CONCEPTUAL ANALYSIS
published: 16 July 2020
doi: 10.3389/fpsyg.2020.01590

Reproducibility in Cognitive Hearing Research: Theoretical Considerations and Their Practical Application in Multi-Lab Studies

Antje Heinrich[1]* and *Sarah Knight*[2]

[1] Manchester Centre for Audiology and Deafness, University of Manchester, Manchester, United Kingdom, [2] Department of Psychology, University of York, York, United Kingdom

OPEN ACCESS

Edited by:
Claude Alain,
Rotman Research Institute (RRI),
Canada

Reviewed by:
Marco Pitteri,
University of Verona, Italy
Phillip Evan Gander,
The University of Iowa, United States

***Correspondence:**
Antje Heinrich
antje.heinrich@manchester.ac.uk

Specialty section:
This article was submitted to
Consciousness Research,
a section of the journal
Frontiers in Psychology

Received: 25 September 2019
Accepted: 15 June 2020
Published: 16 July 2020

Citation:
Heinrich A and Knight S (2020)
Reproducibility in Cognitive Hearing
Research: Theoretical Considerations
and Their Practical Application
in Multi-Lab Studies.
Front. Psychol. 11:1590.
doi: 10.3389/fpsyg.2020.01590

In this article, we consider the issue of reproducibility within the field of cognitive hearing science. First, we examine how retest reliability can provide useful information for the generality of results and intervention effectiveness. Second, we provide an overview of retest reliability coefficients within three areas of cognitive hearing science (cognition, speech perception, and self-reported measures of communication) and show how the reporting of these coefficients differs between fields. We argue that practices surrounding the provision of retest coefficients are currently most rigorous in clinical assessment and that basic science research would benefit from adopting similar standards. Finally, based on a distinction between direct replications (which aim to keep materials as close to the original study as possible) and conceptual replications (which test the same purported mechanism using different materials), we discuss new initiatives which address the need for both. Using the example of the auditory Stroop task, we provide practical illustrations of how these theoretical issues can be addressed within the context of a multi-lab replication study. By illustrating how theoretical concepts can be put into practice in empirical research, we hope to encourage others to set up and participate in a wide variety of reproducibility-related studies.

Keywords: reproducibility, replication, retest reliability, multi-laboratory collaboration, cognitive hearing science, Stroop

INTRODUCTION

Reproducibility is a core requirement for the accrual of scientific knowledge and the advancement of a field. The concept derives its importance from the fact that in an ideal world we would expect repeated measurement of the same variable to lead to the same result. However, this is often not the case, and it is unclear whether divergent results occur because measurement conditions were not comparable in some essential aspect or because of random error. Differentiating between these two potential explanations has important implications for test selection, experimental setup, and interpretation of results from single studies, as well as for the assessment of scientific progress as a whole.

We acknowledge that a lively discussion exists within the wider field of behavioral science regarding the theoretical question of what exactly constitutes "reproducibility" and how best to assess it (Schmidt, 2009; Goodman et al., 2016). We would like to extend this discussion to the field of cognitive hearing science and suggest ways of incorporating relevant theoretical concepts into empirical practice. Cognitive hearing science is an interdisciplinary field that aims to understand how auditory and cognitive processes combine to allow speech understanding in complex listening environments (Arlinger et al., 2009). Before examining cognitive hearing science specifically, however, we first propose that the following general distinctions are crucial when discussing replication: the *level* of replication and the *type* of replication. In terms of level, two levels of replication exist: individual-level replications, which concern the reproducibility of individual differences, often in the form of retest reliability; and group-level replications, which concern the reproducibility of effect sizes, often expressed as differences between group means. The former refers to the similarity of an individual's test scores at different points in time when no intervention has been applied. The latter refers to the likelihood that an experimental effect based on group-level differences will replicate. In terms of type, two types of replication exist: direct and conceptual. This distinction follows theoretical analyses by Hendrick (1991) and Schmidt (2009) who, among other things, identified the following two important classes of variables that shape the nature of an experiment: the *primary information focus* and the *contextual background*. The primary information focus of an experiment refers to both the hypothesis to be tested (the *immaterial* focus) and also the instructions, materials, and events experienced by participants (the *material realization*). The contextual background into which the primary information focus is embedded includes participant characteristics, the physical setting, the particular experimenter, minute material differences (called *specific task variables*; e.g., different screen resolutions), and also the procedures for the selection and allocation of participants [which Schmidt (2009) defines as a separate class]. Based on this framework, we define *direct replication* studies as those which aim to keep the material realization of the primary information focus as close to the original study as possible. These studies vary contextual background in an effort to discount sampling error or artifacts as explanations for published findings. *Conceptual replications*, on the other hand, aim to verify the underlying hypotheses of previous studies by constructing an experiment which tests the same purported mechanisms but uses different materials. Thus, conceptual replications address the same immaterial primary information focus, but vary its material realization. Often, type and level of replication vary together. For instance, retest reliability measures tend to involve individual-level scores from direct replications. Conceptual replications, on the other hand, typically involve the replicability of group means, but may also investigate individual-level scores.

It is important to note that these distinctions are not entirely unproblematic. In the case of replication type (direct versus conceptual), the distinction is a matter of degree and centers around questions such as: Up to what point is a replication still direct? When does it become conceptual? Which changes in material, setup or test population are consequential? In the case of replication level (individual versus group), the distinction is conceptually and mathematically clear – but these different types of replication have directly opposing requirements for sample selection and experimental set-up. Experimental studies are based on average responses, which require homogeneous samples in order to reduce unwanted between-subject variability and increase the chance that even small experimental effects will reach a significance threshold. By contrast, correlational studies, which examine phenomena based on individual differences within a population, require large between-subject variability to be most sensitive. Indeed, low between-subject variability in this type of study adversely affects the reliability of individual differences, decreasing the likelihood of replicability for results based on correlations with other factors (Hedge et al., 2018). In other words, a fundamental methodological tension exists between two commonly-used research designs. Being aware of this tension is particularly important when both approaches are combined within a single study, a practice increasingly common in cognitive hearing science (Lunner and Sundewall-Thorén, 2007; Heinrich et al., 2010; Schneider et al., 2016).

In this article, we discuss reproducibility in the context of three types of measures commonly used within cognitive hearing science: behavioral measures of cognition, behavioral measures of speech perception, and self-report measures of communication ability and speech perception. In the following section, we first discuss individual-level replications before turning to group-level replications.

REPLICATION LEVEL: INDIVIDUAL DIFFERENCES – THE CASE OF RETEST RELIABILITY

Replication of individual differences concerns the stability of a score over time – that is, how likely it is that an individual's score will replicate when the individual has not undergone any intervention. Typically, such replications are carried out as direct replications (i.e., the same test is performed repeatedly) and the resulting measure of consistency is termed "retest reliability." Retest reliability provides information about the robustness and precision of a test on the level of individual scores. Such information is crucial in experimental studies because the predictive value of variables is limited by their precision, as measured by reliability (Spearman, 1904; Nunnally, 1970). In clinical diagnosis and prognosis, retest reliability is a prerequisite for accurate assessment and the monitoring of interventions over time. Despite the importance of retest reliability, it is rarely measured and reported, a situation that is not unique to cognitive hearing science (Watson, 2004).

One important question in retest reliability is how to adequately account for measurement error. Error can be systematic – arising in a relatively predictable fashion from sources such as practice effects, fatigue, or use of specific raters – or it may be random. Furthermore, strategies to control for systematic and random error can be employed on a design level and/or at the analysis stage. For example, one common design-level strategy to avoid systematic error arising from memorization of stimulus materials is to use similar but non-identical stimulus lists when tests are re-administered. Strategies at the analysis stage, meanwhile, usually center around the choice of statistical measure. Retest reliability of interval-scaled and normally distributed data is typically assessed using either of the following two parameters: Pearson's Product-Moment Correlation (PPMC) or the Intraclass Correlation Coefficient (ICC). The PPMC has historically been a popular parameter of retest reliability despite its problems with accounting for systematic error and bias, and its overestimation of reliability (Heise, 1969). The ICC, on the other hand, explicitly estimates both systematic and random error, thereby allowing for a distinction between estimates of agreement and consistency (Aldridge et al., 2017). In this context, agreement refers to the extent to which observed raw scores obtained by a measurement tool for one individual match between raters or time-points in the absence of any actual (systematic) change in the outcome being measured. In contrast, consistency refers to how similar the relative rank of an individual's score within the group is across raters/times; the actual value of the raw scores itself is unimportant. In addition to estimating retest reliability, the ICC allows the estimation of minimal difference scores. Minimal difference scores are vital in the context of intervention effectiveness, since they indicate the smallest difference that can be considered a significant change (Jacobson and Truax, 1991; Chelune et al., 1993) rather than arising from measurement error. Regardless of the strategy used to control for error, calculated values for retest reliability are typically judged according to the guidance given by Cicchetti and Sparrow (1981): retest reliability below 0.40 is poor, between 0.40 and 0.59 is fair, between 0.60 and 0.74 is good, and 0.75 and above is excellent.

Supplementary Table S1 gives example values of retest reliability scores for a number of tests used in cognitive hearing science. In particular, we focus on the three types of measure mentioned above: behavioral measures of cognition[1], behavioral measures of speech perception, and self-report measures of communication ability and speech perception. The table gives the name of the test, the number of participants and make-up of the sample on which the retest reliability value is based, the time period between administrations, which type of retest coefficient was used, its value, and whether any systematic differences in means were reported. Providing this information enables the reader to assess for themselves whether systematic error has occurred and how well it was taken into account for a particular estimate.

Note that behavioral tests used for clinical assessment are typically more rigorously assessed for retest reliability than tests used in lab-based research, and speech testing is no exception in this regard. Some clinical speech tests have undergone extensive validation in order to construct equivalent but non-identical forms or stimulus lists, such as the CUNY NTS (Dubno and Dirks, 1982), NU-6 word test (Causey et al., 1983), and the QuickSIN (Killion et al., 2004). Speech tests developed exclusively for research purposes, on the other hand, normally undergo less stringent validation, although there are notable exceptions such as the SPIN-R (Bilger et al., 1984). In fact, there are countless examples, including from our own research, where speech material was newly developed and used to investigate differences between groups of interest without first rigorously testing the accuracy of the material as an outcome measure (Heinrich et al., 2008, 2010; Heinrich and Schneider, 2011; Knight and Heinrich, 2017, 2019).We suggest that lab-based science research would benefit from aspiring to similar standards to clinical assessment when it comes to the measurement and reporting of retest reliability and test validity.

A Group-Level View of Individual Differences

In addition to examining retest reliability estimates from single studies, it is possible to examine how well retest reliability estimates themselves replicate across different studies – in other words, to take a group-level view of individual-level retest reliability. To give a sense of the insights gained from such an approach, we re-print in **Supplementary Table S2** a subset of information from **Supplementary Table S1** and provide more detailed descriptions of the studies involved, including sample composition, test administration, and statistical details. Note that retest reliability estimates often vary by about 0.2. For some tests (Letter Number Sequencing test, semantic fluency), it appears unclear which, if any, of the methodological differences caused this disparity. For the Digit Span test, it is possible that the difference in retest reliability values was caused by the varying retest intervals (days versus a year) but it may also be due to other unreported differences between the studies. For phonological fluency, it is troubling to note that, although the participant groups had similar characteristics and the same coefficient was used to estimate retest reliability (PPMC), the estimates still varied between 0.63 and 0.82. An additional concern for the phonological fluency estimates is that the one study which examined differences between the first and second tests found systematic differences. These were then not adequately taken into account, thus possibly leading to an overestimate of retest reliability. As at least one other study also used PPMC estimates without testing for systematic differences, it is possible that other values are overestimated as well (Heise, 1969). Finally, the Trail Making A&B tests were the most frequently replicated. The results from these studies suggest that values around 0.6 may be more representative of this test's retest reliability in many situations than 0.8 or 0.9.

Methodological variations in testing are likely to explain some of the differences found for retest reliability estimates. However, how much they explain and how much is due to error remains to be established by systematic investigation. There is a clear need for better validation of experimental measures, resulting in more reliable and comparable tests. Such a shift in practices would represent one means of tackling the replication crisis currently facing psychological science in general

[1] We only include here cognitive tests that have previously been shown to have a link with speech perception and/or communication.

(Open Science Collaboration, 2015) and most likely cognitive hearing science too. Given that it is not always clear how robust scores are in various populations when they are repeatedly assessed, either using identical tests or comparable but non-identical stimulus lists, it can be difficult to know what results mean and whether interventions and manipulations have had the intended effect. We therefore advocate for retest scores of identical and comparable stimulus lists to be routinely included as standard measurements. This would enable researchers to assess robustness of scores and list equivalence more easily, make more informed choices regarding outcomes measures, and also encourage methods for improving robustness, particularly of non-identical lists.

REPLICATION LEVEL: GROUP EFFECTS

Besides individual differences, replication of group differences is another essential aspect of scientific practice. Sometimes, one or two key conditions from a previous study are included in a new study in order to verify the premise of the basic effect (Studdert-Kennedy and Shankweiler, 1970; Cutting, 1974; Amitay et al., 2002; Ziegler et al., 2005)—although these replications do not always yield the intended result (Baker et al., 2008; Arsenault and Buchsbaum, 2016). Publishing complete direct replications (and their failure) has been a longstanding problem, since publication guidelines of scientific journals have traditionally stated that the scope of their publications is innovative, new, or original research. This almost exclusive focus on novelty as practiced by many scientific journals in the past has given rise to a number of concerns. In particular, due to publication and other biases (Ioannidis, 2005; Ioannidis et al., 2014), only positive results tended to be published (Scheel et al., 2020). Such a practice historically made it difficult to explore whether replication failure was due to inadvertent consequential changes in the paradigm or to random error. Additionally, the replication of previous work typically represents a minor focus of a given publication, making it difficult to track the state of replications in a field (see also Rosenthal, 1979).

However, this practice is in the process of changing, with more journals now stating that they value the internal (i.e., within study) and external (i.e., across study) replication of results (e.g., *Journal of Psychology: General; Psychological Science; Royal Society*). The recent change in approach to replication in psychological research can be illustrated by the publication of two large-scale replication projects: the Reproducibility Project undertaken by the Open Science Collaboration (Open Science Collaboration, 2012, 2015) and the Many Labs projects (Klein et al., 2014, 2018). In addition to being the first large-scale replication projects in psychology, they also illustrate the different approaches that can be taken to multi-site replication work. The OSC project is an example of a "broad-and-shallow" approach to direct replications, in which single replications of many different findings were carried out, each at a different site. The OSC conducted replications of 100 experimental and correlational studies from cognitive and social psychology. They reported that effect sizes were approximately half the magnitude of the original effects, and only 37% of replications showed significant results (Open Science Collaboration, 2015). In contrast, the two Many Labs projects are an example of a "narrow-and-deep" approach, in which the authors seek to replicate the same small group of findings across a number of sites with some variation, mainly in testing population. The findings varied across the two Many Labs projects, but in both cases the authors concluded that replicability depended more upon the effect being studied than the sample or setting used to study it (Ebersole et al., 2016; Klein et al., 2018). Besides these two recent efforts, it is also worth noting that the idea of multi-site replication is now starting to become embedded in undergraduate education, for example via the establishment of the GW4 Undergraduate Psychology Consortium in the United Kingdom (Button et al., 2019).

As described above, we define *direct replication* studies as those which aim to keep the material realization of the primary information focus as close to the original study as possible and *conceptual replications* as those, which test the same purported mechanisms with different materials. Both the Reproducibility Project and the Many Labs projects are direct replication studies, which closely reproduce the material realization while varying the contextual background. Indeed, the replication protocols were developed whenever possible in collaboration with the original authors, even including the use of original materials. However, direct replications are not a panacea. Among other things, simply reproducing methodologies without considering theoretical underpinnings [what Phaf (2020) calls "mechanical" replications] runs the risk of perpetuating, rather than unearthing, problems. As Gelman and Carlin (2014) explain, "Consistent findings could take on the status of confirmed truths, when they actually reflect failings in study design, methods or analytical tools." (p400).

A number of suggestions have been advanced to improve the quality of replications. Phaf (2020) suggested that experimental work should always be complemented by thorough theoretical analyses. In the case of unsuccessful replications, this would allow for the discovery of potentially crucial (and as yet unexamined) factors that may explain the result. A second related suggestion is to formulate competing theoretical hypotheses that focus on the disproof and exclusion of alternative explanations rather than the traditional presence or absence of a statistical effect (for detailed discussions see Platt, 1964; Phaf, 2020). Adopting this approach in the field of cognitive hearing science would minimize the existence of null results and replication failures. Such a change in hypothesis generation would, however, necessitate development of and closer engagement with underlying theoretical concepts. A third approach focuses on paying closer attention to the types of errors that occur as part of incorrect statistical inferences and effect size estimation. Two types of errors are often differentiated: errors of magnitude (in which the effect size is exaggerated) and errors of direction. Both types of errors can be surprisingly high for underpowered studies, even when the statistical results are significant (Gelman and Carlin, 2014). In the context of our discussion, this means that if studies are underpowered, their results may not only be non-significant (thereby leading to replication failure) but may also give rise to effects in the unexpected direction. In a study with theoretically-motivated hypotheses, such a misdirected effect would likely be discounted regardless of its significance. However, in a study that only predicts an effect of a variable without specifying its direction (common in regression-type analyses of individual differences), it is much harder to identify errors of direction. For a detailed discussion of the probabilities for these types of errors see Kirby and Sonderegger (2018). Finally, some researchers advocate the adoption of "big data" and machine learning to enhance reproducibility in psychological research – approaches which, among other things, involve very large sample sizes (Yarkoni and Westfall, 2017). This may, the authors suggest, involve using existing datasets or corpora, or it may involve "large, multilab, collaborative projects" (p. 1110), such as the OSC and Many Labs projects – a point to which we return below. Of course, not all researchers will wish or be able to involve machine learning in their work; however, regardless of whether or not one takes an AI-based approach, it is clear that increased sample sizes are vital in order to avoid errors resulting from underpowering or overfitting to local noise. Similarly, clearly defined and pre-determined stopping rules for data collection must be implemented to reduce the prevalence of false-positive results (Simmons et al., 2011). Recent developments in the area of stopping rules have shown them to have important implications in both frequentist and Bayesian hypothesis testing (Rouder, 2014; Sanborn and Hills, 2014).

Conceptual replications, meanwhile, have also been subject to criticism. For example, Pashler and Harris (2012) argue that results from direct replications always have the power to advance the field: successes strengthen the trust in the phenomenon, while failures will slowly erode it. However, while successful conceptual replications provide new information by extending the reach of the phenomenon, failures of conceptual replications will not necessarily erode the trust in a phenomenon and thus not provide useful information. Failures will only be interpreted as showing that the material realization was not close enough to the original study. Such an interpretation cannot exclude the possibility that the phenomenon itself (with the same material realization) may not have been replicable in the first place. In this sense, conceptual replications may have less information value than direct replications. Arguing along similar lines, Nosek and Errington (2020) suggest that many conceptual replications are in practice actually generalizability tests, in which failures "are interpreted, at most, as identifying boundary conditions" (p. 5).

Nevertheless, we argue that – regardless of whether they are viewed as "replications" proper or as generalizability tests – conceptual replications have both practical and theoretical value. From a practical perspective, many attempts at replication are conceptual to some extent: materials and methods are often based only on the descriptions given in the experimental report, and these are typically underspecified (Open Science Collaboration, 2015). It is therefore important to determine whether the level of change in materials used in a replication study mean that it is a "meaningful" conceptual replication (if successful), or whether the changes are simply unavoidable but non-critical variability in the material realization of the primary information focus, thus making the study effectively a direct replication. Such conceptual replications are vital if researchers want to know which particular implementation of a given task is likely to produce the most robust effect in their participant pool, and/or which specific details of a set-up are vital and which can be safely varied or omitted.

From a theoretical perspective, conceptual replications are important because they can add further support to the original hypotheses and/or proposed mechanisms underlying a particular effect; indeed, by identifying boundary conditions in terms of experimental protocol, they can actually help to clarify and refine the original interpretation and explanation of an effect. In order for conceptual replications to provide all of this information, they need to be carried out in a systematic and incremental fashion, altering only one aspect of a single class of variable at a time. Unfortunately, as Schmidt (2009) observes, conceptual replications are relatively unpopular with reviewers and editors, and as a result, the process of conceptual replication is often not explicit – and therefore somewhat haphazard.

However, conceptual replications do present a theoretical complication. A pure conceptual replication should vary the material realization of the primary information focus, not the contextual background; therefore, strictly speaking, they should be carried out using an identical participant sample to the original study (Schmidt, 2009). In reality, of course, this is not possible or practical. In order to carry out conceptual replications in as meaningful a way as possible, one should therefore perform them over a large enough sample and variety of sites to demonstrate *both* robustness of the concept itself *and* its replicability over multiple contexts. Such a large-scale study would both (i) function as a conceptual replication that explores in a controlled and systematic fashion the necessary and sufficient material conditions required for an effect to emerge and reveals meaningful boundary conditions and also (ii) use large enough sample sizes to be able to discount sampling error, artifacts and lack of power as explanations for the effects. One way to address these issues is to run a series of systematic multi-lab conceptual replications [along the lines of the large, collaborative projects advocated by Yarkoni and Westfall (2017) see above]. In the following section, we present one example of how such an approach might work in practice, focusing on a test commonly used to assess inhibition – the Stroop task.

DIRECT AND CONCEPTUAL REPLICATIONS OF GROUP EFFECTS AND INDIVIDUAL DIFFERENCES IN THE CONTEXT OF STROOP TASKS

Stroop tasks are widely used to assess inhibition – the ability to suppress goal-irrelevant information (Stroop, 1935; MacLeod, 1991). In its classic form, the Stroop task assesses inhibition in the visual domain via color-word interference. Participants are required to name the ink color of a string of characters while ignoring the characters themselves. In the neutral condition, these characters are meaningless or irrelevant; in the incongruent condition, they spell out a conflicting color word (e.g., BLUE printed in red). The difference in reaction times between the incongruent and neutral conditions is typically taken as a measure of inhibitory ability and termed Stroop interference (SI).

The visual Stroop task is an example of a task with a rich conceptual replication history, particularly as concerns the testing materials. For example, some studies enhance the visibility of the color by replacing font color with a larger patch of color underneath a superimposed word (Janse, 2012; Knight and Heinrich, 2017). For the control condition, some studies use a string of Xs as their irrelevant characters, while others use unrelated words or even simply blank patches of color (MacLeod, 1991). In the incongruent condition, meanwhile, some studies have used only the first letters of the incongruent color words (such as "R" instead of "RED"; Regan, 1978). Such conceptual replications have been shown to vary the size of the interference effect, but as a general rule such modifications "only modestly affect its magnitude, not its qualitative form" (MacLeod, 1991, p166). Indeed, even with substantial changes to experimental protocol, Stroop-type tasks still produce an interference effect; such changes include – to name just a few – spatial separation of color patches and words, using different response modalities (oral vs. manual), using color-related (as opposed to actual color) words in the incongruent condition (e.g., lemon and sky), and asking participants to sort stimuli into categories rather than simply naming or otherwise responding to their basic properties.

This rich and robust replication history stands in contrast to auditory versions of the Stroop task. Although such versions have been successfully used (Green and Barber, 1981; Morgan and Brandt, 1989), their replication appears to be less successful if we take as an indication the rarity of published studies reporting them. In auditory Stroop tasks, participants are typically required to respond to some perceptual feature of a sound while ignoring the semantic content, which – as in the visual version – can be either irrelevant or conflicting. For example, participants may be required to respond to the speaker's gender regardless of the word spoken, which in the control condition will be neutral (e.g., "cat") and in the incongruent condition will be conflicting (e.g., "woman" spoken by a man). In addition to gender, other auditory dimensions have been used including pitch ("high" vs. "low"), location ("left" vs. "right"), loudness ("loud" vs. "soft"), and even time ("fast" vs. "slow") (Hamers and Lambert, 1972; Pieters, 1981; Morgan and Brandt, 1989; Roberts and Hall, 2008; Whitton et al., 2017). As well as fewer studies reporting the use of the task, there are also direct reports of non-replication. For example, Morgan and Brandt (1989) report an auditory Stroop interference effect only in the pitch domain, but not in the time domain. Additionally, Knight and Heinrich (2017) found a modest auditory Stroop interference effect using a gender-based task only on the group level, but could not replicate this effect for every participant or indeed for every one of the four speakers used in their materials.

Auditory versions of the Stroop task are particularly attractive when the main outcome variable of interest is itself auditory – for example, speech-in-noise perception – and have been used both alone and alongside visual Stroop tasks (Sommers and Danielson, 1999; Knight and Heinrich, 2017; Whitton et al., 2017). In many cases, it is implied that the auditory Stroop task is essentially equivalent to a visual version: for example, immediately beneath the heading "Audio/Visual Stroop," Whitton et al. (2017) simply state that "The Stroop effect provides a well-established measure of inhibitory control." Here, the auditory Stroop task is being treated as a conceptual equivalent of the visual Stroop: a task which, despite the very different material realization of the primary information focus, nevertheless produces the same group-level effects (and presumably therefore taps into the same underlying mechanism) as the classic visual version.

However, in the case of the auditory Stroop task, this is in fact far from clear. In 1991, MacLeod asked "How equivalent are all of these tasks that superficially resemble the Stroop task? Even for the very prevalent alternatives [. . .] we do not know [. . .] Obviously, though, it is of theoretical importance to know whether similar processes are invoked in these many variations, but we have insufficient evidence at present." (MacLeod, 1991, p. 170). This remains true for the auditory Stroop task nearly 30 years later: although it is often assumed to tap the same underlying domain-general inhibitory ability as the visual task, the extent to which this is true is unclear. Crucially, the extent of overlap appears to depend on the exact implementation of the two tasks. For example, Roberts and Hall (2008), using extremely carefully chosen and closely matched tasks, demonstrated similar patterns of neural activation and correlated behavioral responses for Stroop tasks presented across different modalities, suggesting that visual and auditory versions do indeed tap shared inhibitory processes. Conversely, when auditory and visual Stroop tasks were less closely matched and arguably more *conceptually* similar than methodologically similar, the auditory and visual versions have not been found to correlate at all between individuals (Shilling et al., 2002; Knight and Heinrich, 2017).

In short, then, there are a number of reproducibility issues regarding the auditory Stroop task that need to be addressed. First, at the level of group effects, more conceptual replications are needed of the auditory Stroop task in isolation in order to investigate which specific material realizations (e.g., gender- vs. pitch-based tasks) produce a reliable group-level effect in the auditory domain. Second, at the level of individual differences, direct replications of individual scores (i.e., retest reliability) need to be considered. We are not aware of any studies that have provided these data for auditory Stroop tasks, and not having this information limits our understanding of the extent to which correlations between different Stroop tasks are limited by retest reliability (Hedge et al., 2018). Finally, even if measures of auditory Stroop interference are replicable across different material realizations and reliably assess behavior on an individual basis, the question remains of whether or not they assess the same underlying mechanism as the visual Stroop task. Therefore, research needs to assess whether participants' individual scores are correlated (i.e., replicate) across the two types of task in their different material realizations. Only if this is true can the visual and auditory Stroop tasks be considered *conceptually* equivalent.

Besides these theoretical considerations, there is also a strong practical aspect to such a project: if researchers know that auditory Stroop tasks do (or do not) produce similar results to their visual counterparts and are aware of which auditory

Stroop implementations produce the largest and/or most visual-like results, then they can confidently select the best type of Stroop task for their purposes. We believe that these questions could be fruitfully addressed using a many-labs-style replication project and outline in the following section how such a project could be implemented.

SETTING UP A MULTI-LAB REPLICATION FOR THE AUDITORY STROOP

Any number of implementations of the auditory Stroop task could be tested, but in the first instance it seems reasonable to attempt to replicate a small number of tasks already reported by existing studies. Such a practice would also be consistent with the traditional approach to replication: that is, selecting previously reported key results and seeking to replicate their group effects as closely as possible. In addition to the selected auditory Stroop tasks, we would suggest the inclusion of a classic color-word visual Stroop, since this is in many ways the "gold standard" version of the task (MacLeod, 1992). In terms of participants, a conceptual replication would imply the use of a sample as closely matched demographically to the original studies as possible – in this case, undergraduate students. However, a straightforward extension of the replication could see the demographic requirements for participation relaxed and the influence of demographic variables on the Stroop effect investigated in its own right. In this case, it would be necessary to collect demographic information about all participants, along with measures of visual and auditory acuity.

The selected auditory Stroop tasks, the color-word visual Stroop task and the collection of relevant demographic data would serve as a core package carried out across all sites involved in the study. Following the Hendrick (1991) and Schmidt (2009) framework, we suggest that this core package keeps the immaterial realization of the primary information focus constant while varying its material realization – in other words, it provides a conceptual replication by testing the robustness and replicability of results across different Stroop tasks. Rolling out this core package across multiple labs also varies critical aspects of the contextual background, testing replicability across different participant groups, physical settings and experimenters. This approach therefore fulfills the need, outlined above, for conducting conceptual replications over a large enough sample and variety of sites to demonstrate *both* robustness of the concept itself *and* its replicability over multiple contexts. It is nevertheless desirable to minimize those aspects of the contextual background that Schmidt refers to as *specific task variables*: minor variations in materials such as paper color, headphone type, screen resolution and so forth. To minimize these effects in our core package, we suggest using the same stimuli across all labs involved in the project, and using shared calibration procedures and close collaboration during task set-up. Online repositories for sharing materials – such as that hosted by the Open Science Framework[2] –are of great help in this regard: participating labs can easily and remotely access not only stimuli, but also details of calibration procedures and code for running the tasks, thus ensuring that set-up, instructions, and procedure are as close as possible across the different participating sites.

Besides the closely prescribed core package, which would be relatively brief, participating laboratories could also have flexibility in adding their own tasks and collecting supplementary, single-lab datasets relevant to their needs and interests. One key addition, as discussed above, would be to ask participants to perform the same tasks multiple times to assess retest reliability. Individual labs may also be interested in running the tasks on different listener groups or adding additional tasks to explore the relationship of Stroop scores to other measures. A further extension could be a comparison of data collected online with that collected in the laboratory. The growing popularity of online recruitment and/or testing platforms such as Gorilla[3], Prolific[4], and Amazon's Mechanical Turk[5] has opened up possibilities for collecting data from a much broader range of participants than those typically involved in laboratory studies[6]. Another line of extensions could explore the limits of replicability by changing aspects of the set-up in ways that theoretical analyses suggest alter the task in a conceptually meaningful manner. Such replications (or replication failures) would help delineate the extent to which generalizations can reasonably be made [see the suggestions of Phaf (2020) and others, discussed above].

CONCLUSION

In this article, we have discussed theoretical and practical aspects of the reproducibility crisis in science and how they might be tackled. In particular, we have suggested that one way to improve reproducibility, particularly when assessing individual differences, is to encourage researchers to include retest reliability measures of their quantitative assessment methods as a routine aspect of testing, analysis, and reporting. The gradual collection of retest coefficients of commonly-used tests in a variety of situations would allow researchers to better judge the reliability of tests, which in turn should influence both the planning stage of studies as well as the interpretation of results. We have also advocated for both direct replications – those which address the contextual background of a task while preserving the material realization of the primary information focus as far as possible – and also conceptual replications. In particular, we have focused on the benefits of large-scale systematic conceptual replications – that is, systematically varying the material realization of the primary information focus while

[2] https://osf.io/

[3] www.gorilla.sc

[4] www.prolific.ac

[5] www.mturk.com

[6] Early indications are that data quality and reliability is high for online studies (Casler et al., 2013; Gould et al., 2015) but disparities between lab and online samples do emerge for some tasks (Crump et al., 2013). As a result, a new facet of replicability has been added to the contextual background of a study: how well results replicate across lab and online cohorts. Both visual and auditory Stroop tasks can be set up in such a way that the same task can be run online and in the lab, thus allowing this additional aspect of reproducibility to be assessed.

nevertheless collecting large enough, multi-site sample sizes to account for contextual variation. Such replications can only be achieved through close collaboration on multi-lab projects.

AUTHOR CONTRIBUTIONS

AH and SK wrote the manuscript with oversight and conceptual guidance from AH. AH also produced the final structure of the article. Both authors contributed to the article and approved the submitted version.

FUNDING

AH was supported by the NIHR Manchester Biomedical Research Centre.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01590/full#supplementary-material

REFERENCES

Aldridge, V. K., Dovey, T. M., and Wade, A. (2017). Assessing test-retest reliability of psychological measures: persistent methodological problems. *Eur. Psychol.* 22, 207–218. doi: 10.1027/1016-9040/a000298

Amitay, S., Ben-Yehudah, G., Banai, K., and Ahissar, M. (2002). Disabled readers suffer from visual and auditory impairments but not from a specific magnocellular deficit. *Brain* 125, 2272–2285. doi: 10.1093/brain/awf231

Arlinger, S., Lunner, T., Lyxell, B., and Pichora-Fuller, M. K. (2009). The emergence of cognitive hearing science. *Scand. J. Psychol.* 50, 371–384. doi: 10.1111/j.1467-9450.2009.00753.x

Arsenault, J. S., and Buchsbaum, B. R. (2016). No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. *Psychon. Bull. Rev.* 23, 1231–1240. doi: 10.3758/s13423-015-0988-z

Baker, R. J., Jayewardene, D., Sayle, C., and Saeed, S. (2008). Failure to find asymmetry in auditory gap detection. *Laterality* 13, 1–21. doi: 10.1080/13576500701507861

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., and Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *J. Speech Hear. Res.* 27, 32–48. doi: 10.1044/jshr.2701.32

Button, K. S., Chambers, C. D., Lawrence, N., and Munafò, M. R. (2019). Grassroots training for reproducible science: a consortium-based approach to the empirical dissertation. *Psychol. Learn. Teach.* 19, 77–90. doi: 10.1177/1475725719857659

Casler, K., Bickel, L., and Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav.* 29, 2156–2160. doi: 10.1016/j.chb.2013.05.009

Causey, G. D., Hermanson, C. L., Hood, L. J., and Bowling, L. S. (1983). A comparative evaluation of the Maryland NU 6 auditory test. *J. Speech Hear. Disord.* 48, 62–69. doi: 10.1044/jshd.4801.62

Chelune, G. J., Naugle, R. I., Lüders, H. O., Sedlak, J. M., and Awad, I. A. (1993). Individual change after epilepsy surgery: practice effects and base-rate information. *Neuropsychology* 7, 41–52. doi: 10.1037/0894-4105.7.1.41

Cicchetti, D. V., and Sparrow, S. S. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am. J. Ment. Defi.* 86, 127–137.

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's mechanical turk as a tool for experimental behavioral research. *PLoS One* 8:e57410. doi: 10.1371/journal.pone.0057410

Cutting, J. E. (1974). Two left-hemisphere mechanisms in speech perception. *Percept. Psychophys.* 16, 601–612. doi: 10.3758/bf03198592

Dubno, J. R., and Dirks, D. D. (1982). Evaluation of hearing-impaired listeners using a nonsense-syllable test. I. Test reliability. *J. Speech Hear. Res.* 25, 135–141. doi: 10.1044/jshr.2501.135

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., et al. (2016). Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* 67, 68–82. doi: 10.1016/j.jesp.2015.10.012

Gelman, A., and Carlin, J. (2014). Beyond power calculation: assessing type S (sign) andtype M (magnitude) errors. Perspectives on. *Psychol. Sci.* 9, 641–651. doi: 10.1177/1745691614551642

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Sci. Transl. Med.* 8:341s12. doi: 10.1126/scitranslmed.aaf5027

Gould, S. J. J., Cox, A. L., Brumby, D. P., and Wiseman, S. (2015). Home is where the lab is: a comparison of online and lab data from a time-sensitive study of interruption. *Hum. Comput.* 2, 45–67. doi: 10.15346/hc.v2i1.4

Green, E. J., and Barber, P. J. (1981). An auditory stroop effect with judgments of speaker gender. *Percept. Psychophys.* 30, 459–466. doi: 10.3758/BF03204842

Hamers, J. F., and Lambert, W. E. (1972). Bilingual interdependencies in auditory perception. *J. Verbal Learn. Verbal Behav.* 11, 303–310. doi: 10.1016/S0022-5371(72)80091-4

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186. doi: 10.3758/s13428-017-0935-1

Heinrich, A., Flory, Y., and Hawkins, S. (2010). Influence of English r-resonances on intelligibility of speech in noise for native English and German listeners. *Speech Commun.* 52, 1038–1055. doi: 10.1016/j.specom.2010.09.009

Heinrich, A., and Schneider, B. A. (2011). Elucidating the effects of ageing on remembering perceptually distorted word pairs. *Q. J. Exp. Psychol.* 64, 186–205. doi: 10.1080/17470218.2010.492621

Heinrich, A., Schneider, B. A., and Craik, F. I. M. (2008). Investigating the influence of continuous babble on auditory short-term memory performance. *Q. J. Exp. Psychol.* 61, 735–751. doi: 10.1080/17470210701402372

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *Am. Sociol. Rev.* 34, 93–101.

Hendrick, C. (1991). "Replication, strict replications, and conceptual replications: are they important?," in *Replication Research in the Social Sciences*, ed. J. W. Neuliep (Newbury Park: Sage), 41–49.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn. Sci.* 18, 235–241. doi: 10.1016/j.tics.2014.02.010

Jacobson, N. S., and Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J. Consult. Clin. Psychol.* 59, 12–19. doi: 10.1037/0022-006x.59.1.12

Janse, E. (2012). A non-auditory measure of interference predicts distraction by competing speech in older adults. *Neuropsychol. Dev. Cogn. Section BAgingNeuropsychol. Cogn.* 19, 741–758. doi: 10.1080/13825585.2011.652590

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., and Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 116, 2395–2405. doi: 10.1121/1.1784440

Kirby, J., and Sonderegger, M. (2018). Mixed-effects design analysis for experimental phonetics. *J. Phonet.* 70, 70–85. doi: 10.1016/j.wocn.2018.05.005

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., and Bahník, S. (2014). Data from investigating variation in replicability: a "Many Labs" replication project. *J. Open Psychol. Data* 2:e4. doi: 10.5334/jopd.ad

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., et al. (2018). Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* 1, 443–490. doi: 10.1177/2515245918810225

Knight, S., and Heinrich, A. (2017). Different measures of auditory and visual stroop interference and their relationship to speech intelligibility in noise. *Front. Psychol.* 8:230. doi: 10.3389/fpsyg.2017.00230

Knight, S., and Heinrich, A. (2019). Visual inhibition measures predict speech-in-noise perception only in people with low levels of education. *Front. Psychol.* 9:2779. doi: 10.3389/fpsyg.2018.02779

Lunner, T., and Sundewall-Thorén, E. (2007). Interactions between cognition, compression, and listening conditions: effects on speech-in-noise performance in a two-channel hearing aid. *J. Am. Acad. Audiol.* 18, 604–617. doi: 10.3766/jaaa.18.7.7

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychol. Bull.* 109, 163–203. doi: 10.1037/0033-2909.109.2.163

MacLeod, C. M. (1992). The Stroop task: the "gold standard" of attentional measures. *J. Exp. Psychol. Gen.* 121, 12–14. doi: 10.1037/0096-3445.121.1.12

Morgan, A. L., and Brandt, J. F. (1989). An auditory Stroop effect for pitch, loudness, and time. *Brain Lang.* 36, 592–603. doi: 10.1016/0093-934X(89)90088-6

Nosek, B. A., and Errington, T. M. (2020). What is replication? *PLoS Biol.* 18:e3000691. doi: 10.1371/journal.pbio.3000691

Nunnally, J. C. (1970). *Introduction to Psychological Measurement*. New York, NY: McGraw-Hill.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.* 7, 657–660. doi: 10.1177/1745691612462588

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Sci. Transl. Med.* 349:aac4716. doi: 10.1126/science.aac4716

Pashler, H., and Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* 7, 531–536. doi: 10.1177/1745691612463401

Phaf, R. H. (2020). Publish less, read more. *Theory Psychol.* 30, 263–285. doi: 10.1177/0959354319898250

Pieters, J. M. (1981). Ear asymmetry in an auditory spatial Stroop task as a function of handedness. *Cortex: J. Devoted Study Nervous Syst. Behav.* 17, 369–379. doi: 10.1016/S0010-9452(81)80024-X

Platt, J. R. (1964). Strong inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* 146, 347–353. doi: 10.1126/science.146.3642.347

Regan, J. (1978). Involuntary automatic processing in color-naming tasks. *Percept. Psychophys.* 24, 130–136. doi: 10.3758/BF03199539

Roberts, K. L., and Hall, D. A. (2008). Examining a supramodal network for conflict processing: a systematic review and novel functional magnetic resonance imaging data for related visual and auditory stroop tasks. *J. Cogn. Neurosci.* 20, 1063–1078. doi: 10.1162/jocn.2008.20074

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* 86, 638–641. doi: 10.1037/0033-2909.86.3.638

Rouder, J. N. (2014). Optional stopping: no problem for bayesians. *Psychon. Bull. Rev.* 21, 301–308. doi: 10.3758/s13423-014-0595-4

Sanborn, A. N., and Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychon. Bull. Rev.* 21, 283–300. doi: 10.3758/s13423-013-0518-9

Scheel, A. M., Schijen, M., and Lakens, D. (2020). An excess of positive results: comparing the standard psychology literature with registered reports. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/p6e9c

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. doi: 10.1037/a0015108

Schneider, B. A., Avivi-Reich, M., Leung, C., and Heinrich, A. (2016). How age and linguistic competence affect memory for heard information. *Front. Psychol.* 7:618. doi: 10.3389/fpsyg.2016.00618

Shilling, V. M., Chetwynd, A., and Rabbitt, P. M. (2002). Individual inconsistency across measures of inhibition: an investigation of the construct validity of inhibition in older adults. *Neuropsychologia* 40, 605–619. doi: 10.1016/S0028-3932(01)00157-9

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. doi: 10.1177/0956797611417632

Sommers, M. S., and Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: the interaction of lexical competition and semantic context. *Psychol. Aging* 14, 458–472. doi: 10.1037/0882-7974.14.3.458

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1412159

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J. Exp. Psychol.* 18, 643–662. doi: 10.1037/h0054651

Studdert-Kennedy, M., and Shankweiler, D. (1970). Hemispheric specialization for speech perception. *J. Acoust. Soc. Am.* 48, 579–594. doi: 10.1121/1.1912174

Watson, D. (2004). Stability versus change, dependability versus error: issues in the assessment of personality over time. *J. Res. Personal.* 38, 319–350. doi: 10.1016/j.jrp.2004.03.001

Whitton, J. P., Hancock, K. E., Shannon, J. M., and Polley, D. B. (2017). Audiomotor perceptual training enhances speech intelligibility in background noise. *Curr. Biol.* 27, 3237–3247. doi: 10.1016/j.cub.2017.09.014

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Ziegler, J. C., Pech-Georgel, C., George, F., Alario, F.-X., and Lorenzi, C. (2005). Deficits in speech perception predict language learning impairment. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14110–14115. doi: 10.1073/pnas.0504446102

Conflict of Interest: 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.