.* 7, 139–147.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika* 19, 149–161.

Harman, H. (1976). *Modern Factor Analysis.* Chicago, IL: University of Chicago Press.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185.

Hotelling, H. (1933a). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417–441.

Hotelling, H. (1933b). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 498–520.

IEA. (2007). *PIRLS 2006 Assessment.* Boston, MA: TIMSS & PIRLS International Study Center.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika* 31, 165–178.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* 32, 443–482.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 187–200.

Kaiser, H., and Dickman, K. (1959). Analytic determination of common factors. *Am. Psychol.* 14, 425.

Kasper, D. (2012). *Klassische und Dichotome Faktorenanalyse auf dem Prüfstand: Zum Einfluss der Fähigkeits- und Aufgabenparameter auf die Schätzgenauigkeit verschiedener Faktorenmodelle [Classical and Dichotomous Factor Analysis put to Test: On the Impact of the Ability and Item Parameters on the Estimation Precision of Various Factor Models].* Münster: Waxmann.

Kelley, T. L. (1935). *Essential Traits of Mental Life.* Cambridge, MA: Harvard University Press.

Klauer, K. (1987). *Kriteriumsorientierte Tests: Lehrbuch der Theorie und Praxis lehrzielorientierten Messens [Criterion-Referenced Tests: Textbook on Theory and Practice of Teaching Goal Oriented Measuring].* Göttingen: Hogrefe.

Lienert, G., and Raatz, U. (1998). *Testaufbau und Testanalyse [Test Construction and Test Analysis].* Weinheim: Psychologie Verlags Union.

MacCallum, R. (2009). "Factor analysis," in *The SAGE Handbook of Quantitative Methods in Psychology*, eds R. E. Millsap and A. Maydeu-Olivares (London: Sage Publications), 123–147.

Maraun, M. D. (1996). Metaphor taken as math: indeterminacy in the factor analysis model. *Multivariate Behav. Res.* 31, 517–538.

Martin, M., Mullis, I., and Kennedy, A. (2007). *PIRLS 2006 Technical Report.* Boston, MA: TIMSS & PIRLS International Study Center.

Mattson, S. (1997). How to generate non-normal data for simulation of structural equation models. *Multivariate Behav. Res.* 32, 355–373.

Maydeu-Olivares, A. (2005). "Linear item response theory, nonlinear item response theory and factor analysis: a unified framework," in *Contemporary Psychometrics*, eds A. Maydeu-Olivares and J. J. McArdle (Mahwah, NJ: Lawrence Erlbaum Associates), 73–102.

McDonald, R. (1985). *Factor Analysis and Related Methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mulaik, S. (2009). *Foundations of Factor Analysis.* Boca Raton, FL: Chapman & Hall/CRC.

Mullis, I. V., Kennedy, A. M., Martin, M. O., and Sainsbury, M. (2006). *PIRLS 2006 Assessment Framework and Specifications.* Boston, MA: TIMSS & PIRLS International Study Center.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika* 43, 551–560.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *J. Econom.* 22, 43–65.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 49, 115–132.

Muthén, B. (1989). Dichotomous factor analysis of symptom data. *Sociol. Methods Res.* 18, 19–65.

Muthén, B., and Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika* 46, 407–419.

Muthén, B., and Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: a note on the size of the model. *Br. J. Math. Stat. Psychol.* 45, 19–30.

OECD. (2005). *PISA 2003 Technical Report.* Paris: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report.* Paris: OECD Publishing.

Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philos. Mag.* 2, 559–572.

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., and Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics* 21, 201–214.

Reinartz, W. J., Echambadi, R., and Chin, W. W. (2002). Generating non-normal data for simulation of structural equation models using Mattson's method. *Multivariate Behav. Res.* 37, 227–244.

Roznowski, M., Tucker, L., and Humphreys, L. (1991). Three approaches to determining the dimensionality of binary items. *Appl. Psychol. Meas.* 15, 109–127.

Seier, E. (2002). Comparison of tests for univariate normality. Retrieved from: http://interstat.statjournals.net/YEAR/2002/articles/0201001.pdf [March 4, 2013].

Shapiro, S., and Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: Chapman & Hall/CRC.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *Am. J. Psychol.* 15, 201–292.

Sturzbecher, D., Kasper, D., Bönninger, J., and Rüdel, M. (2008). *Evaluation der theoretischen Fahrerlaubnisprüfung: Methodische Konzeption und Ergebnisse des Revisionsprojekts [Evaluation of the Theoretical Driving License Test: Methodological Concepts and Results of the Revision Project].* Dresden: TÜV|DEKRA arge tp 21.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychol. Rev.* 38, 406–427.

Thurstone, L. L. (1965). *Multiple Factor Analysis.* Chicago, IL: University of Chicago Press.

van der Linden, W., and Hambleton, R. (eds). (1997). *Handbook of Modern Item Response Theory.* New York: Springer.

Velicer, W., and Jackson, D. (1990). Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. *Multivariate Behav. Res.* 25, 1–28.

Wang, C. (2001). *Effect of Number of Response Categories and Scale Distribution on Factor Analysis.* Master's thesis, National Taiwan University, Taipei.

Weng, L., and Cheng, C. (2005). Parallel analysis with unidimensional binary data. *Educ. Psychol. Meas.* 65, 697–716.

Widaman, K. F. (2007). "Common factors versus components: principals and principles, errors and misconceptions," in *Factor Analysis at 100: Historical Developments and Future Directions*, eds R. Cudeck and R. C. MacCallum (Mahwah, NJ: Lawrence Erlbaum Associates), 177–204.

Wirth, R., and Edwards, M. (2007). Item factor analysis: current approaches and future directions. *Psychol. Methods* 12, 58–79.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 June 2012; accepted: 17 February 2013; published online: 27 March 2013.

Citation: Kasper D and Ünlü A (2013) On the relevance of assumptions associated with classical factor analytic approaches. Front. Psychol. ***4****:109. doi:10.3389/fpsyg.2013.00109*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 06 March 2012
doi: 10.3389/fpsyg.2012.00056

How predictable are "spontaneous decisions" and "hidden intentions"? Comparing classification results based on previous responses with multivariate pattern analysis of fMRI BOLD signals

Martin Lages* and Katarzyna Jaworska

School of Psychology, University of Glasgow, Glasgow, UK

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Eric-Jan Wagenmakers, University of Amsterdam, Netherlands
Judit Antal, College Board, USA

****Correspondence:***
Martin Lages, Department of Psychology, University of Glasgow, 8 Hillhead Street, Scotland, G12 8QB, Glasgow, UK.
e-mail: martin.lages@glasgow.ac.uk

In two replication studies we examined response bias and dependencies in voluntary decisions. We trained a linear classifier to predict "spontaneous decisions" and in the second study "hidden intentions" from responses in preceding trials and achieved comparable prediction accuracies as reported for multivariate pattern classification based on voxel activities in frontopolar cortex. We discuss implications of our findings and suggest ways to improve classification analyses of fMRI BOLD signals that may help to reduce effects of response dependencies between trials.

Keywords: prediction accuracy, classification, sequential dependencies, response bias, decision making, free will, frontal cortex

The prospect of decoding brain activity to predict spontaneous or free decisions captivates not only the neuroscientific community (Haggard, 2008) but increasingly inspires researchers in other disciplines (Mobbs et al., 2007; Heisenberg, 2009).

The purpose of this paper is to draw attention to possible confounds and improved data analyses when decoding neural correlates to predict behavior. We focus on a specific task and set of results but we believe that the problem of sequential dependencies is pervasive and needs to be considered carefully when applying machine learning algorithms to predict behavior from brain imaging data.

In two replication studies we illustrate how individual response bias and response dependencies between trials may affect the prediction accuracy of classification analyses. Although our behavioral results are not sufficient to dismiss the original findings based on fMRI BOLD signals they highlight potential shortcomings and suggest alternative ways to analyze the data.

In recent studies Soon et al. (2008) and Bode et al. (2011) used a self-paced free or voluntary decision task to study unconscious determinants preceding "spontaneous" motor decisions. In both studies subjects spontaneously pressed a left or right response button with their corresponding index finger in a series of trials. Brain activity was measured by fMRI BOLD signals and the pattern of voxel activity before (and after) the decision was used to predict binary motor responses. Soon et al. (2008) applied a linear multivariate pattern classifier and searchlight technique to localize patterns of predictive voxel activities (Haxby et al., 2001) and achieved 60% prediction accuracy in a localized region of frontopolar cortex (FPC). Bode et al. (2011), using a high-resolution scan of the prefrontal cortex, reported 57% prediction accuracy for the same task. The authors conclude that patterns of voxel activities in FPC constitute the neural correlate of unconscious determinants preceding spontaneous decisions.

In both studies the activation patterns in FPC occurred up to 10–12 s *before* participants reported their conscious decisions. If validated this finding would dramatically extend the timeline of pre-SMA/SMA (Lau et al., 2004; Leuthold et al., 2004) as well as results on readiness potentials for voluntary acts (Libet et al., 1983; see however Trevena and Miller, 2010) with far-reaching implications (Mobbs et al., 2007; Heisenberg, 2009).

Soon et al. (2008) and Bode et al. (2011) made considerable attempts to control carry-over effects from one trial to the next and selected a subset of trials with balanced left and right responses to eliminate response bias. We argue that despite these precautions response dependencies in combination with response bias in the raw data may have introduced spurious correlations between patterns of voxel activities and decisions.

Naïve participants have difficulties to generate random sequences and sequential dependencies across trials are commonly observed in binary decisions (Lages and Treisman, 1998; Lages, 1999, 2002) as well as random response tasks (Bakan, 1960; Treisman and Faulkner, 1987). In some tasks, these dependencies are not just simple carry-over effects from one trial to the next but reflect stimulus and response dependencies (Lages and Treisman, 1998) as well as contextual information processing (Lages and Treisman, 2010; Treisman and Lages, 2010), often straddling across several trials and long inter-trial-intervals (Lages and Paul, 2006). These dependencies may indicate involvement of memory and possibly executive control (Luce, 1986), especially in self-ordered tasks where generation of a response requires monitoring previously executed responses (Christoff and Gabrieli, 2000).

In order to address the issue of sequential dependencies in connection with response bias we conducted two behavioral replication studies and performed several analyses on individual behavioral data, the results of which are summarized below.

STUDY 1: SPONTANEOUS MOTOR DECISIONS

In this replication study we investigated response bias and response dependency of binary decisions in a spontaneous motor task. We closely replicated the study by Soon et al. (2008) in terms of stimuli, task, and instructions (Soon et al., 2008) but without monitoring fMRI brain activity since we are mainly interested in behavioral characteristics.

METHODS

Subjects were instructed to relax while fixating on the center of a screen where a stream of random letters was presented in 500 ms intervals. At some point, when participants felt the urge to do so, they immediately pressed one of two buttons with their left or right index finger. Simultaneously, they were asked to remember the letter that appeared on the screen at the time when they believed their decision to press the button was made. Shortly afterward, the letters from three preceding trials and an asterisk were presented on screen randomly arranged in a two-by-two matrix. The participants were asked to select the remembered letter in order to report the approximate time point when their decision was formed. If the participant chose the asterisk it indicated that the remembered letter was not among the three preceding intervals and the voluntary decision occurred more than 1.5 s ago. Subjects were asked to avoid any form of preplanning for choice of movement or time of execution.

PARTICIPANTS

All participants ($N = 20$, age 17–25, 14 female) were students at Glasgow University. They were naïve as to the aim of the study, right-handed, and with normal or corrected-to-normal visual acuity. The study was conducted according to the Declaration of Helsinki ethics guidelines. Informed written consent was obtained from each participant before the study.

RESULTS

Following Soon et al. (2008) we computed for each participant the frequency of a left or right response. If we assume that the spontaneous decision task produces *independent* responses then the process can be modeled by a binomial distribution where probability for a left and right response may vary from participant to participant.

$$P[t(x) = s|\theta] = \binom{n}{s} \theta^s (1-\theta)^{n-s}$$

The observed data $t(x)$ is simply the sum of s left (right) responses, n is the total number of responses, and θ is a parameter that reflects the unknown probability of responding Left (Right) with $\theta \in [0,1]$. The hypothesis of a balanced response corresponds to a response rate of $\theta = 0.5$. Rather than trying to affirm this null hypothesis we can test whether the observed number of left (right) responses deviates significantly from the null hypothesis by computing the corresponding p-value (two-sided).

$$p_{(\text{two - sided})} = \sum_{x_i=0}^{n-s} t\,(x_i) + \sum_{x_i=s}^{n} t\,(x_i)$$

We found that response frequencies of 4 out of 20 participants (2 out of 20 if adjusted for multiple tests according to Sidak–Dunn) significantly deviated from a binomial distribution with equal probabilities ($p < 0.05$, two-sided). Soon et al. (2008) excluded 24 out of 36 participants who exceeded a response criterion that is equivalent to a binomial test with $p < 0.11$ (two-sided). Bode et al. (2011) applied a similar response criterion but did not document selection of participants. They reported exclusion of a single participant from their sample of $N = 12$ due to relatively unbalanced decisions and long trial durations; responses from the remaining 11 subjects were included in their analyses. In the present study 8 out of 20 participants did not meet Soon et al.'s response criterion (for details see **Table A1** in Appendix).

Selection of participants is a thorny issue. While the intention may have been to select participants who made truly spontaneous and therefore independent decisions they selected participants who generated approximately balanced responses. This assumption is fallible since subjects' response probabilities are unlikely to be perfectly balanced and the null hypothesis of $\theta = 0.5$ can be difficult to affirm.

Excluding 2/3 of the subjects reduces generalizability of results and imposing the assumption of no response bias on the remaining subjects seems inappropriate because these participants can still have true response probabilities θ that are systematically different from 0.5.

To give an example of how a moderate response bias may affect prediction accuracy of a trained classifier, consider a participant who generates 12 left and 20 right responses in 32 trials. Although this satisfies the response criterion mentioned above, a classifier trained on this data is susceptible to response bias. If the classifier learns to match the individual response bias prediction accuracy may exceed the chance level of 50%. (If, for example, the classifier trivially predicts the more frequent response then this strategy leads to 62.5% rather than 50% correct predictions in our example.)

To alleviate the problem of response bias Soon et al. (2008) and Bode et al. (2011) not only selected among participants but also designated equal numbers of left (L) and right (R) responses from the experimental trials before entering the data into their classification analysis. It is unclear how they sampled trials but even if they selected trials randomly the voxel activities before each decision are drawn from an experiment with unbalanced L and R responses. As a consequence the problem does not dissipate with trial selection. After selecting an equal number of L and R responses from the original data set this subsample still has an unbalanced number of L and R responses in the *preceding* trials so that the distribution of all possible pairs of successive responses in trial $t-1$ and trial t (LL, LR, RR, RL) is not uniform. Since there are more Right responses in the original data set we are more likely to sample more RR "stay" trials and less LR "switch" trials as well as more RL "switch" trials compared to LL "stay" trials. The exact transition probabilities for these events depend on the individual response pattern. Switching and staying between successive responses creates a confounding variable that may introduce spurious correlations between voxel activities from previous responses and the predicted responses. This confound may be picked up when training a linear support vector machine

(SVM) classifier to predict current responses from voxel activities in previous trials.

Similar to Soon et al. (2008) and Bode et al. (2011) we first computed the length and frequency of the same consecutive responses (L and R runs) for each participant and fitted the pooled and averaged data by an exponential function. However, here we fitted the pooled data with the (single-parameter) exponential probability distribution function

$$f(x) = \lambda e^{-\lambda x}, \text{ for } x \geqslant 0$$

(see **Figure 1A**). We found reasonable agreement with the exponential distribution ($R = 0.89$) as an approximation of the geometric distribution. The estimated parameter $\lambda = -0.805$ is equivalent to a response rate of $\theta = 1 - e^{-\lambda} = 0.553$, slightly elevated from $\theta = 0.5$.

Although the exponential distribution suggests an independent and memoryless process such a goodness-of-fit does not qualify as evidence for *independence* and *stationarity* in the individual data. Averaging or pooling of data across blocks and participants can hide systematic trends and patterns in *individual* data.

To avoid these sampling issues we applied the Wald–Wolfowitz or Runs test (MatLab, MathWorks Inc.) to each individual sequence of 32 responses. This basic non-parametric test is based on the number of runs above and below the median and does not rely on the assumption that binary responses have equal probabilities (Kvam and Vidakovic, 2007). Our results indicate that 3 out of the 12 selected participants in our replication study show statistically significant ($p < 0.05$) departures from stationarity (2 out of 12 adjusted for multiple tests). Similarly, approximating the binomial by a normal distribution with unknown parameters (Lilliefors, 1967), the Lilliefors test detected 4 out of 20 (4 out of the selected 12) statistically significant departures from normality in our replication study (3 out of 20 and 1 out of 12 adjusted for multiple tests). These violations of stationarity and normality point to response dependencies between trials in at least some of the participants (for details see **Table A1** in Appendix).

CLASSIFICATION ANALYSIS

In analogy to Soon et al. (2008) and Bode et al. (2011) we also performed a multivariate pattern classification. To assess how much discriminative information is contained in the pattern of previous responses rather than voxel activities, we included up to two preceding responses to predict the following response within an individual data set. Thereto, we entered the largest *balanced* set of left/right response trials and the *unbalanced* responses from the preceding trials into the analysis and assigned every 9 out of 10 responses to a training data set. This set was used to train a linear SVM classifier (MatLab, MathWorks Inc.). The classifier estimated a decision boundary separating the two classes (Jäkel et al., 2007). The learned decision boundary was applied to classify the remaining sets and to establish a predictive accuracy. This was repeated 10 times, each time using a different sample of learning and test sets, resulting in a 10-fold cross validation. The whole procedure was bootstrapped a 100 times to obtain a mean prediction accuracy and a measure of variability for each individual data set (see **Figure 1B**; **Table A1** in Appendix).

One participant (Subject 11) had to be excluded from the classification analysis because responses were too unbalanced to train the classifier. For most other participants the classifier performed better when it received only one rather than two preceding responses to predict the subsequent response and therefore we report only classification results based on a single preceding response.

If we select $N = 12$ participants according to the response criterion employed by Soon et al. (2008) prediction accuracy for a response based on its preceding response reaches 61.6% which is significantly higher than 50% ($t(11) = 2.6$, CI = [0.52–0.72], $p = 0.013$). If we include all participants except Subject 11 then

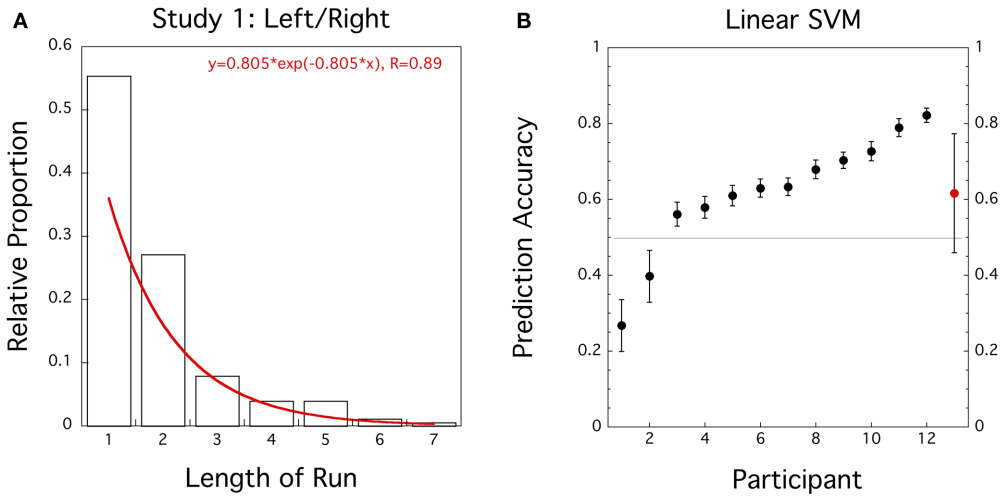


FIGURE 1 | Study 1 ($N = 12$): Left/Right motor decision task. (A) Histogram for proportion of length of response runs pooled across participants. Superimposed in Red is the best fitting exponential distribution function. **(B)** Prediction accuracies of linear SVM trained on preceding and current responses for individual data sets (black circles, error bars = ± 1 SD bootstrapped) and group average (red circle, error bar = ± 1 SD).

average prediction accuracy based on the preceding response was reduced to 55.4%($t(18) = 1.3$, CI = [0.47–0.64], $p = 0.105$).

Our classification analysis illustrates that a machine learning algorithm (linear SVM) can perform better than chance when predicting a response from its preceding response. The algorithm simply learns to discriminate between switch and stay trials. In our replication study this leads to prediction accuracies that match the performance of a multivariate pattern classifier based on voxel activities in FPC (Soon et al., 2008; Bode et al., 2011).

DISCUSSION

Although our behavioral results show that response bias and response dependency in individual data give rise to the same prediction accuracy as a multivariate pattern classifier based on voxel activities derived from fMRI measurements, our behavioral results are not sufficient to dismiss the original findings.

In particular, the temporal emergence of prediction accuracy in voxel patterns as observed in Soon et al. (2008) and Bode et al. (2011) seems to contradict the occurrence of sequential or carry-over effects from one trial to the next because prediction accuracy from voxel activity in FPC starts at zero, increases within a time window of up to 10–12 s before the decision and returns to zero shortly after the decision.

It should be noted however that their results are based on non-significant changes of voxel activity in FPC averaged across trials and participants. We do not know how reliably the predictive pattern emerged in each of the participants and across trials. It is also noteworthy that the hemodynamic response function (HRF) in FPC, modeled by finite impulse response (FIR), peaked 3–7 s *after* the decision was made. Although the voxel activity above threshold did not discriminate between left and right responses, this activity in FPC must serve a purpose that is different from generating spontaneous decisions.

The FIR model for BOLD signals makes no assumption about the shape of the HRF. Soon et al. estimated 13 parameters at 2 s intervals and Bode et al. (2011) used 20 parameters at 1.5 s intervals for each of the approximately 5 by 5 by $5 = 125$ voxels in a spherical cluster. The cluster moved around according to a modified searchlight technique to identify the most predictive region. Although a linear SVM with a fixed regularizer does not invite overfitting the unconstrained FIR model can assume unreasonable HRF shapes for individual voxels. If these voxels picked up residual activity related to the preceding trial, especially in trials where the ITI was sufficiently short, then this procedure carries the risk of overfitting. Activity in the left and right motor cortex, for example, showed prediction accuracies of up to 75% 4–6 s after a decision was reported (Soon et al., 2008).

It may be argued that at least some predictive accuracy should be maintained throughout ITIs if carry-over effects were present between trials. However, voxel activities were sampled at different ITIs due to the self-paced response task. It seems reasonable to assume that ITIs between "spontaneous" decisions are not uniformly distributed. Indeed the individual response times in our replication study were skewed toward shorter intervals, approximating a Poisson distribution that is typical for behavioral response times (Luce, 1986). When voxel activation is temporally aligned with a decision then this effectively creates a time window in which on average residual activation from a previous response is more likely to occur. As a consequence, and despite relatively long average trial durations, the FIR model parameters may pick up residual activation from the previous trial in a critical time window before the next decision (Rolls and Deco, 2011).

Although we would like to avoid a discussion of the difficult philosophical issues of "free will" and "consciousness," the present task implies that participants monitor the timing of their own conscious decisions while generating "spontaneous" responses. The instruction to perform "spontaneous" decisions may be seen as a contradiction in terms because the executive goal that controls behavior in this task is to generate decisions without executive control (Jahanshahi et al., 2000; Frith, 2007). Participants may have simplified this task by maintaining (fluctuating) *intentions* to press the left or right button and by reporting a decision *when* they actually decided to press the button (Brass and Haggard, 2008; Krieghoff et al., 2009). This is not quite compatible with the instructions for the motor task but describes a very plausible response strategy nevertheless.

STUDY 2: HIDDEN INTENTIONS

Interestingly, in an earlier study with $N = 8$ participants Haynes et al. (2007) investigated neural correlates of hidden intentions and reported an average decoding accuracy of 71% from voxel activities in anterior medial prefrontal cortex (MPFCa) and 61% in left lateral frontopolar cortex (LLFPC) before task execution.

We replicated this study in order to test whether response bias and dependency also match the prediction accuracies for delayed intentions. 12 participants (age 18–29, nine female) freely chose between addition and subtraction of two random numbers before performing the intended mental operation after a variable delay (see Haynes et al., 2007 for details). Again, we closely replicated the original study in terms of stimuli, task, and instructions but without monitoring fMRI BOLD signals.

RESULTS

As in Study 1 we tested for response bias and 4 out of 12 participants significantly deviated from a binomial distribution with equal probabilities ($p < 0.05$, two-sided). Since Haynes et al. (2007) do not report response bias and selection of participants we included all participants in the subsequent analyses.

The exponential probability distribution with $\lambda = -0.562$ (equivalent to a response probability $\theta = 0.430$) fitted the pooled data of sequence lengths well ($R = 0.97$; see **Figure 2A**) but the Wald–Wolfowitz or Runs test on each individual sequence of 32 responses indicated that 5 out of 12 participants violated stationarity in the delayed addition/subtraction task (one participant when adjusted for multiple tests). Similarly, a Lilliefors test detected five significant violations of normality (three adjusted for multiple tests, see **Table A2** in Appendix).

One participant (Subject 6) was excluded because responses were too unbalanced to train the linear SVM classifier. We then performed a classification analysis on selected trials with a balanced number of addition/subtraction responses using the preceding response as the only predictor. Averaged across $N = 11$ participants the prediction accuracy of the SVM classifier reached 64.1% which is significantly different from 50% ($t(10) = 3.39$,

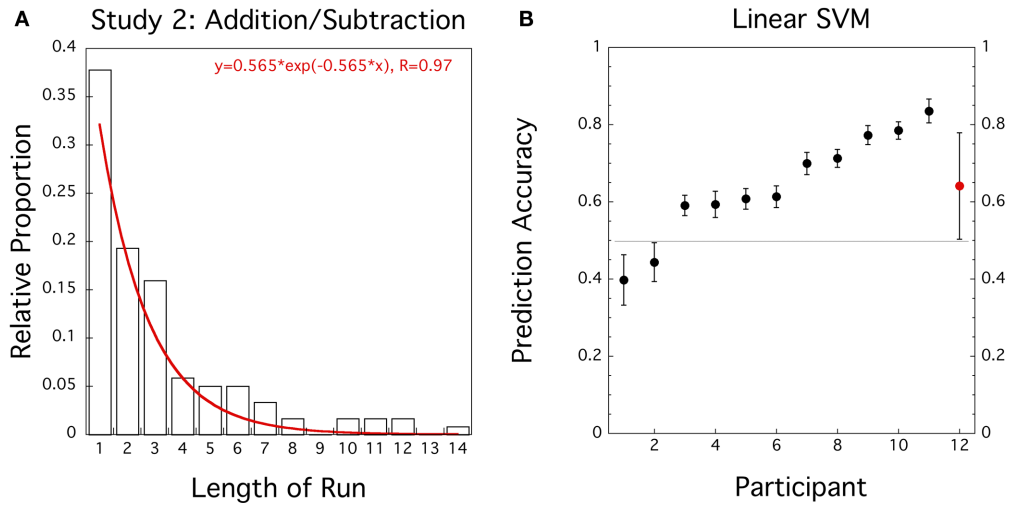


FIGURE 2 | Study 2 ($N = 12$): addition/subtraction in delayed intention task. (A) Histogram for proportion of length of response runs pooled across participants. Superimposed in Red is the best fitting exponential distribution function. **(B)** Prediction accuracies of linear SVM trained on preceding and current responses for individual data sets (black circles, error bars = ± 1 SD bootstrapped) and group average (red circle, error bar = ± 1 SD).

CI = [0.55–0.73], $p = 0.0035$). The classification results are summarized in **Figure 2B** and **Table A2** in the Appendix.

DISCUSSION

It has been suggested that the FPC is implicated in tasks requiring high-level executive control, especially in tasks that involve storing *conscious* intentions across a delay (Sakai and Passingham, 2003; Haynes et al., 2007), processing of internal states (Christoff and Gabrieli, 2000), modulation of episodic memory retrieval (LePage et al., 2000; Herron et al., 2004), prospective memory (Burgess et al., 2001), relational reasoning (Christoff et al., 2001; Kroger et al., 2002), the integration of cognitive processes (Ramnani and Owen, 2004), cognitive branching (Koechlin and Hyafil, 2007), as well as alternative action plans (Boorman et al., 2009). How much a participant can be consciously aware of these cognitive operations is open to discussion but they seem to relate to strategic planning and executive control rather than random generation of responses.

In an attempt to relate activity in the FPC to contextual changes in a decision task, Boorman et al. (2009) reported a study where subjects decided freely between a left and right option based on past outcomes but random reward magnitudes. In this study participants were informed that the reward magnitudes were randomly determined on each trial, so that it was not possible to track them across trials; however, participants were also told that reward probabilities depended on the recent outcome history and could therefore be tracked across trials, thus creating an effective context for directly comparing FPC activity on self-initiated (as opposed to externally cued) switch and stay trials. Increased effect size of relative unchosen probability/action peaked twice in FPC: shortly after the decision and a second time as late as 20 s after trial onset. Boorman et al. (2009) suggest that FPC tracks the relative advantage associated with the alternative course of action over trials and, as such, may play a role in switching behavior. Interestingly, in their analyses of BOLD signal changes the stay trials (LL, RR) differed significantly from the switch trials (LR, RL).

Following neuroscientific evidence (Boorman et al., 2009) and our behavioral results we recommend that multivariate pattern classification of voxel activities should be performed not only on trials with balanced responses but on balanced combinations of previous and current responses (e.g., LL, LR, RL, and RR trials) to reduce hidden effects of response dependencies. Similarly, it should be checked whether the parameters of an unconstrained FIR model describe a HRF that is anchored on the same baseline and shows no systematic differences between switch and stay trials. This will inform whether or not the FIR model parameters pick up residual activity related to previous responses, especially after shorter ITIs.

CONCLUSION

Applying machine learning in form of a multivariate pattern analysis (MVPA) of voxel activity in order to localize neural correlates of behavior brings about a range of issues and challenges that are beyond the scope of this paper (see for example Hanson and Halchenko, 2008; Kriegeskorte et al., 2009; Pereira et al., 2009; Anderson and Oates, 2010; Hanke et al., 2010).

In general, a selective analysis of voxel activity can be a powerful tool and perfectly justified when the results are statistically independent of the selection criterion under the null hypothesis. However, when applying machine learning in the form of MVPA the danger of "double dipping" (Kriegeskorte et al., 2009), that is the use of the same data for selection and selective analysis, increases with each stage of data processing (Pereira et al., 2009) and can result in inflated and invalid statistical inferences.

In a typical behavioral study, for example, it would be seen as questionable if the experimenter first rejected two-thirds of the participants according to an arbitrary response criterion, sampled trials to balance the number of responses from each category in each block, searched among a large number of multivariate predictors and reported the results of the classification analysis with the highest prediction accuracy.

In conclusion, it seems possible that the multivariate pattern classification in Soon et al. (2008) and Bode et al. (2011) was compromised by individual response bias in preceding responses and picked up neural correlates of the intention to switch or stay during a critical time window. The moderate prediction accuracies for multivariate classification analyses of fMRI BOLD signals and our behavioral results call for a more cautious interpretation of findings as well as improved classification analyses.

A fundamental question that may be put forward in the context of cognitive functioning is whether the highly interconnected FPC generates voluntary decisions independently of contextual information, like a *homunculus* or ghost in the machine. After all, the frontal cortex as part of the human cognitive system is highly integrated and geared toward strategic planning in a structured environment. In this sense it seems plausible that neural correlates of "hidden intentions" and "spontaneous decisions" merely reflect continuous processing of contextual information.

ACKNOWLEDGMENTS

Martin Lages was supported by Leverhulme Trust F00-179/BG (UK) and ESRC/MRC RES-060-25-0010 (UK), Katarzyna Jaworska was supported by a Wellcome Trust scholarship and Nuffield Foundation undergraduate bursary.

REFERENCES

Anderson, M. L., and Oates, T. (2010). "A critique of multi-voxel pattern analysis," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Portland.

Bakan, P. (1960). Response-tendencies in attempts to generate random binary series. *Am. J. Psychol.* 73, 127–131.

Bode, S., He, A. H., Soon, C. S., Trampel, R., Turner, R., and Haynes, J. D. (2011). Tracking the unconscious generation of free decisions using ultra-high field fMRI. *PLoS ONE* 6, e21612. doi:10.1371/journal.pone.0021612

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., and Rushworth, M. F. S. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* 62, 733–743.

Brass, M., and Haggard, P. (2008). The what, when, whether model of intentional action. *Neuroscientist* 14, 319–325.

Burgess, P. W., Quayle, A., and Frith, C. D. (2001). Brain regions involved in prospective memory as determined by positron emission tomography. *Neuropsychologia* 39, 545–555.

Christoff, K., and Gabrieli, J. D. (2000). The frontopolar cortex and human cognition: evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology* 28, 168–186.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., and Gabrieli, J. D. E. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage* 14, 1136–1149.

Frith, C. (2007). *Making up the Mind: How the Brain Creates our Mental World.* New York: Blackwell.

Haggard, P. (2008). Human volition: towards neuroscience of will. *Nat. Rev. Neurosci.* 9, 934–946.

Hanke, M., Halchenko, Y. O., Haxby, J. V., and Pollmann, S. (2010). Statistical learning analysis in neuroscience: aiming for transparency. *Front. Neurosci.* 4:38–43. doi:10.3389/neuro.01.007.2010

Hanson, S. J., and Halchenko, Y. O. (2008). Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. *Neural Comput.* 20, 486–503.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430.

Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Curr. Biol.* 17, 323–328.

Heisenberg, M. (2009). Is free will an illusion? *Nature* 459, 164–165.

Herron, J. E., Henson, R. N., and Rugg, M. D. (2004). Probability effects on the neural correlates of retrieval success: an fMRI study. *Neuroimage* 21, 302–310.

Jahanshahi, M., Dirnberger, G., Fuller, R., and Frith, C. D. (2000). The role of the dorsolateral prefrontal cortex in random number generation: a study with positron emission tomography. *Neuroimage* 12, 713–725.

Jäkel, F., Schölkopf, B., and Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *J. Math. Psychol.* 51, 343–358.

Koechlin, E., and Hyafil, A. (2007). Anterior prefrontal function and the limits of human decision-making. *Science* 318, 594–598.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540.

Krieghoff, V., Brass, M., Prinz, W., and Waszak, F. (2009). Dissociating what and when of intentional actions. *Front. Hum. Neurosci.* 3:3. doi:10.3389/neuro.09.003.2009

Kroger, J. K., Sabb, F. W., Fales, C. I., Bookheimer, S. Y., Cohen, M. S., and Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cereb. Cortex* 12, 477–485.

Kvam, P. H., and Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering.* Hoboken, NJ: Wiley Inc.

Lages, M. (1999). *Algebraic Decomposition on of Individual Choice Behavior. Materialien aus der Bildungsforschung' No. 63.* Berlin: Max Planck Institute for Human Development.

Lages, M. (2002). Ear decomposition for pair comparison data. *J. Math. Psychol.* 46, 19–39.

Lages, M., and Paul, A. (2006). Visual long-term memory for spatial frequency? *Psychon. Bull. Rev.* 13, 486–492.

Lages, M., and Treisman, M. (1998). Spatial frequency discrimination: visual long-term memory or criterion setting? *Vision Res.* 38, 557–572.

Lages, M., and Treisman, M. (2010). A criterion setting theory of discrimination learning that accounts for anisotropies and context effects. *Seeing Perceiving* 23, 401–434.

Lau, H. C., Rogers, R. D., Haggard, P., and Passingham, R. E. (2004). Attention to intention. *Science* 303, 1208–1210.

LePage, M., Ghaffar, O., Nyberg, L., and Tulving, E. (2000). Prefrontal cortex and episodic memory retrieval mode. *Proc. Natl. Acad. Sci. U.S.A.* 97, 506–511.

Leuthold, H., Sommer, W., and Ulrich, R. (2004). Preparing for action: inferences from CNV and LRP. *J. Psychophysiol.* 18, 77–88.

Libet, B., Gleason, C. A., Wright, E. W., and Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain* 106, 623–642.

Lilliefors, H. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.* 62, 399–402.

Luce, R. D. (1986). *Response Times. Their Role in Inferring Elementary Mental Organization.* Oxford: Oxford University Press.

Mobbs, D., Lau, H. C., Jones, O. D., and Frith, C. D. (2007). Law, responsibility, and the brain. *PLoS Biol.* 5, e103. doi:10.1371/journal.pbio.0050103

Pereira, F., Mitchell, T., and Botvinick, M. M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, S199–S209.

Ramnani, N., and Owen, A. M. (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nat. Rev. Neurosci.* 5, 184–194.

Rolls, E. T., and Deco, G. (2011). Prediction of decisions from noise in the brain before the evidence is provided. *Front. Neurosci.* 5:33. doi:10.3389/fnins.2011.00033

Sakai, K., and Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nat. Neurosci.* 6, 75–81.

Soon, C. S., Brass, M., Heinze, H. J., and Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nat. Neurosci.* 11, 543–545.

Treisman, M., and Faulkner, A. (1987). Generation of random sequences by human subjects: cognitive operations or psychophysical process. *J. Exp. Psychol. Gen.* 116, 337–355.

Treisman, M., and Lages, M. (2010). Sensory integration across modalities: how kinaesthesia integrates with vision in visual orientation discrimination. *Seeing Perceiving* 23, 435–462.

Trevena, J., and Miller, J. (2010). Brain preparation before voluntary action: evidence against unconscious movement initiation. *Conscious. Cogn.* 19, 447–456.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 November 2011; paper pending published: 06 February 2012; accepted: 13 February 2012; published online: 06 March 2012.

Citation: Lages M and Jaworska K (2012) How predictable are "spontaneous decisions" and "hidden intentions"? Comparing classification results based on previous responses with multivariate pattern analysis of fMRI BOLD signals. Front. Psychology ***3****:56. doi: 10.3389/fpsyg.2012.00056*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.

APPENDIX

Table A1 | Replication study 1: left/right key press in spontaneous motor decision task.

Subject	Resp. bias left/right	Bino. test *p*-value	No of Runs	Runs test *p*-value	Lilliefors *p*-value	SVM pred acc
[1]	21/11	0.1102	11	0.1194	0.0957	[0.5875]
2	16/16	1.0	17	1.0	0.1659	0.3973
3	13/19	0.3771	16	1.0	0.1116	0.2672
[4]	11/21	0.1102	15	1.0	0.0017**	[0.4970]
5	14/18	0.5966	26	0.0009***	0.4147	0.8218
6	15/17	0.8601	15	0.6009	0.0893	0.5607
[7]	22/10*	0.0501(*)	17	0.4906	>0.5	[0.4259]
8	20/12	0.2153	18	0.5647	>0.5	0.6295
[9]	10/22*	0.0501(*)	15	1.0	0.001***	[0.2845]
10	18/14	0.5966	21	0.1689	0.0936	0.6789
[11]	31/1***	0.0001***	3	1.0	>0.5	[NA]
12	19/13	0.3771	20	0.2516	0.4147	0.5790
13	20/12	0.2153	22	0.0280(*)	0.3476	0.7267
14	14/18	0.5966	21	0.1689	0.1659	0.6098
15	16/16	1.0	21	0.2056	0.1116	0.6330
[16]	4/28***	0.0001***	5	0.0739	0.0017***	[0.7513]
17	17/15	0.8601	22	0.0989	>0.5	0.7026
[18]	11/21	0.1102	16	0.9809	0.0893	[0.2985]
19	12/20	0.2153	7	0.00092***	0.001***	0.7891
[20]	11/21	0.1102	16	0.9809	0.0936	[0.2844]
Tot/Avg	**4.25#**	**4 (2)**	**16.2**	**3 (2)**	**4 (4)**	**0.5539**

Participant excluded according to Soon et al.'s (2008) response criterion (binomial test $p < 0.11$); #average deviation: $\Sigma_i|x_i - n/2|/N$ for $n = 32$ and $N = 20$.
$*p < 0.05$, $**p < 0.01$, $***p < 0.001$; (·) number of violations adjusted for multiple tests after Sidak–Dunn.

Table A2 | Replication study 2: addition/subtraction in hidden intention task.

Subject	Resp. bias add/sub	Bino. test *p*-value	No of Runs	Runs test *p*-value	Lilliefors *p*-value	SVM pred acc
1	11/21	0.1102	9	0.0184(*)	0.0112(**)	0.7725
2	14/18	0.5966	10	0.0215(*)	0.0271(*)	0.7124
3	11/21	0.1102	8	0.0057(**)	0.1153	0.6991
4	14/18	0.5966	9	0.0071(**)	0.001***	0.7847
5	21/11	0.1102	12	0.2405	0.001***	0.5932
6	2/30***	0.0001***	5	1.0	>0.5	N/A
7	17/15	0.8601	18	0.8438	0.3267	0.6074
8	20/12	0.2153	12	0.1799	0.0016**	0.5904
9	6/26***	0.0005***	4	0.0006***	0.4443	0.8351
10	18/14	0.5966	16	0.9285	0.2313	0.3972
11	9/23*	0.0201(*)	13	0.8488	0.4030	0.4434
12	9/23*	0.0201(*)	10	0.1273	>0.5	0.6131
Tot/Avg	**5.33#**	**4 (2)**	**10.5**	**5 (1)**	**5 (3)**	**0.6408**

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$; #average deviation: $\Sigma_i|x_i - n/2|/N$ for $n = 32$ and $N = 12$. (·) Number of violations adjusted for multiple tests after Sidak–Dunn.

frontiers in
PSYCHOLOGY

METHODS ARTICLE
published: 25 September 2012
doi: 10.3389/fpsyg.2012.00354

Using classroom data to teach students about data cleaning and testing assumptions

***Kevin Cummiskey*[1]*, *Shonda Kuiper*[2] *and *Rodney Sturdivant*[1]**

[1] *Department of Mathematical Sciences, United States Military Academy, West Point, NY, USA*
[2] *Department of Mathematics and Statistics, Grinnell College, Grinnell, IA, USA*

Edited by:
Jason W. Osborne, Old Dominion University, USA

Reviewed by:
Steven E. Stemler, Wesleyan University, USA
Fernando Marmolejo-Ramos, University of Adelaide, Australia
Martin Dempster, Queen's University Belfast, UK

***Correspondence:**
Kevin Cummiskey, Department of Mathematical Sciences, United States Military Academy, West Point, NY, USA.
e-mail: kevin.cummiskey@usma.edu

This paper discusses the influence that decisions about data cleaning and violations of statistical assumptions can have on drawing valid conclusions to research studies. The datasets provided in this paper were collected as part of a National Science Foundation grant to design online games and associated labs for use in undergraduate and graduate statistics courses that can effectively illustrate issues not always addressed in traditional instruction. Students play the role of a researcher by selecting from a wide variety of independent variables to explain why some students complete games faster than others. Typical project data sets are "messy," with many outliers (usually from some students taking much longer than others) and distributions that do not appear normal. Classroom testing of the games over several semesters has produced evidence of their efficacy in statistics education. The projects tend to be engaging for students and they make the impact of data cleaning and violations of model assumptions more relevant. We discuss the use of one of the games and associated guided lab in introducing students to issues prevalent in real data and the challenges involved in data cleaning and dangers when model assumptions are violated.

Keywords: Guided Interdisciplinary Statistics Games and Labs, messy data, model assumptions

INTRODUCTION

The decisions that researchers make when analyzing their data can have significant impacts on the conclusions of a scientific study. In most cases, methods exist for checking model assumptions, but there are few absolute rules for determining when assumptions are violated and what to do in those cases. For example, when using t-tests or ANOVA, decisions about normality, equal variances, or how to handle outliers are often left to the discretion of the researcher. While many statistics courses discuss model assumptions and data cleaning (such as removing outliers or erroneous data), students rarely face data analysis challenges where they must make and defend decisions. As a result, the impacts of these decisions are rarely discussed in detail.

The topics of data cleaning and testing of assumptions are particularly relevant in light of the fact that there have been several high-profile retractions of articles published in peer-reviewed psychology journals because of data related issues. In June 2012, Dr. Dirk Smeesters resigned from his position at Erasmus University and had a paper retracted from the *Journal of Experimental Psychology* after it was determined his data was statistically highly unlikely. He admitted to removing some data points that did not support his hypothesis, claiming that this practice is common in psychology and marketing research (Gever, 2012). Simmons et al. (2011) show that even when researchers have good intentions, they control so many conditions of the experiment that they are almost certain to show statistically significant evidence for their hypothesis in at least one set of conditions. These conditions include the size of the sample, which outliers are removed, and how the data is transformed. They argue that the ambiguity of how to make these decisions and the researcher's desire to obtain significant results are the primary reasons for the large number of false positives in scientific literature (Simmons et al., 2011).

Replication studies are designed to ensure the integrity of scientific results and may help detect issues associated with the data and model assumptions. However, replication is not a panacea. Miller (2012) shows that the probability of replicating a significant effect is essentially unknowable to the researcher so the scientific community may not always correctly interpret replication study results. Furthering the difficulty in assessing the quality of researchers' decisions involving data is the fact that relatively few replication studies are done on published research. Journals prefer to publish original research and rarely publish replications of previous studies even if the replication shows no effect. Therefore, there is little incentive for researchers to replicate others' results which increases the likelihood that studies resulting in false positives are accepted into scientific journals. Dr. Brian Nosek and a group working on the ambitious Reproducibility Project are attempting to replicate every study published in three major psychology journals in 2008 (Barlett, 2012).

Through student generated datasets, we demonstrate how students in the same class, with the same raw dataset, and using the same statistical technique can draw different conclusions without realizing the assumptions they made in their analysis. There is much truth to Esar's (1949) humorous saying "Statistics [is] the only science that enables different experts using the same figures to draw different conclusions." Instead of having our students believe that statistics is simply a set of step-by-step calculations,

we emphasize the influence a researcher's judgment can have on study conclusions.

Reforms in statistics education encourage an emphasis on understanding of concepts, interpretation, and data analysis instead of formulas, computation, and mathematical theory (Garfield et al., 2007; DeVeaux and Velleman, 2008). Curricula based on these reforms move away from teaching statistics as a collection of facts. Instead, they encourage the scientific process of interdisciplinary data analysis as statistics is actually practiced. Paul Velleman states, "It seems that we have not made [this] clear to others – and especially not to our students – that good statistical analyses include judgments, and we have not taught our students how to make those judgments" (Velleman, 2008). Our classroom activities and corresponding datasets demonstrate the importance of emphasizing these points and offer ideas for those teaching courses involving data analysis and experimental design to introduce the discussion in the classroom.

THE TANGRAMS GAME AND LAB

Tangrams is an ancient Chinese puzzle where players arrange geometrically shaped pieces into a particular design by flipping, rotating, and moving them. The online Tangrams game and the web interface, shown in **Figure 1**, allow students the opportunity to play many versions of the original game.

Prior to starting the game, the class decides upon one or more research questions they want to investigate as a group. For example, students may decide to test whether the game completion time depends on the type of music played in the background, or they could test if one gender is more likely to use hints. Students then design the experiment by determining appropriate game settings and conditions for collecting the data. After the student researchers design the experiment, they become subjects in the study by playing the game. The website collects the players' information and records their completion times. The data is available for immediate use through the website. If one research study is designed for the entire class, every student plays the game under similar conditions and a large sample of data is immediately available through the website for analysis. The students return to their role of researcher using the data that they just collected.

Next, students (as a class or in small groups) make decisions about data cleaning, check assumptions, perform a statistical test of significance, and state their conclusions. Classroom testing of the Tangrams game and associated labs over the last three semesters has given us a rich data set demonstrating the impacts of data cleaning and the importance of validating the assumptions of statistical tests.

The Tangrams game-based lab gives students exposure to the entire research process: developing research questions, formulating hypotheses, designing experiments, gathering data, performing statistical tests, and arriving at appropriate conclusions. This lab is a fun and effective way for instructors to transition from textbook problems that focus on procedures to deeper learning experiences that emphasize the importance of proper experimental design and understanding assumptions.

IMPACTS OF DATA CLEANING

Figure 2 shows boxplots of data collected at West Point in the fall semester of 2011 for one research question. The dependent variable is the time to successfully complete a specified Tangrams puzzle. The independent variable is athlete (Yes means the student plays on a collegiate team while No means the student does not play on a collegiate team). Students were given an overview of the game by their instructor and then allowed to practice the game once on a different puzzle in order to get familiar with the game controls. For both groups, the distributions of completion times appear unimodal with a large positive skew. There are several outliers present within the dataset. Discussing the data with students tends to provide the following reasons for at least some of the very high times:

1. The student did not fully understand the object of the game even after the practice game.

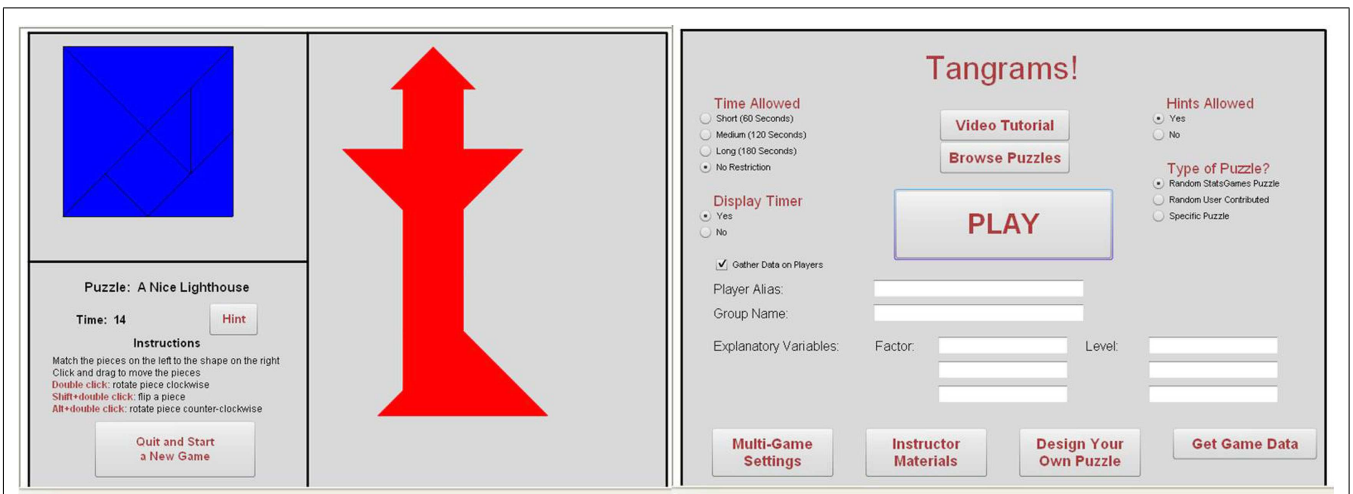


FIGURE 1 | Tangrams web interface.

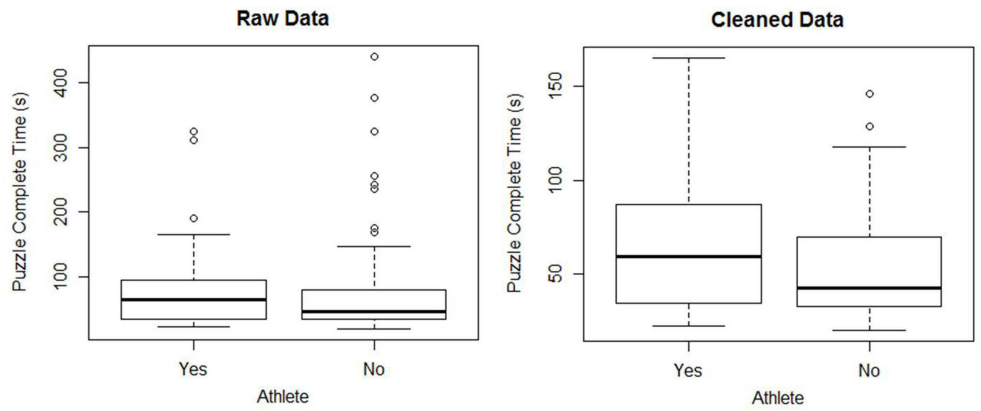


FIGURE 2 | Side-by-side boxplots of the completion time of the Tangrams game for raw and cleaned data.

2. The student did not fully understand how to manipulate the pieces even after the practice game (some puzzle shapes require a piece to be flipped while others do not).
3. The full attention of the student was not on the game during the entire time recorded.

Any one of these reasons could justify removing that observation from the analysis. Before conducting their analysis, some students removed the positive outliers shown in the boxplots of the raw data in **Figure 2**[1] (Carling, 2000; Ueda, 2009). For the rest of this paper, we will refer to the data set after removing these outliers as the cleaned data. Through class discussion of the data after the experiment, students recognized issues with the conduct of the experiment and the importance of understanding the data collection mechanism. They were able to formulate recommendations for improving the future experiments such as better control of extraneous variables and including a method for getting feedback from players to determine if their results were erroneous.

The decision on whether or not to keep outliers or erroneous data in the analysis has a very clear impact on the results. For example, **Table 1** shows that removing the outliers identified within the boxplots in **Figure 2** can change the p-value from 0.478 to 0.058 for a one-way ANOVA. Most students found the difference in p-values surprising, especially given that the sample sizes of both groups are larger than 30. Many researchers would interpret a p-value of 0.058 as being small enough to conclude that there is some evidence that there is a difference between the two population means. This conclusion is clearly different than the one we would reach with all the data points.

IMPACTS OF INVALID MODEL ASSUMPTIONS

In addition to considering impacts of cleaning data, results of these classroom experiments show the impact of model assumptions on the conclusions of the study. The two sample t-test and one-way ANOVA are both parametric hypothesis tests used to determine

Table 1 | Summary statistics for raw and cleaned data.

	Raw data		Cleaned data (outliers removed)	
	Athlete	**Non-athlete**	**Athlete**	**Non-athlete**
Sample size	36	92	33	84
Sample mean	82.72	72.50	65.23	53.02
SD	72.00	73.50	39.35	27.11
	***p*-value = 0.478** (one-way ANOVA on difference in means)		***p*-value = 0.058** (one-way ANOVA on difference in means)	

if there is a difference between the means of two populations. In our case, we want to see if the difference between the means of the athletes and non-athletes is statistically significant. The null (H_0) and alternate (H_a) hypotheses are:

$$H_0 : \mu_A = \mu_N$$
$$H_a : \mu_A \neq \mu_N$$

where μ_A and μ_N are the means of the athlete and non-athlete populations. Both tests assume that we have random samples from their respective populations and that each population is normally distributed. The one-way ANOVA also assumes equal variances. However, the two-sample t-test can be conducted without the equal variance assumption (sometimes called Welch's t-test). In this section, we will discuss the equal variance and normality assumptions.

Some texts suggest that formal tests should be used to test for equal variances. However, some tests, such as Bartlett's test (and the F-test), are very sensitive to non-normality. Even with the outliers removed, the cleaned data is still strongly skewed right (see **Figure 2**). Box criticized using Bartlett's test as a preliminary test for equal variances, saying "To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave

[1] In introductory courses, graphical methods and rules of thumb are usually used to detect outliers. Formal tests such as the Ueda method and Carling's method could be used in more advanced courses.

port" (Box, 1953). Levene's test of homogeneity of variance is less sensitive to departures from normality (Levene, 1960; Brown and Forsythe, 1974). For the cleaned data, Levene's test gives a p-value of 0.023, indicating there is evidence of unequal variances.

A commonly used informal test is to reject the equal variance assumption when the ratio of SD (largest SD over the smallest SD) is greater than 2 (Moore and McCabe, 2003). Using this rule of thumb on our data, the ratios for the raw and cleaned data are 1.02 and 1.45, respectively. Using this informal rule, we fail to reject the assumption of equal variances for both the raw and cleaned data.

Whether or not the researcher decides to assume the two populations have equal variances will contribute to the choice of which statistical test to perform and has a surprisingly large impact on the p-value of the two sample t-test. Without assuming equal variances, the p-value of the two sample t-test on the cleaned data is 0.109, which is considerably larger than the p-value of 0.058 found when assuming equal variances. Note that the p-value for the t-test assuming equal variances and the one-way ANOVA are mathematically equivalent and result in the same p-value.

To explain the impacts of this equal variance assumption, we need to recognize the influence of unequal sample sizes. When the group with the smallest sample size has a larger SD, the mean square error (or pooled SD) is likely to underestimate the true variance. Then ANOVA is likely to incorrectly reject the null hypothesis (conclude that there are differences when there really are no differences between group means).

A second assumption that a researcher should validate is that both samples are from normally distributed populations. From the boxplots in **Figure 2**, a student should suspect that the population distributions are not normal. Additional tools such as histograms and normal probability plots clearly show that the sample data is not normally distributed. For both athletes and non-athletes, the Shapiro–Wilks test for normality rejects the null for both samples with p-values less than 0.001, providing further evidence that the assumption of normality is not valid (see **Table 2** for a summary of the results of various Shapiro–Wilks tests).

When faced with data that indicates the normality assumption is not valid, transforming the data is one method to allow the analyst to proceed with the analysis. In this case, taking the log of the completion times results in plots that appear much closer to the shape of the normal distribution[2]. **Figure 3** shows a boxplot of the cleaned data after the log transformation. It is more appropriate to conduct the statistical test using the transformed data since the normality assumption is more closely met. When the two sample t-test (unequal variances) is performed on the transformed data, the resulting p-value is 0.307 which would lead us to conclude that there is not a significant difference between athletes and non-athletes. These results are somewhat different from the p-value of 0.478 obtained using the raw data, although the difference in the p-value did not change the conclusion of the test.

Researchers using the cleaned data face similar issues. Even after removing the outliers, the cleaned data still is strongly skewed to the right. Once again, a log transform improves the normality assumption. Conducting the two sample t-test on the cleaned log transformed data results in a p-value of 0.18.

We have shown that even when students start their analysis with the same raw dataset, decisions involving data cleaning and validation of model assumptions cause p-values to vary between 0.058 and 0.478. **Table 3** summarizes the different p-values based on the assumptions we discussed in the last two sections. This clearly demonstrates that model assumptions need to be checked before any statistical conclusions are drawn. It also shows that a researcher determined to find significant results can do so by choosing a set of assumptions resulting in the smallest p-value.

In introductory statistics courses, this dataset can be used to focus on tests that are based on the equal variance assumption and the normality assumption (t-tests and ANOVA) and how the violation of these assumptions can influence p-values as shown in **Table 3**. However, there are several other statistical techniques that are typically beyond the scope of an introductory course that can be discussed with this dataset. In addition to Levene's test and the Shapiro–Wilks test shown above, instructors could discuss the following:

- More advanced methods, such as the Box-Cox power transformation, can be used to find a better transformation (Osborne, 2002; Olivier and Norberg, 2010).

[2]While the log transformation is helpful, both groups still show some evidence of lack of normality. Other, less common, transformations did better fit the normality assumption, however, the resulting p-values were fairly similar.

Table 2 | Summary of Shapiro–Wilks normality test under various conditions.

	Athlete	Non-athlete
Raw data	<0.001	<0.001
Cleaned data	0.00157	<0.001
Log-transformed raw data	0.118	<0.001
Log-transformed cleaned data	0.153	0.0521

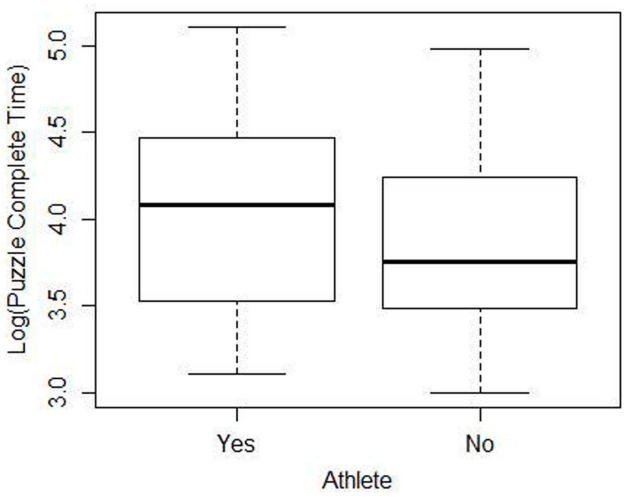


FIGURE 3 | Side-by-side boxplots of the log of completion time for the cleaned data.

Table 3 | Summary of *p*-values for the difference in means between Tangrams completion times of athletes and non-athletes under various assumptions.

	Raw data (*p*-value)	Cleaned data (*p*-value)
Two-sample *t*-test assuming equal variances	0.478	0.058
Two-sample *t*-test without assuming equal variances	0.475	0.109
Two-sample *t*-test assuming equal variance using the log-transformed data	0.307	0.139
Two-sample *t*-test assuming equal variance using the log-transformed data	0.323	0.180

- Response time is often modeled with the exponentially modified Gaussian distribution (ex-Gaussian distribution). This game typically provides a relevant example of data that is better modeled with a distribution other than the normal distribution (Marmolejo-Ramos and González-Burgos, 2012).
- Many students believe that if the sample size is greater than 30, the Central Limit Theorem guarantees that tests based on the normal distribution are appropriate to use. Tim Hesterberg has written an easy-to-understand article that is freely available online that challenges this assumption (Hesterberg, 2008). Having students read this article helps them understand that while the "sample size greater than 30" is a nice guideline, it is not an absolute rule.
- The data can be used to demonstrate better tools to enhance the visibility of data that is not normally distributed. For example, the shifting boxplot and the violin plot with confidence interval around the mean provide additional information not displayed in the basic boxplot (Marmolejo-Ramos and Matsunaga, 2009; Marmolejo-Ramos and Tian, 2010).
- The analysis of the Tangrams data is equally suited for courses that emphasize other concepts, such as Bayesian approaches to data analysis (Kruschke, 2011).
- The data can be analyzed with techniques that are not based on the normality assumption, such as randomization and permutation tests.

RANDOM SAMPLING

In our experience, the assumptions of statistical tests such as those discussed in the previous section are at least covered in typical statistics textbooks and classes. An assumption that is addressed far less often but that should always be validated before any statistical conclusions are drawn about populations is that each observation is a sample randomly selected from the population. Homework-style problems do not give enough information about how the data was collected for students to consider the quality of the sample. When students conduct their own experiment, the students are more aware of this issue in the resulting data.

Some sources of bias are easily identifiable. For example, this sample gathered at West Point is most certainly not generalizable to all colleges nationwide, as only about 15% of cadets are female. In addition, it is important to discuss the impact of a researcher acting as a subject within their own research study. Other sources of bias are not so easily identifiable but warrant discussion. This is especially the case with observational studies such as ours where the students enrolled in the course are acting as subjects in the study. For example, it could be possible that there are other factors that the athletes in our sample share that were the true reason for differences in their times when playing the puzzle game. One possibility is that the athletes have early morning practice and are therefore more tired than the non-athletes. Also, because students decided and knew which variables were being investigated, there is the possibility of stereotype threat, where groups are primed to perform better or worse.

From our experience, the following discussion questions are effective at addressing some of the important issues involved with random sampling and designing experiments.

- If you had a chance to play the game under different circumstances, would you be able to perform better? Describe any factors that may have kept you from doing your best while playing the game.
- Do you think outside factors (like the ones you or others mentioned in the previous question) could impact the results of a study? Should researchers provide some type of motivation for their subjects in various studies to do their best?
- If you were conducting this study again, how would you control for any key factors that may influence each subject's performance?
- Ask small groups to design their own study, addressing other not so obvious conjectures.
  - Are there any variables (such as academic major or gender) that explain why some players solve Tangrams puzzles faster than others?
  - What type of student improves the most when playing this type of game?
  - Is it helpful for a second student to provide hints?

ADDITIONAL DATA SETS

The Tangrams game and corresponding labs are designed to be flexible so that students can design studies related to their interests. In another semester at West Point, we investigated the relationship between academic major and Tangrams performance. Students majoring in math, science, and engineering disciplines (MSE) were compared to those majoring in other disciplines (non-MSE). In this study, the raw data resulted in a *p*-value of 0.353 for the two sample *t*-test. When outliers were removed from the data, the *p*-value decreased to 0.071. Other classes have tested two-way or three-way factorial designs. For any variables that students are interested in testing, the nature of the Tangrams game tends to produce positively skewed data and outliers. Thus, any study conducted by students with this game provides opportunities to demonstrate the importance of data cleaning and model assumptions. **Table 4** contains a list of suggested independent variables that could be used to explain Tangrams performance.

In this paper, we focused on the time to complete the puzzle as the response or dependent variable. The Tangrams game offers

Table 4 | Candidate independent variables to explain Tangrams performance.

Variable	Research question
Gender	Do males or females perform better at tangrams?
Academic major	Do students majoring in science, technology, engineering, and mathematics perform better at tangrams than other students?
Type of high school attended	Do students who attended private or public high schools perform better at tangrams?
Athlete	Do college athletes perform better at tangrams than non-athletes?
Political affiliation	Do students who affiliate with the democratic, republican, or other parties perform better at tangrams?
Academic performance	Do students who made the dean's list perform better at tangrams than those that did not?

Table 5 | Candidate dependent variables.

Variable	Description
Puzzle completion time	Time to complete a tangrams puzzle
Puzzle success or failure	Given a fixed amount of time, whether or not a student can complete the puzzle
Number of moves	Number of moves (a flip or rotation) required to solve the puzzle
Time to quit	Time before a student quits a puzzle that is impossible to solve
Time to receive a hint	Time until a student asks the game for a hint
Number of puzzles solved	Given a fixed amount of time, the number of puzzles that a student can solve

many more dependent variables that can be investigated. **Table 5** contains a list of some of the other dependent variables that can be investigated using Tangrams.

In more advanced courses, small groups of student researchers have used these online games to develop more complex research studies. For example, one group built upon Butler and Baumeister's (1998) findings that a friendly observer (compared to neutral or hostile observers) had a detrimental effect on the ability of participants to accurately complete difficult tasks. They conducted a study with a repeated measures design to determine whether this effect would be the same if the friendly observer actively offered advice on how to solve Tangrams puzzles.

We have developed a series of online games and associated labs like the one discussed in this paper. Multiple independent variables can be used to test memory or spatial reasoning skills. Students can choose variables (such as using hints, amount of time allowed, or number of pieces) that are very likely to result in small p-values even with small sample sizes. Other independent variables, such as gender or major, are less likely to have small p-values. When multiple studies are conducted, we typically find no more than one or two of the studies will show significant differences among gender. This can lead to very interesting discussions of publication bias that can occur when only research studies with small p-values are published. Each game-based lab or research project allows for quick, anonymous, and automated data collection that can be used to demonstrate the importance of impacts of data cleaning and model assumptions on the results of a study.

STUDENT COMMENTS AND ASSESSMENT

In statistics education, it is often challenging to have students experience the true complexities of conducting a large research study. Designing an experiment, collecting and analyzing data, and deciding upon and carrying out the appropriate statistical test are usually too time-consuming, costly, or impractical for an introductory statistics course. As a result, most students learn statistical calculations without a tie to the context of the scientific research and become disinterested in, and even cynical toward, statistics. An alternative to lectures and textbook style problems is to incorporate research-like experiences in the classroom.

The Classroom Undergraduate Research Experience (CURE) survey of undergraduate science evaluated courses that contain research-like experiences. Research-like experiences are activities that contain "Group work, reading primary literature, data collection and analysis... students conduct research in which the outcome is not known (even to the course instructor) and students have at least some input into the research topic and design of the methodological approach." Results of the CURE survey show that "Students in high research-like courses report learning gains similar in kind and degree to gains reported by students in dedicated summer research programs" (Lopatto, 2010). In addition, Wei and Woodin (2011) found "Evidence is emerging that these approaches are introducing more underrepresented minorities to scientific research in the classroom." These game-based labs are designed so that, with technology, students can gain many of the benefits of research-like experiences.

A formal assessment of student attitudes and learning was conducted throughout the 2011–2012 school year. These materials were taught in five sections of an introductory statistics course. When we asked students what they liked best about the lab, typical responses were:

- The data set was real and we played a part in creating it.
- To be able to see an actual scenario where what we learned can be used.
- The fact that we could collaborate...
- ...work at own pace, ask questions as needed.
- I liked that getting the data was very quick and easy.
- Playing the game!

As a group, students enjoyed playing the games. Even though the online game is fairly simple with plain graphics, it was considered a welcome break from normal classroom activities. The level of interest in trying to explain potential biases in Tangrams performance was very high. Ideas ranged from number of hours of sleep to SAT scores to the age of the player. This activity also seemed

Table 6 | Survey results for 115 students after completing the Tangrams lab.

Survey question	Strongly agree (%)	Agree (%)	Neutral (%)	Disagree (%)	Strongly disagree (%)
The Tangrams lab was a good way of learning about hypothesis testing	43	38	8	7	3
Students who do not major in science should not have to take statistics courses	5	10	23	37	24
Statistics is essentially an accumulation of facts, rules, and formulas	10	34	30	19	6
Creativity plays a role in research	30	47	12	7	4
If an experiment shows that something does not work, the experiment was a failure	9	2	5	31	52
The tangrams lab had a possible effect on my interest in statistics	17	38	32	13	1

to truly engage students who were otherwise quiet throughout the semester.

Many students commented that they liked using "real" data for the *first* time in their course. This comment came as a surprise because the instructors had used data from economics, sports, scientific, and military websites in lessons prior to this lab. However, to the students, if they are not involved in the collection of the data, it is not real to them. Involving students in data collection makes them much more interested in the outcome of a statistical process. In addition, messy data makes the decision process more real to students.

In many courses using the lab, students had yet to actively experience the context for the statistical procedures they were learning. They had only seen textbook type questions that give them the research question, the experiment, the data, and the statistical procedure to use. After completing the lab, many students commented that they saw how statistical procedures are actually used by people outside the statistics classroom. Survey results suggest that students enjoyed the lab and felt like they had learned from the experience. In this assessment, 81% of students either agreed or strongly agreed that the Tangrams lab was a good way of learning about hypothesis testing, while only 10% disagreed. seventy-four percent either agreed or strongly agreed that the Tangrams lab improved their understanding of using statistics in research. Complete results of the survey are displayed in **Table 6**.

Although our students have laptop computers and are required to bring them to class, the Tangrams lab has been implemented in other classroom conditions. In large sections, where each student does not have a laptop, students can play the game outside of class in preparation for the next class period. If no computers are available in class, the guided labs that we have available are detailed enough to allow students to do most of the computational work outside the classroom, where most students presumably have access to a computer. The instructor can then use class time to discuss results and interpretations of findings.

CONCLUSION

Online games and guided labs such as Tangrams are fun and effective ways to incorporate a research-like experience into an introductory course in data analysis or statistics. The labs leverage their natural curiosity and desire to explain the world around them so they can experience both the power and limitations of statistical analysis. They are an excellent way for instructors to transition from textbook problems that focus on procedures to a deeper learning experience that emphasizes the importance of proper experimental design and understanding assumptions. While playing the role of a researcher, students are forced to make decisions about outliers and possibly erroneous data. They experience messy data that make model assumptions highly questionable. These labs give students the context for understanding and discussing issues surrounding data cleaning and model assumptions, topics that are generally overlooked in introductory statistics courses.

ACKNOWLEDGMENTS

Sample materials and datasets are freely available at http://web.grinnell.edu/individuals/kuipers/stat2labs/. A project-based textbook that incorporates some of these games is also available at http://www.pearsonhighered.com/kuiper1einfo/index.html. These materials were developed with partial support from the National Science Foundation, NSF DUE# 0510392, and NSF TUES DUE #1043814. The original games were developed by Grinnell College students under the direction of Henry Walker and Sam Rebelsky. John Jackson and William Kaczynski of the United States Military Academy provided significant input to the classroom materials.

REFERENCES

Barlett, T. (2012). Is psychology about to come undone. *The Chronicle of Higher Education*. Available at: http://chronicle.com/blogs/percolator/is-psychology-about-to-come-undone/29045 (accessed April 17, 2012).

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* 40, 333.

Brown, M. B., and Forsythe, A. B. (1974). Robust tests for equality of variances, *J. Am. Stat. Assoc.* 69, 364–367.

Butler, J. L., and Baumeister, R. F. (1998). The trouble with friendly faces: skilled performance with a supportive audience. *J. Pers. Soc. Psychol.* 75, 1213–1230.

Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Comput. Stat. Data Anal.* 33, 249–258.

DeVeaux, R., and Velleman, R. (2008). Math is music: statistics is literature. *Amstat News* 375, 54.

Esar, E. (1949). "The dictionary of humorous quotations," in *Mathematics 436 – Finely Explained (2004)*. ed. R. H. Shutler, 3. Victoria: Trafford Publishing.

Garfield, J., Aliaga, M., Cobb, G., Cuff, C., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., and Witmer, J. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE): College Report*. Alexandria: American Statistical Association.

Gever, J. (2012). Stats can trip up scientific data fabricators. *MedPage Today*. Available at: http://www.medpagetoday.com/PublicHealthPolicy/Ethics/33655 (accessed July 07, 2012).

Hesterberg, T. (2008). "It's time to retire the 'n>=30' Rule," in *Proceedings of the American Statistical Association, Statistical Computing Section (CD-ROM)* (Alexandria: American Statistical Association).

Kruschke, J. (2011). *Doing Bayesian Data Analysis*. Amsterdam: Academic Press.

Levene, H. (1960). "Robust Tests for Equality of Variances," in *Contributions to Probability and*

Statistics: Essays in Honor of Harold Hotelling, ed. I. Olkin (Stanford: Stanford University Press), 278–292.

Lopatto, D. (2010). Undergraduate research as a high-impact student experience. *Peer Rev.* 12, 27–30.

Marmolejo-Ramos, F., and González-Burgos, J. (2012). A power comparison of various tests of univariate normality on ex-Gaussian distributions. *Methodology (Gott)*. doi: 10.1027/1614-2241/a000059

Marmolejo-Ramos, F., and Matsunaga, M. (2009). Getting the most from your curves: Exploring and reporting data using informative graphical techniques. *Tutor. Quant. Methods Psychol.* 5, 40–50.

Marmolejo-Ramos, F., and Tian, T. S. (2010). The shifting boxplot. A boxplot based on essential summary statistics around the mean. *Int. J. Psychol. Res.* 3, 37–45.

Miller, J. (2012). What is the probability of replicated a statistically significant effect? *Psychon. Bull. Rev.* 16, 617–640.

Moore, D. S., and McCabe, G. P. (2003). *Introduction to the Practice of Statistics*, 4th Edn, New York: W. H. Freeman and Company, 755.

Olivier, J., and Norberg, M. M. (2010). Positively skewed data: revisiting the Box-Cox power transformation. *Int. J. Psychol. Res.* 3, 68–75.

Osborne, J. (2002). Notes on the use of data transformation. *Pract. Assess. Res. Eval.* 8. Available at: http://pareonline.net/getvn.asp?v=8&n=6

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366.

Ueda, T. (2009). A simple method of the detection of outliers, Electron. *J. Appl. Stat. Anal.* 2, 67–76.

Velleman, P. (2008). Truth, damn truth, and statistics. *J. Stat. Educ.* 16, 14.

Wei, C. A., and Woodin, T. (2011). Undergraduate research experiences in biology: alternatives to the apprenticeship model. *CBE-Life Sci. Educ.* 10, 130.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 July 2012; paper pending published: 06 August 2012; accepted: 02 September 2012; published online: 25 September 2012.

*Citation: Cummiskey K, Kuiper S and Sturdivant R (2012) Using classroom data to teach students about data cleaning and testing assumptions. Front. Psychology **3**:354. doi: 10.3389/fpsyg.2012.00354*

This article was submitted to Frontiers in Quantitative Psychology and Measurement, a specialty of Frontiers in Psychology.