Challenges for quantitative psychology and measurement in the 21st century

Jason W. Osborne*

North Carolina State University, Raleigh, NC, USA
*Correspondence: jason_osborne@ncsu.edu

Quantitative researchers exist in the exciting nexus where knowledge is created from raw data. Through quantitative study of the human condition, we hope to gain insight into basic, fascinating questions that humans have pondered for millennia. We (and the quantitative psychologists that have preceded us) are therefore optimists above all else. We believe that through systematic, rigorous study, we are able to gain insight into behavior, psychological processes, and important outcomes that ultimately can benefit the world and its inhabitants. Yet the promise of quantitative study of psychology is also one of its greatest challenges: demonstrating in a convincing way that quantification of behavioral, cognitive, biological, and psychological processes is valid, and that the analyses we subject the numbers to are honest efforts at elucidation rather than obfuscation.

We enter a new era of possibility as previously unimaginable technologies become available to us. One example is fMRI, that some believe quantifies processes that reveal functions of parts of the brain. Previously only imagined in science fiction, fMRI may be the ultimate tool for the study of psychology, yet there are significant questions as to what exactly it is that fMRI reveals, and how best to analyze and present those data (e.g., Haller and Bartsch, 2009; Hemmelmann et al., 2009; e.g., Wang et al., 2009; Yuanqing et al., 2009). Those who want to use this potentially paradigm-changing methodology need to convince the community of science that what they are quantifying and reporting really reflects what they say it does. In the same way, scientists who want to study student achievement, intelligence, attitudes, overt behavior, intentions, beliefs, emotions, stress, race/ethnicity, and indeed even health outcomes (which are just a few of the important variables we as social scientists are interested in measuring and analyzing) must redouble their efforts to convince the community of consumers of

science that our numbers really represent what we assume or propose that they represent. At stake is nothing less than the integrity and future of our field.

Most of us have never seriously questioned whether the numbers we report are meaningful, whether they represent the attributes and processes we believe them to. Our field has a long history of loyal skeptics who question our assumptions, challenging our tacit beliefs, debating important points. For example, one of the most common procedures performed in our field, Null Hypothesis Statistical Testing (NHST) grew out of vigorous debates between Ronald Fisher and the collaborative team of Jerzey Neyman and Egon Pearson; (Fisher, 1925; Neyman and Pearson, 1936), influencing how we perform statistical inference throughout much of the 20th century (and today). However, NHST is also an example of why our field needs to periodically revisit our assumptions and legacies to determine if they are still valid. Today, NHST serves as a 20th century methodological legacy that is increasingly being challenged (e.g., Killeen, 2008 and many others). Other traditions and practices (e.g., creating sum scores for psychological scales via simple averaging, excluding cases with missing values, to name but two of many) deserve close scrutiny as to whether they are justified as best practices. To blindly accept the dogma of the field without scholarly examination is to diminish what we do. If we cannot convincingly demonstrate that the quantifications we work with are substantively meaningful, that the procedures and strategies we use are the best way to do things, if we cannot cogently answer the skeptics and critics, we have a problem. I believe the greatest challenge to our field is to continue to demonstrate convincingly that what we do is meaningful, important, and relevant. And I believe that we can successfully rise to this challenge, and in the process become stronger as a field. In order to encourage this rare type of

collegial discourse, I have invited a prominent, scholarly skeptic to join the editorial board of impressive quantitative scholars to serve as the "loyal opposition" raising questions and challenging assumptions. Those of you who are not on the editorial board but are interested in this epistemological debate are encouraged to use this journal as a forum where we can thoughtfully explore and (hopefully) defend our most important assumptions in the field.

In the introduction to my book, Best Practices in Quantitative Methods (Osborne, 2008), I argue that quantitative researchers are under a moral and ethical imperative to apply their skills in such a way to produce the most defensible, unbiased, generalizable, and applicable results possible. Why? Because what we do has the potential to make a difference (for better or worse). Inappropriate or misapplied quantitative techniques, lack of attention to data quality, and inappropriate generalization can result in unfortunate consequences: governments and organizations can waste resources on sub-optimal interventions and decisions, educators can be inspired to abandon tried-and-true methods for novel (yet inferior) pedagogies, health care workers can utilize sub-optimal treatments, etc. Our profession has the potential to make a tremendous, continual contribution to the well-being of humanity. But when we lose sight of the reason we want to do research, we have the potential to do great harm. Just as getting a new drug to market is valuable *only* if that drug actually improves the human condition in some way, pet theories and lengthy publishing histories are all well and good, but they are only valuable to the extent they make the world a better place in some small (or not so small) way. We must be vigilant, as researchers, to keep this lesson foremost in our minds, to keep challenging ourselves to make a difference, to practice our profession using only superior methodology, and to continue questioning and examining our tacit assumptions.

Psychology as a field, and quantitative psychology and measurement in particular, has experienced explosive progress recently in terms of the choices of analytic techniques and measurement options available. At the dawn of the 20th century, Student's *t* test was just being broadly disseminated (Student, 1908), and most psychologists had to perform calculations by hand, with paper and pencil. By the time I entered my doctoral program in 1990, the field was embracing tools and techniques unimaginable decades earlier: multivariate statistics, latent variable modeling, modern measurement methodologies, multilevel modeling, sophisticated meta-analytic techniques, new estimation procedures, and even tools that appear to assess physiological indicators of psychological activity. I wonder what tools and techniques will be available to scholars at the end of this century, and whether we would be able to comprehend them.

Our job is to help the field move toward this unknowable future. At this, the dawn of the 21st century, there are remarkably promising signs. Researchers are beginning to understand that strict null hypothesis statistical testing (NHST) is limiting and provides an incomplete picture of results. More journals now require effect sizes, confidence intervals, and other practices one might argue are well overdue. We have more computing power in our cell phones these days than in the university and corporate mainframes I started out programming 30 years ago. Our software tools are so powerful and sophisticated that we now can ask questions of our data that were barely imagined even a decade or two ago. We have ways of understanding measurement that allow us to create ever more sophisticated quantifications of human attributes and behaviors. Truly, this is a wonderful time to be a quantitative researcher. I believe we must use these ever more effective tools to renew and freshen the field of quantitative methods through evidence-based promotion of best practices. Research-based conclusions are only as good as the evidence they are based on, and only to the extent that the analyses are done in the best way possible. It seems every year we are hearing about new, expensive "miracle drugs" that initially looked quite promising from the available evidence, but then are found to either cause serious, sometimes deadly side effects, or turn out to be no more effective than simple, cheap, commonly available medicines. Sometimes it is better to do nothing for a patient. Sometimes standard practice or even archaic practice (using leeches, honey, or aspirin, e.g.,) is more effective than snazzy new drugs or procedures. And sometimes the newest is best. We need to be able to clearly, empirically demonstrate the best, most defensible way to do things (best practices) and motivate practicing researchers to use them. Our goal should be to leverage our skill at quantitative methodology to study our own tools; what techniques give us the best, most replicable, most powerful, least error-prone outcomes, and under what conditions? And what do researchers need to do to make sure their analyses turn out as well as possible? We, as a field, need to move beyond turf wars, opinion, petty careerism, and evangelism to an evidence-based body of knowledge that researchers in other areas of the discipline can use to improve the odds that their work will have the best possible outcome. We need to allow certain archaic or sub-optimal techniques to sunset, retaining and promoting best practices, whether they are new or a century old.

It is my hope that this journal can help us move toward just such an evidence-based, clearly articulated future, and I hope you will join the efforts of this tremendously talented, diverse, international editorial board to make it happen. I believe that we will be able to meet these challenges, leverage these technologies, and leave a legacy of excellence for future generations of scholars to follow.

Yet we cannot forget that the path to the unknowable future is rarely clear and easy. Our field has seen an unprecedented contraction in recent years. Quantitative training needs are expanding exponentially, yet doctoral programs in quantitative research methods (and students interested in specializing in those methods) are declining in numbers. For example, the American Psychological Association's Task Force on Quantitative Psychology reported just 23 Quantitative Psychology doctoral programs in North America, each with a handful (or fewer) faculty, and many with unused capacity to train more students than they had qualified applicants. At a time when we have the power to leverage tremendous amounts of data to answer important questions, why does there seem to be a lack of interest in specializing in this discipline? Is it possible that because our tools are so easy to use, with point-and-click interfaces, that there is now a perception that students do not need as much training in quantitative methods? Of course, the reverse is true. The more sophisticated the software has become, the more training quantitative researchers need to make informed choices about what they are doing and ensure appropriate interpretation of the results. Our challenge is to maintain a dialogue with our students and colleagues about the ever-increasing need for methodological training, and to define what training is necessary and sufficient for a scholar in the 21st century.

I wonder if the lack of interest in Quantitative training has to do with a very real lack of diversity in the field. At least within North America, the vast majority of faculty in quantitative methods are Caucasian males, and almost two-thirds of students in these programs are Caucasian as well. Do we have a diversity issue in the field? If so, how do we address it? The APA Task Force notes that Quantitative Psychology lags behind the sciences and engineering in diversity. Our editorial board is one of the most diverse I have seen, which is a tremendous asset. I challenge us (and our colleagues reading this) to constructively examine and address this apparent gap in our field in some meaningful, scholarly way. Let Frontiers in Quantitative Psychology and Measurement be a forum not only for discussion of methods and best practices, excellence in application and debate as to epistemology, but perhaps as important, scholarship and debate around the training of quantitative psychologists, statisticians, psychometricians, and researchers in the social sciences. Our field needs a forum to explore important trends, discuss troubling issues, and investigate possible solutions. If our field continues down this path, all social science will suffer.

As our field has developed increasingly sophisticated and interesting options for analysis of data, we become increasingly at risk for making errors of inference if we stop attending to basic issues such as data quality. Our software is now seductive in that we can immediately begin clicking and analyzing data without realizing that our results might be substantially biased or invalidated by poor data quality. As point of reference, one of my recent publications pointed out that in top educational psychology journals,

almost no authors reported testing assumptions or data quality in their articles. This troubles me, and I hope it troubles you. We must continue to motivate researchers to attend to basics before moving to the fun, advanced analytic techniques available to us. But it also points out a larger issue- software has become increasingly complex and sophisticated in many ways. One challenge I would like us to meet is to create a series of articles that guide readers on best practices in using particular software packages. I have been working to build bridges between FQPM and communities that specialize in using statistical software, and I hope that in the near future we will see this journal become a repository of specialized information on how to get the most of the incredibly rich software we have access to.

In this journal you will probably find concepts foreign to you, and probably some things you don't agree with. That's exactly my goal. The world doesn't need another journal promulgating 20th century thinking, genuflecting at the altar of $p < 0.05$. I challenge us to challenge tradition. Shrug off the shackles of 20th century methodology and thinking, and the next time you sit down to examine your hard-earned data, challenge yourself to implement *one new methodology* that represents a best practice. Use Rasch measurement or IRT rather than averaging items to form scale scores. Calculate $p_{(rep)}$ in addition to power and p. Use HLM to study change over time, or use propensity scores to create more sound comparison groups. Use meta-analysis to leverage the findings of dozens of studies rather than merely adding one more to the literature. Choose just one best practice, and use it. And each time afterward, add one more.

There it is. The gauntlet has been cast down. Do you pick it up, accepting my challenge? I and the board of editors look forward to reading your articles!

REFERENCES

Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh, Oliver & Boyd.

Haller, S., and Bartsch, A. J. (2009). Pitfalls in fMRI. *Eur. Radiol.* 19, 2689–2706.

Hemmelmann, D., Leistritz, L., Witte, H., and Galicki, M. (2009). Identification of neural activity based on fMRI data: a simulation study. *J. Physiol. Paris* 103, 353–360.

Killeen, P. R. (2008). Replication Statistics. In Best Practices in Quantitative Methods, J. W. Osborne, ed. (Thousand Oaks, CA, Sage), pp. 103–124.

Neyman, J., and Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses. *Stat. Res. Mem.*, 1, 1–37.

Osborne, J. W. (2008). Best Practices in Quantitative Methods. Thousand Oaks, CA, Sage Publishing.

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

Wang, T.-T., Mo, L., and Shu, S.-Y. (2009). The brain mechanism of memory encoding and retrieval: A review on the fMRI studies. *Sheng Li Xue Bao*, 61, 395–403.

Yuanqing, L., Namburi, P., Zhuliang, Y., Cuntai, G., Jianfeng, F., and Zhenghui, G. (2009). Voxel selection in fMRI data analysis based on sparse representation. *IEEE Trans. Biomed. Eng.*, 56, 2439–2451.