Exploring Heterogeneity in Factor Analytic Results

by

Patrick Don Manapat

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2022 by the
Graduate Supervisory Committee:

Michael C. Edwards, Co-Chair
Samantha F. Anderson, Co-Chair
Kevin J. Grimm
Roy Levy

ARIZONA STATE UNIVERSITY

August 2022

ABSTRACT

The last two decades have seen growing awareness of and emphasis on the replication of empirical findings. While this is a large literature, very little of it has focused on or considered the interaction of replication and psychometrics. This is unfortunate given that sound measurement is crucial when considering the complex constructs studied in psychological research. If the psychometric properties of a scale fail to replicate, then inferences made using scores from that scale are questionable at best. In this dissertation, I begin to address replication issues in factor analysis – a widely used psychometric method in psychology. After noticing inconsistencies across results for studies that factor analyzed the same scale, I sought to gain a better understanding of what replication means in factor analysis as well as address issues that affect the replicability of factor analytic models. With this work, I take steps toward integrating factor analysis into the broader replication discussion. Ultimately, the goal of this dissertation was to highlight the importance of psychometric replication and bring attention to its role in fostering a more replicable scientific literature.

DEDICATION

To Danielle and Parker, for their unconditional love and unwavering support. Everything I do is for you two.

To my parents, Armie and Cezar, for their sacrifices that have paved the way for this PhD.

To my sister, Carmella, for giving me a reason to set a good example. I look up to you now.

ACKNOWLEDGMENTS

As a student of ASU, I acknowledge that ASU would not exist without the theft of the lands of the Indigenous peoples of the Phoenix area. ASU sits on the unceded territory of the Akimel O'odham and Pee Posh peoples. I also acknowledge the immeasurable amount of harm that the theft of these lands has done to the Akimel O'odham and Pee Posh peoples including loss of language, food sources, culture, kinship systems, and relationships to the land. I acknowledge my status as a guest on these lands and hope to one day see the return of these lands to the original peoples. The Akimel O'odham and Pee Posh peoples are strong and have been resilient despite the active genocide of their people which continues today. I support their sovereignty and will always advocate and center their voices in their original territory.

I would like to express my deepest gratitude to my advisor, Mike. Thank you for giving me a shot. It has been an honor to be your student. Through your training, I have gained invaluable skills for success within and, more importantly, beyond academia.

I am also grateful to my co-chair, Samantha, as well as my dissertation committee members, Kevin and Roy. Thank you for serving on my defense committee and your guidance throughout the process.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Over the past few decades, a renewed interest in replication has led to significant progress in how we think about and address the replicability of scientific research. In psychology specifically, the heightened awareness around replication has led to a great deal of progress (e.g., Shrout & Rodgers, 2018) including improvements to procedural, statistical, and editorial aspects of psychological science (e.g., McNutt, 2014; Nosek et al., 2015; Wicherts et al., 2016). However, one important aspect that has been largely overlooked is psychometrics (Flake & Fried, 2020; Pargent et al., 2019; Shaw et al., 2020). This is unfortunate given that sound measurement is crucial when considering the complex constructs involved in human cognition, emotion, and behavior. If the psychometric properties of a scale fail to replicate, then inferences made using scores from that scale are questionable and/or uninterpretable (Flake et al., 2017, 2022). As the foundation for many scales and assessments used in psychological studies, it is crucial that the replicability of psychometric work be integrated into the broader replication discussion.

At the intersection of replication and psychometrics, I was interested in the replicability of factor analytic results. Factor analysis is a widely used psychometric method for scale development and evaluation in psychology. It is primarily used to assess dimensionality (i.e., how many latent variables underlie a set of measured variables/indicators/items) and learn how measured variables relate to the underlying latent variables (i.e., factor structure). I propose that the replicability of factor analyses is an important consideration for the plausibility of factor structures. To start, I use a very

broad definition of replication within the context of factor analysis: A practical equivalence between studies in the (1) number of factors, (2) factor structure, and/or (3) factor labels. If studies factor analyzing the same scale fail to reach a practical equivalence for any of these aspects, then the use of that scale may need to be called into question.

As part of a psychometric evaluation of the Brief Self-Control Scale (BSCS; Chapter 2; Manapat et al., 2021), I noticed inconsistencies across results for studies that factor analyzed the BSCS (Tangney et al., 2004). The BSCS is a scale that asks respondents to rate 13 statements (e.g., *I am good at resisting temptation*) on a scale ranging from one (*not at all like me*) to five (*very much like me*). Scores on the BSCS are purported to represent levels of self-control and have often been used to study other important constructs (e.g., academic performance, self-esteem, health-related behaviors, substance abuse; Hagger et al., 2018; Morean et al., 2014; Tangney et al., 2004). Across studies that had factor analyzed the BSCS, there was disagreement about the number of factors, factor structure, and factor labels. This was troubling because many studies have used the BSCS to measure self-control despite the inconsistent factor analytic results in the literature. I was interested in examining whether the inconsistencies observed for the BSCS were common and extended to other scales in psychology.

I extended the review of factor analytic studies to assess whether there was similar disagreement across results for the Center for Epidemiologic Studies-Depression Scale (CES-D) and the NEO Personality Inventory-Revised (NEO-PI-R; Costa Jr. & McCrae, 1992). As was observed for the BSCS, studies factor analyzing the CES-D and NEO-PI-R also failed to agree on the number of factors, factor structure, and factor

labels. Having found these issues to persist beyond a single scale, I sought to further this investigation of replicability in factor analysis.

Although the inconsistencies observed in factor analytic results for the three psychological scales could be conceptualized as failures to replicate, it is not reasonable to expect results across replications to be the same. The replication literature has moved past binary assessments of replication (success vs. failure) and evolved into more nuanced, meta-scientific commentaries on what should (and should not) be expected of replication studies (e.g., carefully defining replication goals, moving beyond single-study evaluations; Anderson & Maxwell, 2016; Hedges & Schauer, 2019b, 2019a; Steiner et al., 2019). To effectively address factor analytic replication, it is important to interpret results in light of anticipated heterogeneity. Across replication studies, differences in results should be expected to at least some degree (Kenny & Judd, 2019; McShane et al., 2019; McShane & Böckenholt, 2014). I sought to identify the factors contributing to the heterogeneity and drew insights from what replication researchers have already found to be impacting replicability outside of psychometrics.

An issue I found to be prevalent for replication more broadly was methodological flexibility. When conducting a statistical analysis, there are many methodological options to consider. Different methodological choices have the potential to produce different statistical results (inject heterogeneity), regardless of defensibility (e.g., Klein et al., 2018; Silberzahn et al., 2018). I was interested in evaluating methodological flexibility in factor analysis, but first needed a coherent framework for evaluation. Unfortunately, what I found was inconsistent, vague, and overlapping use of extant language. There was an opportunity to clarify definitions for terms such as *researcher degrees of freedom* (RDFs;

e.g., Simmons et al., 2011) and *questionable research practices* (QRPs; e.g., John et al., 2012), which would enhance conceptual clarity and allow for more effective evaluations of research practices (for psychometrics and in general).

To facilitate evaluations of methodological flexibility, I developed a revised and expanded taxonomy for assessing research practices (Chapter 3; Manapat et al., 2022). This taxonomy was based on revised definitions for RDFs and QRPs that reduced overlap and focused on distinguishing between practices that are empirically supported (e.g., RDFs) and practices known to be suboptimal (e.g., QRPs). The goals of this taxonomy were to help researchers understand the different sources of heterogeneity, provide a structure for organizing and evaluating research practices that contribute to unnecessary heterogeneity, help researchers navigate the complex series of decisions throughout the research process, and drive research toward areas where there is a need for more methodological guidance. My primary aim was to deploy this taxonomy in subsequent evaluations of factor analytic replication in the empirical literature as well as simulation-based studies.

In the chapters that follow, I outline an exploration of heterogeneity in factor analytic results. Chapter 2, entitled *A Psychometric Analysis of the Brief Self-Control Scale* (Manapat et al., 2021), is the study where I first documented heterogeneity across factor analytic results. Chapter 3, entitled *A Revised and Expanded Taxonomy for Understanding Heterogeneity in Research and Reporting Practices* (Manapat et al., 2022), is the study where I propose a new taxonomy for addressing methodological flexibility, a source of heterogeneity in research results. To conclude, I summarize the

progress made in addressing heterogeneity in factor analytic results and describe two extensions of the current work.

CHAPTER 2

A PSYCHOMETRIC ANALYSIS OF THE BRIEF SELF-CONTROL SCALE

Abstract

The Brief Self-Control Scale (BSCS) is a widely used measure of self-control, a construct associated with beneficial psychological outcomes. Several studies have investigated the psychometric properties of the BSCS but have failed to reach consensus. This has resulted in an unstable and ambiguous understanding of the scale and its psychometric properties. The current study sought resolution by implementing scale evaluation approaches guided by modern psychometric literature. Additionally, our goal was to provide a more comprehensive item analysis via the item response theory (IRT) framework. Results from the current study support both unidimensional and multidimensional factor structures for the 13-item version of the BSCS. The addition of an IRT analysis provided a new perspective on item- and test-level functioning. The goal of a more defensible psychometric grounding for the BSCS is to promote greater consistency, stability, and trust in future results.

Introduction

As defined by Tangney, Baumeister, and Boone (2004), self-control is "the ability to override or change one's inner responses, as well as to interrupt undesired behavioral tendencies and refrain from acting on them," (p. 274). This construct is extensively researched in the literature and numerous scales have been created intending to measure it (see Sharma, Markon, & Clark, 2014). The Brief Self-Control Scale (BSCS), developed by Tangney et al. (2004), is one measure that is widely used (Table 1). The BSCS is most commonly used for investigations of the association between self-control and various positive outcomes (Ferrari, Stevens, & Jason, 2009; Maloney, Grawitch, & Barber, 2012; Tangney et al., 2004). Despite the breadth of studies that use the BSCS, researchers have yet to agree on the dimensionality of the scale. Using principal components analysis, Tangney et al. (2004) found the original 36-item Self-Control Scale (SCS) to represent five components: "self-discipline," "deliberative/nonimpulsive action," "healthy habits," "work ethic," and "reliability." However, the authors state these five components did not improve prediction of outcomes (e.g., academic performance, psychological adjustment) and recommend use of the total score which is suggestive of unidimensionality.

Subsequent analyses of the more common BSCS seem to suggest that one factor may not be adequately representative. This disagreement can be attributed, at least in part, to differences in methodology which is summarized in Table 2 (along with the methodology used for the current study). This variation in methods has likely contributed to different conceptualizations of the BSCS. Ferrari et al. (2009) found a two-factor structure to be most appropriate, comprising factors of "self-discipline" (Items 2, 3, 4, 5,

7, 9, 10, 12, and 13) and “impulse control” (Items 1, 6, 8, and 11). It is important to note that this separation maps onto how the items are phrased (positive vs. negative). De Ridder, De Boer, Lugtig, Bakker, and Van Hooft (2011) also found a two-factor structure to be most appropriate, comprising factors of “inhibitory self-control” (Items 1, 2, 5, 6, 9, and 12) and “initiatory self-control” (Items 3, 10, 11, and 13). There was a third factor consisting of “generic nature” items (Items 4, 7, and 8) but these items were removed prior to analysis. Maloney et al. (2012) settled on a final model with two factors: “impulsivity” (Items 5, 9, 12, and 13) and “restraint” (Items 1, 2, 7, and 8). Items 3, 4, 6, 10, and 11 were removed. Also, deciding on a two-factor structure, Morean et al. (2014) identified “self-discipline” (Items 5, 9, 12, and 13) and “impulse control” (Items 1, 8, and 11). Items 2, 3, 4, 6, 7, and 10 were removed.

In summary, each study produced a unique factor structure for the BSCS. Some studies retained all 13 items, while others removed items. Some authors used substantive arguments to justify item removal (e.g., De Ridder et al., 2011), while others removed items based on empirical findings (e.g., Maloney et al., 2012). There are also differences in the conceptualizations of the factors as well as which particular items relate to those corresponding factors. The study conducted by Lindner, Nagy, and Retelsdorf (2015) attempted to resolve the disparities between four of the five studies previously mentioned. It is assumed that these authors excluded the Morean et al. (2014) study since it was likely unavailable when they did the research.

Lindner et al. (2015) found the Ferrari et al. (2009) factor structure to be the most plausible model. However, this model is potentially confounded by valence of phrasing, where all negative items loaded onto one factor and all positive items loaded onto the

other. In other words, the factors may represent “negatively phrased” and “positively phrased” instead of “self-discipline” and “impulse control,” respectively. Lindner et al. (2015) found the Maloney et al. (2012) model to be the next best which fit well in a sample of apprentices in vocational training. However, this model did not fit in a sample of university students. Ultimately, it was inconclusive whether a unidimensional or multidimensional factor structure better served the BSCS (Lindner et al., 2015). The authors concluded with a recommendation of using the BSCS total score since the one-factor model outperformed the two-factor model in outcome prediction (e.g., life satisfaction, grades, dropout intention, self-assessed achievement). Across the 11 years (2004-2015) of research on the BSCS, studies have toggled between unidimensional and multidimensional conceptualizations leaving the nature of this scale unresolved.

As previously mentioned, these studies differed in both their choice of methods/procedures and the order in which these methods/procedures were applied (Table 2). Two studies used a single method to assess dimensionality. Ferrari et al. (2009) conducted an exploratory factor analysis (EFA) using maximum likelihood (ML) estimation with an orthogonal varimax rotation and De Ridder et al. (2011) conducted a confirmatory factor analysis (CFA) but did not specify the estimator. Two studies used a combination of EFA and CFA. Maloney et al. (2012) utilized an EFA (estimator not specified) with an oblique direct oblimin rotation followed by a CFA with ML estimation. In contrast, Morean et al. (2014) used these methods in reverse order. These authors first conducted a CFA with robust ML estimation on the factor structure proposed by Maloney et al. (2012). The model did not fit the data well so Morean et al. (2014) attempted to establish an improved factor structure by means of an EFA via robust

ML estimation with an oblique varimax rotation. Lindner et al. (2015) used CFA (estimator not specified) to test the factor structures proposed by Ferrari et al. (2009), Maloney et al. (2012), and De Ridder et al. (2011). Within each, Lindner et al. (2015) examined both a unidimensional model and a two-factor model.

In addition to the lack of clarity surrounding the BSCS and its dimensionality, there has yet to be an examination of the scale under the item response theory (IRT) framework. According to Edwards (2009), IRT is a collection of latent variable models that seek to uncover the underlying process that influences responses to observed variables. This is driven by properties about individuals (i.e., θ; corresponds to level on the latent variable or construct of interest such as self-control in the present context) and properties about the items. IRT is a preferred method because it extracts more detailed parameters about items. This gives IRT scale scores a number of benefits including a common, easily interpretable metric, greater score variability than summed scores, conditional standard errors, and straightforward equating (De Ayala, 2008; Embretson & Reise, 2000).

Samejima's (1969) graded response model (GRM) is the most popular for psychological scales such as the BSCS, where individual items consist of more than two response options (e.g., Likert-type). The GRM is formulated as follows:

$$P(x_j = c \mid \theta) = \frac{1}{1+\exp\left[-a_j\left(\theta - b_{cj}\right)\right]} - \frac{1}{1+\exp\left[-a_j\left(\theta - b_{(c+1)j}\right)\right]}$$

which represents the probability of endorsing response option c given θ which produces the observed response (x_j) to item j.

The *a*-parameter represents the slope for item *j* and is also referred to as the discrimination parameter. This parameter describes the relationship between an item and the latent construct (e.g., self-control). Higher slopes indicate that more of the variability in item responses can be attributed to differences in the latent construct. This may also be interpreted as having a stronger relationship with the latent construct. The *b*-parameter represents the threshold of response option *c* for item *j* and has been historically referred to as the severity parameter. For the BSCS, which contains five response options, there are four ($c - 1$) thresholds. These threshold parameters indicate how much self-control is required for a respondent to endorse a particular response option. So, the higher the threshold, the higher the level of self-control required to endorse the response option.

Additional benefits of working under the IRT framework for scale evaluation are outlined in Edwards (2009), Embretson and Reise (2000), and Thissen and Steinberg (2009). For the purposes of the current study, we focus on the benefits associated with scoring and score precision. Scoring in IRT involves weighting response patterns using the item parameters. As mentioned previously, higher slopes indicate that an item tells us more about the latent construct. IRT scale scores take this item-construct relationship into account instead of weighting all of the items equally as is done for summed scores. IRT scale scores are also more variable and allow for a greater degree of differentiation between individuals. Also related to scoring is equating. When items are properly calibrated, IRT scale scores may be directly compared regardless of the specific items administered.

Score precision, closely related to reliability, is also approached differently in IRT. Rather than assuming all scores are equally reliable, IRT defaults to an assumption

that scores vary in their precision/reliability. IRT characterizes precision/reliability using a number of different metrics. One such metric is Fisher information (hereafter called information). Information is provided at the item-level through item information functions (IIFs) as well as the test-level through test information functions (TIFs). Both functions plot the amount of information provided by the item or test across the entire range of the latent construct. In this way, we are able to assess how informative an item or a test is at differing levels of the latent construct (e.g., self-control). Standard errors under IRT are inversely related to information ($\sqrt{1/\text{INF}}$) and may also be computed for any value across the range of the latent construct (Thissen & Wainer, 2001). Because information and standard errors are provided for every value of the latent construct, the IRT framework paints a better picture in terms of score precision. Other overviews and examples, with some being more in-depth, of scale development under the IRT framework may be found in DeVellis (2012), DeWalt et al. (2013), and Preston et al. (2018).

One aspect across the previous BSCS studies that was relatively stable was internal consistency which was computed per factor (one reliability estimate for a one-factor solution and two reliability estimates for a two-factor solution). Most reported using coefficient alpha (Cronbach, 1951) but one study failed to specify which measure of internal consistency was used (Maloney et al., 2012). For studies that used reduced versions of the BSCS, reliabilities ranged from .65 to .78 (De Ridder et al., 2011; Maloney et al., 2012; Morean et al., 2014). However, these values were based on different versions of the BSCS where number of items and conceptualization of factors differed. Therefore, it would be inappropriate to directly compare these measures of

internal consistency. Of the studies that retained the original version of the BSCS and used all 13 items, reliabilities ranged from .69 to .85 (Ferrari et al., 2009; Tangney et al., 2004). There is agreement (at least between two studies) and support for adequate internal consistency of the original BSCS.

The purpose of the current study was to evaluate the BSCS using evaluation methods guided by the psychometric literature. The primary aims were to bring clarity to the dimensionality dispute by removing methodology as a source of variability in factor analytic results. Specifically, we were interested in the dimensionality of the BSCS which attempts to measure self-control and not the conceptual dimensionality of self-control as a construct. We planned to accomplish this through the use of psychometric best practices as well as detailed and transparent reporting of these best practices. Additionally, we planned to provide a more comprehensive item analysis via the IRT framework which offers advantages above and beyond the methods historically used for psychometric evaluations of the BSCS.

Currently, evidence for the factor structure of the BSCS is inconclusive which has resulted in a compromised understanding of the measurement instrument (Morean et al., 2014). The lack of agreement is likely driven, in part, by the inconsistent methodology applied across studies which could explain why different researchers have suggested different factor structures. Because there is no consensus over the number and nature of the BSCS factors, Morean et al. (2014) note that interpretations of this measure and its relations with other variables would be questionable at best. These authors stressed the importance of a sound measurement instrument which they consider to be a prerequisite for drawing valid conclusions from study results. The current study intended to establish

a more trustworthy conceptualization of the BSCS with a set of new, evidence-based psychometric approaches. The long-term goal is to promote consistency in methods in an effort to unify results.

Proper development of a self-control measure is extremely valuable due to the importance of this construct for overall psychological well-being. Self-control is described by Tangney et al. (2004) as highly adaptive and essential for happiness and good health. Therefore, it is paramount to ensure the BSCS is adequately measuring what it was intended to measure. Additionally, many studies have investigated associations between self-control and favorable outcomes. These include greater academic performance, increased impulse control, better psychological adjustment, higher self-esteem, healthy interpersonal relationships, well-adjusted emotional patterns, abstaining from substance abuse, positive affect, and other behavioral advantages (De Ridder et al., 2011; Ferrari et al., 2009; Maloney et al., 2012; Tangney et al., 2004). Assuming plausible models are found, we also plan to conduct preliminary validation analyses to better understand the resulting factor(s).

Given the evidence in the literature about self-control and this construct's associations with various outcomes, the following hypotheses guided our evaluation of the BSCS and the validity of its use. First, Ferrari et al. (2009) found a positive association between self-control and abstaining from substance use. Therefore, we hypothesized a negative association between the BSCS and alcohol use. Next, Tangney et al. (2004) found self-control to be related to psychological adjustment and impulse control and Maloney et al. (2012) found associations between self-control and positive behaviors. Based on these two studies, we first hypothesized a negative association

between the BSCS and impulsivity. Then, based on the work by Babinski, Hartsough, and Lambert (1999) who found impulsivity and conduct problems to positively predict arrest records, we hypothesized a negative association between the BSCS and arrests. Last, research conducted by Čubranić-Dobrodolac, Lipovac, Čičević, and Antić (2017) found impulsivity and aggressive behavior to positively predict the occurrence of traffic accidents. Therefore, it was hypothesized that the BSCS would be negatively associated with traffic accidents.

Method

Participants

The first data set (Sample 1) for the current study was collected as part of a research program primarily focused on developing an ontology of self-regulation (Eisenberg et al., 2018). Data were collected using Amazon's Mechanical Turk (MTurk). Recruitment of participants using MTurk allows for more diverse samples than is typically seen with convenience samples drawn from the university undergraduate population. Sample 1 consisted of 522 U.S.-based participants who completed the 13-item BSCS as part of the research program previously described. The mean age for Sample 1 was 33.63 years (*SD* = 7.87), 51% were female, 86% identified themselves as White, and 44% were at least college educated.

The second data set (Sample 2) was an undergraduate sample collected at George Mason University in 2012. Initially, there were 529 participants that responded to the 13-item BSCS. However, participants were removed if they responded to at least one out of three validity probes incorrectly. An example validity probe was "Select 'False, not at all true' as your response to this question." This left a total of 298 participants in Sample 2.

The mean age for Sample 2 was 22.12 years (*SD* = 5.62), 25% were male, 74% were female, and 1% preferred not to answer. No other demographic information was provided about Sample 2.

Measures

Brief Self-Control Scale

The original version of the 13-item BSCS (Tangney et al., 2004) was administered to participants. The 13-item BSCS is a short-form of the full 36-item SCS developed by the same authors. The benefit of using the short-form version is the reduction in participant burden (Morean et al., 2014). Additionally, in previous research, the short-form achieved a reliability very similar to the full version. Tangney et al. (2004) reported coefficient alphas (Cronbach, 1951) for the BSCS of .83 and .85 for their first and second samples, respectively. These values were very close to the reliability of the SCS ($\alpha = .89$) which suggests similar performance between short and long forms. The 13 items of the BSCS all consist of a 5-point rating scale anchored by 1 (*not at all like me*) and 5 (*very much like me*). Reponses were considered as ordinal and all analyses were conducted to account for the categorical nature of the data. Negatively phrased items were recoded so that higher scores indicated higher levels of self-control. The items as well as the valence of phrasing are provided in Table 1. Polychoric correlations between the 13 items for Sample 1 are displayed in Table 3.

Validation

Five measures from Sample 1 were used for a preliminary assessment of validity based on the BSCS factor structures found in this study. These were first collected concurrently with the BSCS at initial data collection. These same five measures were

collected again approximately 111 days later which is the mean number of days between the two data collection waves (Enkavi et al., 2019). The measures included frequency of alcohol use, number of alcoholic drinks per day, the Barratt Impulsiveness Scale (BIS-11; Patton, Stanford, & Barratt, 1995), number of lifetime arrests, and number of lifetime traffic accidents. Frequency of alcohol use was measured with "How often do you have a drink containing alcohol?" and this item was on a 5-point rating scale anchored by 1 (*never*) and 5 (*four or more times a week*). The 30 items of the BIS-11 all consist of a 4-point rating scale anchored by 1 (*rarely/never*) and 4 (*almost always/always*). A summed score was computed for the cognitive impulsivity and behavioral impulsivity factors as a proxy to the structure presented by Reise, Moore, Sabb, Brown, and London (2013). Negatively phrased items were recoded so that higher scores indicated higher levels of cognitive or behavioral impulsivity.

Psychometric Analyses

The current study first sought to assess the dimensionality of the BSCS. To accomplish this, a combination of EFA and CFA was used which is a popular recommendation. For the EFA and CFA, Sample 1 was split to contain half of the participants in each subsample. Specifically, 261 participants (exploratory half of Sample 1) were used in an exploratory phase to find plausible models through EFA and CFA. The remaining 261 participants (holdout half of Sample 1) were used for CFAs to validate the models found in the exploratory phase. The exploratory and holdout samples would then be combined to obtain final estimates. Since Sample 1 was used for both exploration and validation using the split-half approach, we obtained Sample 2 to validate our results in a new and external sample. Plausible CFA models that were tested in the

holdout half of Sample 1 were tested again in Sample 2 (N = 285 after listwise deletion). Given that the models in Sample 1 hold in Sample 2, this study then sought to provide a new perspective on the BSCS through the lens of IRT. Analyses using the GRM would be conducted on the complete Sample 1 (N = 522) on models deemed as most plausible by the EFA and CFA. Last, internal consistency statistics can be calculated for the factor (or separately in the case of multiple factors). Internal consistency was based on the complete Sample 1 (N = 522).

In a case where the EFA and CFA suggest a multidimensional factor structure for a scale, Edwards (2009) recommends two different solutions. First, the dimensionality assessment may be used to guide modifications to the original scale. For example, items may be dropped from the original version to achieve a plausibly unidimensional revised version. However, we sought to find a model suitable for all 13 items so we turned to the second option: multidimensional IRT (MIRT). This solution is appropriate for scales that exceed a single dimension and does not require the assumption of unidimensionality. Given the plausibility of more than one factor, the multidimensional extension of the GRM (multidimensional GRM; MGRM) would be fit to the complete Sample 1 (N = 522). As noted by Wirth and Edwards (2007), multidimensional constructs are common in psychology and MIRT methods allow researchers to properly model scales of this nature. Although these authors mention the challenges posed to the estimation of the more complex MIRT models, remedies are available. In particular, the advent of more recent software such as flexMIRT (Cai, 2017) has addressed these estimation issues. Specifically, a Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010a; Cai, 2010b) allows for higher dimensional IRT models to be estimated. More

details on MIRT may be found in Ackerman, Gierl, and Walker (2003), Monroe and Cai (2015), Reckase (1997), and Wirth and Edwards (2007).

Exploratory Factor Analysis

We decided to use more than one method for determining the number of factors because mechanical rules tend to be arbitrary (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Information from two different methods would better guide our decisions for the number of factors to estimate and prevent overfactoring or underfactoring. Based on recommendations from Fabrigar et al. (1999), a scree plot and parallel analysis (Horn, 1965) were used as these methods have been shown to outperform other approaches. All EFAs were conducted in CEFA (Browne, Cudeck, Tateneni, & Mels, 2010). The exploratory half of Sample 1 was used for the EFAs. Since prior BSCS studies used ML estimation with Pearson correlations which treats the data as continuous or made no mention of the type of correlation, this study is likely the first to account for the categorical nature of the BSCS items. This was done using polychoric correlations and ordinary least squares estimation (Christoffersson, 1975; Edwards, 2009). For models with more than one factor, we used an oblique rotation because orthogonal rotations in psychological research are rarely justifiable (Byrne, 2005; Fabrigar et al., 1999). Direct quartimin was selected for rotation which has exhibited success in obtaining interpretable solutions (Fabrigar et al., 1999).

Confirmatory Factor Analysis

The information gained from the EFAs guided our selection of models to fit using CFA. Factor structures deemed as plausible were fit using the “lavaan” package (Rosseel, 2012) in R (R Core Team, 2018). CFAs were fit to both exploratory and holdout halves

of Sample 1. In the CFAs with the exploratory half, modification indices (MIs) were inspected. The holdout half was used to validate plausible models after modifications were added. For all CFAs, factor variances were set to one for model identification. Polychoric correlations and the diagonally weighted least squares (DWLS; also called WLSMV in "lavaan") estimator were specified to account and correct for the categorical (i.e., ordinal) nature of the data and to produce accurate indices of model fit (Christoffersson, 1975; Flora & Curran, 2004). As with the EFAs, prior BSCS studies ran CFAs treating the BSCS items as continuous or failed to report the estimator. Models validated using the holdout half of Sample 1 were validated again using Sample 2. There was no missing data for Sample 1 and missingness in Sample 2 was handled via listwise deletion resulting in a sample size of 285 for analyses.

Graded Response Model

The item parameters for the GRM were estimated in flexMIRT (Cai, 2017). The analyses under the GRM were conducted on the complete Sample 1. Estimation for unidimensional models was performed using the marginal ML estimator with expectation-maximization (MML-EM). Estimation for multidimensional models was performed using the previously mentioned MH-RM algorithm. All other program specifications were set to the defaults.

Internal Consistency

Coefficient alpha (Cronbach, 1951) and coefficient omega (McDonald, 1970) were computed using the "coefficientalpha" package (Zhang & Yuan, 2015) in R (R Core Team, 2018). Coefficient alpha (unless strict, almost unrealistic test assumptions are met) is a lower bound of reliability (Sijtsma, 2009). Therefore, coefficient omega is provided

as an alternative and provides a better estimate of reliability under more realistic test assumptions (McNeish, 2018).

Validation

IRT scale scores were obtained from flexMIRT (Cai, 2017). Specifically, we used the expected a posteriori (EAP; Bock & Mislevy, 1982) estimates for the level of self-control. These scores are derived from the mean of the posterior distribution. For more technical details on computing EAPs, please refer to Thissen and Wainer (2001). The EAPs from plausible models were entered as a predictor of frequency of alcohol use, number of alcoholic drinks per day, the two types of impulsivity as measured by the BIS-11, number of lifetime arrests, and number of lifetime traffic accidents. Each of these outcomes comprised a separate regression model. For multidimensional models, the EAPs for each factor were entered simultaneously as predictors for separate models with each outcome measure. Poisson regression was used for the following count outcomes: number of alcoholic drinks per day, number of lifetime arrests, and number of lifetime traffic accidents.

Results

Determining the most plausible BSCS factor structure was first guided by a set of EFAs conducted on the exploratory half of Sample 1. The number of factors to estimate was determined using the scree plot and parallel analysis depicted in Figure 1. The scree plot suggested one, two, or three factors. The parallel analysis (scree plot with random component added as the dashed line) suggested one factor. To avoid underfactoring, we estimated a one-, two-, three-, and four-factor model.

Solution selection among these models was based on the idea of simple structure (Thurstone, 1947). A solution that meets simple structure has high variability in loadings within factors and strong loadings for items onto the fewest number of common factors (low factorial complexity). According to Thurstone (1947), simple structure allows factor solutions to be easily interpretable, meaningful, and replicable. The four- and three-factor solutions failed to meet the properties of simple structure. Specifically, the four-factor model had two weak factors composed of only one (Factor 3) or two (Factor 4) items. This was suggestive of overfactoring. For the three-factor model, there was also a "small" factor, with strong loadings to only two items.

For these reasons, the three- and four-factor models were deemed implausible and we focused our attention on the one- and two-factor models. Factor loadings, communalities, and factor correlations for these models are presented in Table 4. According to MacCallum, Widaman, Zhang, and Hong (1999), good recovery is possible with smaller sample sizes (i.e., less than 100) when communalities are high. The relatively large size of factor loadings and communalities across the two models considered here suggest our samples were of sufficient size for accurate estimation. The two-factor model improved on the three- and four-factor models based on the simple structure criteria. However, this model had one item that failed to load onto either of the factors and one cross-loading. The one-factor model, in contrast, had relatively high loadings for all items onto the single factor making it the clearest and least complex model with regard to the relationship between items and their respective factor(s). The one-factor model was also the most parsimonious (i.e., fewest factors). With all models

considered, the results of the EFAs suggested the plausibility of both one- and two-factor models.

Since the one-factor and two-factor models were most promising, we proceeded by fitting both using CFA. These models were evaluated with the following indices of model fit: χ^2, Tucker–Lewis index (TLI), comparative fit index (CFI), standardized root mean square residual (SRMR), and root mean square error of approximation (RMSEA). Guidelines for acceptable model fit include a TLI and CFI greater than .93 (Hu & Bentler, 1999), an SRMR less than .08 (Hu & Bentler, 1999), and an RMSEA less than .10 (Browne & Cudeck, 1992). Because the computation of the χ^2 statistic is dependent on sample size, one may reject an adequately representative model simply due to a large sample size. Thus, decisions concerning model fit should be based on a combination of indices that lend support to the plausibility of a model rather than a single index. A summary of fit for all CFA models tested is presented in Table 5.

First, all plausible models as indicated by the EFA were tested in the exploratory half of Sample 1. This included the one-factor model (1F) and the two-factor model with cross-loadings (2FCL). For the 2FCL model, some of the factor loadings lacked statistical significance. A closer look showed that the cross-loading items each had one strong (and statistically significant) loading and one weak (and nonsignificant) loading. Therefore, a subsequent two-factor model with independent clustering (2FIC) was tested. This model mirrored the 2FCL but removed the nonsignificant cross-loadings for Item 4 and Item 7. Next, modification indices (MIs) were inspected. Across all three models, there were three modifications in the top five when MIs were ordered by magnitude. These included residual covariances between Item 1 and Item 8 (both items involve restraint), Item 5 and

Item 6 (both items include the phrase “bad for me”), and Item 12 and Item 13 (both items involve impulsiveness). These three residual covariances were added to all three models and retested. The 1F model with the three residual covariances (1F-CR) exhibited adequate fit to the data with the exception of RMSEA which was slightly over the cutoff: $\chi^2(62) = 242.40$, $p < .05$; TLI = .984; CFI = .988; SRMR = .075; RMSEA = .106, 90% confidence interval (CI) [.092, .120]. The 2FCL model with the three residual covariances (2FCL-CR) exhibited adequate fit overall: $\chi^2(59) = 159.54$, $p < .05$; TLI = .991; CFI = .993; SRMR = .061; RMSEA = .081, 90% CI [.066, .096]. The 2FIC model with the three residual covariances (2FIC-CR) also exhibited adequate fit overall: $\chi^2(61) = 169.97$, $p < .05$; TLI = .990; CFI = .993; SRMR = .063; RMSEA = .083, 90% CI [.068, .098].

When comparing both two-factor models, the 2FIC-CR fit almost as well as the 2FCL-CR, had a cleaner structure, and all factor loadings were statistically significant ($p < .001$). For these reasons, the 2FIC-CR was retained for validation analyses in addition to the 1F-CR. According to Fabrigar et al. (1999), the goal is a simpler model that accounts for the data nearly as well as a more complex model. Considering both model fit and parsimony, there was not enough evidence to definitively eliminate either the one- or two-factor solution for the BSCS. A CFA for validation was run on both the 1F-CR and 2FIC-CR models using the holdout half of Sample 1. Both models adequately fit the data: $\chi^2(62) = 187.41$, $p < .05$; TLI = .987; CFI = .989; SRMR = .064; RMSEA = .088, 90% CI [.074, .103] and $\chi^2(61) = 140.29$, $p < .05$; TLI = .992; CFI = .993; SRMR = .056; RMSEA = .071, 90% CI [.055, .086], for the one- and two-factor models respectively. Since both factor structures held in the holdout half of Sample 1, the

exploratory and holdout halves were combined and the 1F-CR and 2FIC-CR were fit to the complete Sample 1. The factor loadings and factor correlations for these final models are reported in Table 6.

The final set of CFAs were run to further validate the factor structures found in Sample 1. Our goal was to test whether the 1F-CR and 2FIC-CR identified with our split-half approach would hold in a separate and external sample (Sample 2). Both models fit adequately in the new sample: $\chi^2(62) = 164.48$, $p < .05$; TLI = .981; CFI = .985; SRMR = .065; RMSEA = .076, 90% CI [.062, .091] and $\chi^2(61) = 146.89$, $p < .05$; TLI = .984; CFI = .987; SRMR = .061; RMSEA = .070, 90% CI [.056, .085], for the one- and two-factor models respectively. Interestingly, the difference in model fit between one- and two-factor models was less in this new sample. This increases our confidence in retaining both the one-and two-factor solutions as plausible conceptualizations of the BSCS's structure.

With reasonable evidence in support of both one-factor (1F-CR) and two-factor (2FIC-CR) structures, we first conducted two analyses using the MGRM on Sample 1. The one-factor model was estimated using the MGRM so that residual covariances may be specified. It is not possible to compute covariances between item-level residuals under the IRT framework so instead, an additional factor is specified to account for residual covariances. Residual covariance factors must be orthogonal to all other factors in the model and the loadings from the two items whose residuals covary must be constrained to equality. These residual covariance factors are mathematically equivalent to residual covariances in the structural equation modeling framework and do not change the meaning or interpretation of parameter estimates or results. Technically, the 1F-CR

model was specified as a four-dimensional model (one primary factor and three residual covariances) and the 2FIC-CR was specified as a five-dimensional model (two primary factors and three residual covariances).

The slope (*a*) and intercept (*c*) parameters for the 1F-CR model of the BSCS are provided in Table 7. The item parameters for the 2FIC-CR model are provided in Table 8. In both tables, the slopes for items loading onto the residual covariance factors are denoted as *r*. Intercepts are reported instead of thresholds (*b*) because in higher dimensional models, thresholds lose their straightforward interpretation. However, thresholds of unidimensional items (i.e., items loading onto only one factor) maintain their convenient interpretation which we use below. Variability in the slopes indicates that the 13 items of the BSCS are differentially related to the construct of self-control. Using Items 9 and 4 as an example (regardless of which model), slopes are higher for the former than the latter. One interpretation is that Item 9 tells us more about an individual's level of self-control as compared with Item 4. Depicted in Figure 2, the trace lines for Item 9 are more peaked than the trace lines for Item 4 which gives a clearer picture of how individuals respond to Item 9. There is also variability in the thresholds. For example, the thresholds for Item 11, regardless of which model, are all less than the thresholds for Item 7. A respondent requires a higher level of self-control to endorse a particular response option for Item 7 than that same response option for Item 11. As may be seen graphically in Figure 3, higher thresholds shift the trace lines to the right of the θ continuum.

Information becomes unwieldy in higher dimensional models and thus, impossible to present in an easily interpretable manner. However, the

multidimensionality of the two substantive factors in the 2FIC-CR model could be estimated just as well with two, separate unidimensional IRT models. Because each item loads onto only one factor (i.e., between-item multidimensionality), the set of items for Factor 1 could be modeled in one unidimensional IRT model and the set of items for Factor 2 could be modeled in another. Second, the multidimensionality resulting from the residual covariance factors were specified to account for local dependence (LD). LD is said to occur when, conditional on the latent variables in the model, there is residual covariance between items. Put another way, there remains a correlation between items after controlling for the latent construct. Although it was necessary to model this dependency, the residual covariance factors themselves carried no substantive meaning. More on LD may be found in Chen and Thissen (1997) and Edwards, Houts, and Cai (2018). Assuming between-item multidimensionality, LD may also be accounted for by estimating a sequence of unidimensional IRT models with parameter constraints on locally dependent items.

The result is a unidimensional approximation to the multidimensional solution. This approach accounts for the nuisance multidimensionality while retaining the straightforward interpretability of a unidimensional model. Between the multidimensional solutions and the unidimensional approximations, the EAP scores were correlated 1 for the 1F-CR model, .99 for Factor 1 of the 2FIC-CR model, and .98 for Factor 2 of the 2FIC-CR model. Therefore, we used information from the unidimensional approximations for interpretation. These approximated information functions for the 13-item BSCS are plotted in Figure 4. The solid line represents information for the 1F-CR model, the long-dashed line represents information for Factor 1 of the 2FIC-CR model,

and the short-dashed line represents information for Factor 2 of the 2FIC-CR model. The thresholds dictate where the information functions peak and the slopes dictate the amount of information around the thresholds. If thresholds are higher, the information function will shift toward the right which indicates that the BSCS is more precise (i.e., informative) for respondents with higher levels of self-control. If slopes are higher, the BSCS will be more precise (i.e., informative), on average, across the range of self-control.

Examination of the information functions for the BSCS indicate that this scale has relatively even precision (i.e., is informative) across the range of self-control. This is the case for both the 1F-CR and 2FIC-CR models. On average, the BSCS experiences its most drastic drop in information toward higher levels of self-control. The scale remains relatively informative until about two standard deviations above the mean of θ. The largest deviation from this is Factor 2 of the 2FIC-CR model which becomes relatively uninformative at approximately 1.5 standard deviations above the mean. The scale functions less well for those with high levels of self-control. Information may also be used to compute the standard error of measurement. With an inverse relationship to information, standard errors may be calculated using the following conversion: 1/INF. Standard errors for the BSCS, similar to information, would be conditional on the level of self-control. This measure shows where the BSCS is more or less imprecise. Essentially, the information function and standard errors tell us the same information about the BSCS in different metrics.

Measures of internal consistency were computed for the BSCS using all 13 items for the one-factor solution. The items were split to compute internal consistency for the

two-factor solution. Items 1, 2, 5, 6, 7, 8, 12 comprised Factor 1 and Items 3, 4, 9, 10, 11, and 13 comprised Factor 2. For the one-factor model, coefficient alpha (Cronbach, 1951) was found to be .914 and coefficient omega (McDonald, 1970) was .915. Coefficients alpha and omega for Factor 1 of the two-factor model were .892 and .894, respectively. As for Factor 2 of the two-factor model, alpha was .819 and omega was .826.

Validation Evidence for the BSCS

When looking at the regression analysis with the validity measures at initial data collection (Table 9), the 1F-CR EAPs had statistically significant negative relationships with all measures except for frequency of alcohol use. This measure, however, was significantly predicted by both Factor 1 EAPs and Factor 2 EAPs from the 2FIC-CR model. The remaining models using the EAPs from the 2FIC-CR model had only one factor as a significant predictor or neither factor as a significant predictor. Both cognitive and behavioral impulsivity had a significant negative association with Factor 2 EAPs. In contrast, two of the count variables (number of alcoholic drinks per day and number of lifetime traffic accidents) were significantly and negatively predicted by Factor 1 EAPs. Last, number of lifetime arrests was not significantly predicted by either of the 2FIC-CR factors.

The regression analysis with the validity measures at retest (Table 10) provided similar results. Again, the 1F-CR EAPs were significantly and negatively predictive of all but one of the measures but this time, number of alcoholic drinks was nonsignificant. For the 2FIC-CR EAPs, none of the measures were significantly predicted by both Factor 1 and Factor 2 in the model simultaneously. Again, both cognitive and behavioral impulsivity had a significant negative association with Factor 2 EAPs. Number of

lifetime arrests and number of lifetime traffic accidents were significantly and negatively predicted by Factor 1 EAPs. Last, number of alcoholic drinks per day was not significantly predicted by Factor 1 or Factor 2 EAPs. Although there was slight variability in the results from initial data collection to retest, all results were in the hypothesized direction. The 1F-CR model significantly predicted the majority of outcomes, while the 2FIC-CR model exhibited differentially predictive factors. With the exception of one model, frequency or count outcomes were significantly predicted by Factor 1, whereas impulsivity was significantly predicted by Factor 2.

Discussion

The results from our psychometric evaluation support both a unidimensional and multidimensional factor structure of the BSCS. The one- and two-factor structure that emerged as most plausible were unique to the current study and retained all 13 items of the BSCS keeping the scale intact. The current study also provided a novel perspective on the BSCS via the IRT framework. In terms of item functioning, all 13 BSCS items were informative and each item contributed differentially to the measurement of self-control. As a measurement instrument overall, the BSCS functioned well and was informative across a wide range of self-control. The BSCS was least informative around and beyond 2 standard deviations above of the mean. If the scale were to be improved in terms of covering a wider range of self-control, items that are informative at higher levels of self-control could be added.

Calibration of the BSCS items under the IRT framework offered a new perspective as well as many benefits. As opposed to weighting the item-to-construct relationship equally as is assumed and done with summed scores, the item properties

provided by the IRT analysis account for the differential weighting of each individual BSCS item to the construct of self-control. The slopes and thresholds differ among the 13 items of the BSCS, which supports the notion that items contribute differently to the measurement of self-control. Once item properties are accounted for, researchers are able to make finer distinctions between individuals with IRT scale scores (e.g., EAPs). If equated, then scores may be directly compared when different forms of the scale are administered even if the forms have no items in common. This may be used in applications such as developing an adaptive version of the BSCS. For reliability, the IRT framework provided a practical alternative (i.e., information) to traditional measures. Instead of assuming that the BSCS is equally reliable across the continuum of self-control, we were able to assess the precision of the BSCS conditional on the level of the construct. Last, calibration of items only has to be carried out one time. Once calibrated, future research using the same population may use the item parameters estimated in this study and are not required to conduct additional IRT analyses.

The validation evidence is supportive, in part, of both a one-factor model and two-factor model. For the one-factor model, the BSCS scores were significantly predictive overall. With the two-factor BSCS, the factors were differentially related to outcomes. Specifically, there was often only one significantly predictive factor out of the two. Factor 1 tended to be predictive when an outcome was related to number of occurrences (count or frequency variables). Factor 2 tended to be predictive when an outcome was related to a trait (impulsivity). Which factor was predictive varied depending on the particular outcome. While the one-factor model seems to be sufficient for prediction of outcomes, in general, the two-factor model may be capable of making

finer distinctions—especially with different types of outcomes. This suggests that while interpretations may differ between the one- and two-factor models, both may be viable/of interest.

The current study has taken a step toward establishing a solid foundation for the BSCS and removing methodology as a source of variability in results as is common in the literature. A handful of studies have conducted psychometric evaluations on the BSCS but none have been able to agree on the factor structure. While our results are not definitive in terms of number of factors, we have successfully eliminated many alternatives leaving two factor structures to consider. It may be that either factor structure adequately represents the dimensionality of the BSCS and the decision between one or two factors depends more on the scale's intended use. The evidence is mixed regarding which model is preferable to use for future studies. On one hand, the one-factor model fits that data reasonably well and based on the outcomes examined here, the division into two factors may not add much statistical utility. On the other hand, the two-factor model fits better—if only trivially so—and the interfactor correlation is only .83 (from the MIRT analysis). Although this is a large correlation, it indicates that well over 30% of the variance in each factor is unique. Also, the extent that other validity evidence not considered here (e.g., expert opinion, qualitative work, etc.) supports one model over the other could prove decisive. There has also been disagreement on which items should be dropped from the scale. Our results showed that each item contributed information and suggests that all the items should be retained.

The variability in findings and conclusions regarding the BSCS is not surprising considering that each previous study used a different combination of methods.

Ultimately, this lack of consistency across studies has compromised a clear understanding of the BSCS's psychometric properties. The current study established a more stable conceptualization of the BSCS by: (a) compiling a set of optimal methods directly guided by the psychometric literature, (b) clearly reporting each method used (including computer programs and specifications) with relevant rationale, (c) implementing the novel set of methods, and (d) properly evaluating the results of the psychometric analyses.

Given that this is the first study to empirically support a one-factor model of the BSCS and a two-factor model with item mapping that differed from previous studies, subsequent replications of these models would add to the confidence of our one- and two-factor structures. Although our new two-factor model shared similarities with the De Ridder et al. (2011) model (inhibitory self-control vs. initiatory self-control), additional work would be needed to determine the substantive meaning of each factor (e.g., content analysis). Future work would also benefit from an evaluation of measurement invariance as was done by Morean et al. (2014). Now that the IRT framework has been introduced to the psychometric evaluation of the BSCS, there is an elegant approach for assessing measurement invariance: differential item functioning (DIF). DIF allows researchers to examine whether items function differently across groups. Responses to items in one group may not mean the same thing in another group. Common groups that DIF may be tested within include males versus females and clinical versus general. If DIF is detected, then differences in item parameters may be adjusted for which would avoid any potential bias in measurement.

Future studies of the BSCS would benefit from including measures of types other than just self-report. There is an expanding literature involving behavioral-performance and physiological response indicators which have been contributing to the creation of a multimethod measurement framework of self-control. Examples of studies that have incorporated such measures include Brennan and Baskin-Sommers (2018) and Venables et al. (2018). Although some recent evidence suggests a lack of correspondence between self-report and experimental measures (Eisenberg et al., 2018), behavioral-performance and physiological response indicators would be valuable in further validation work of the BSCS. In combination with self-report measures, a more comprehensive view and understanding of self-control would be achieved.

Other limitations include the MTurk sample and validity measures. Although MTurk has allowed for faster data collection at reasonable costs along with the advantage of more diverse samples, it is not free from flaws. Some issues include representativeness of samples, practice effects, and deception. These may even be exaggerated for incriminating measures such as some of the ones used for the current validity assessment. More on issues with MTurk samples can be found in Follmer, Sperling, and Suen (2017) and Paolacci and Chandler (2014). Although such issues could negatively affect the validity of a study, the consistency of results between our initial MTurk sample and university undergraduate sample have mitigated our concerns. As for the validity measures used in this study, none were representative of the positive outcomes highlighted in the introduction. This may limit generalizations based on the current results. However, since the majority of the BSCS items pertain to a lack of self-control, there could be an argument for more effective prediction of maladaptive outcomes.

Nonetheless, further work with the inclusion of positive outcomes would be an important contribution to the BSCS literature.

To conclude, this study features methods, procedures, and reporting practices for scale evaluation guided by the psychometric literature. Another difference in our approach was that, rather than fixating on a single "right" answer, we acknowledge that statistical models are approximations and it is possible—and perhaps should be more common—that we find more than one model is plausible. While this makes working with the BSCS potentially more complicated, the complexity is reflective of real complexities in the data, the scale, and potentially the construct(s). We hope future studies are able to explore these two different solutions and determine if one is to be preferred or, failing that, *when* one is to be preferred over the other.

References

Ackerman T. A., Gierl M. J., & Walker C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51.

Babinski L., Hartsough C., & Lambert N. (1999). Childhood conduct problems, hyperactivity-impulsivity, and inattention as predictors of adult criminal activity. *Journal of Child Psychology and Psychiatry*, 40, 347-355.

Bock R. D., & Mislevy R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psycho-logical Measurement*, 6, 431-444.

Brennan G. M., & Baskin-Sommers A. R. (2018). Brain-behavior relationships in externalizing: P3 amplitude reduction reflects deficient inhibitory control. *Behavioural Brain Research*, 337, 70-79.

Browne M. W., & Cudeck R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21, 230-258.

Browne M. W., Cudeck R., Tateneni K., & Mels G. (2010). CEFA: Comprehensive exploratory factor analysis, version 3.04 [Computer software and manual]. Retrieved from https://psychology.osu.edu/dr-browne-software

Byrne B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment*, 85, 17-32.

Cai L. (2017). *flexMIRT version 3.51: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

Cai L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33-57.

Cai L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35, 307-335.

Chen W., & Thissen D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.

Christoffersson A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40, 5-32.

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

Čubranić-Dobrodolac M., Lipovac K., Čičević S., & Antić B. (2017). A model for traffic accidents prediction based on driver personality traits assessment. *Promet (Zagreb)*, 29, 631-642.

De Ayala R. J. (2008). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Ridder D. T. D., De Boer B. J., Lugtig P., Bakker A. B., & Van Hooft E. A. J. (2011). Not doing bad things is not equivalent to doing the right thing: Distinguishing between inhibitory and initiatory self-control. *Personality and Individual Differences*, 50, 1006-1011.

DeVellis R. F. (2012). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

DeWalt D. A., Thissen D., Stucky B. D., Langer M. M., Morgan DeWitt E., Irwin D. E., . . . Varni J. W. (2013). PROMIS Pediatric Peer Relationships Scale: Development of a peer relationships item bank as part of social health measurement. *Health Psychology*, 32, 1093-1103.

Edwards M. C. (2009). An introduction to item response theory using the Need for Cognition Scale. *Social and Personality Psychology Compass*, 3, 507-529.

Edwards M. C., Houts C. R., & Cai L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, 23, 138-149.

Eisenberg I. W., Bissett P. G., Canning J. R., Dallery J., Enkavi A. Z., Whitfield-Gabrieli S., . . . Poldrack R. A. (2018). Applying novel technologies and methods to inform the ontology of self-regulation. *Behaviour Research and Therapy*, 101, 46-57.

Embretson S. E., & Reise S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Enkavi A. Z., Eisenberg I. W., Bissett P. G., Mazza G. L., MacKinnon D. P., Marsch L. A., & Poldrack R. A. (2019). A large-scale analysis of test-retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 5472-5477.

Fabrigar L. R., Wegener D. T., MacCallum R. C., & Strahan E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.

Ferrari J. R., Stevens E. B., & Jason L. A. (2009). The role of self-regulation in abstinence maintenance: Effects of communal living on self-regulation. *Journal of Groups in Addiction & Recovery*, 4, 32-41.

Flora D. B., & Curran P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.

Follmer D. J., Sperling R. A., & Suen H. K. (2017). The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher*, 46, 329-334.

Horn J. L. (1965). A rationale and technique for estimating the number of factors in factor analysis. *Psychometrika*, 30, 179-185.

Hu L., & Bentler P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.

Lindner C., Nagy G., & Retelsdorf J. (2015). The dimensionality of the Brief Self-Control Scale: An evaluation of unidimensional and multidimensional applications. *Personality and Individual Differences*, 86, 465-473.

MacCallum R. C., Widaman K. F., Zhang S., & Hong S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-99.

Maloney P. W., Grawitch M. J., & Barber L. K. (2012). The multi-factor structure of the Brief Self-Control Scale: Discriminant validity of restraint and impulsivity. *Journal of Research in Personality*, 46, 111-115.

McDonald R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1-21.

McNeish D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412-433.

Monroe S., & Cai L. (2015). Examining the reliability of student growth percentiles using multidimensional IRT. *Educational Measurement: Issues and Practice*, 34(4), 21-30.

Morean M. E., Demartini K. S., Leeman R. F., Pearlson G. D., Anticevic A., Krishnan-Sarin S., . . . O'Malley S. S. (2014). Psychometrically improved, abbreviated versions of three classic measures of impulsivity and self-control. *Psychological Assessment*, 26, 1003-1020.

Paolacci G., & Chandler J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23, 184-188.

Patton J. H., Stanford M. S., & Barratt E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, 51, 768-774.

Preston K. S. J., Gottfried A. W., Park J. J., Manapat P. D., Gottfried A. E., & Oliver P. H. (2018). Simultaneous linking of cross-informant and longitudinal data involving positive family relationships. *Educational and Psychological Measurement*, 78, 409-429.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reckase M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.

Reise S. P., Moore T. M., Sabb F. W., Brown A. K., & London E. D. (2013). The Barratt Impulsiveness Scale-11: Reassessment of its structure in a community sample. *Psychological Assess-ment*, 25, 631-642.

Rosseel Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/

Samejima F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Sharma L., Markon K. E., & Clark L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, 140, 374-408.

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

Tangney J. P., Baumeister R. F., & Boone A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72, 271-324.

Thissen D., & Steinberg L. (2009). Item response theory. In Millsap R. E., Maydeu-Olivares A. (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 148-177). London, England: Sage.

Thissen D., & Wainer H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.
Thurstone L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

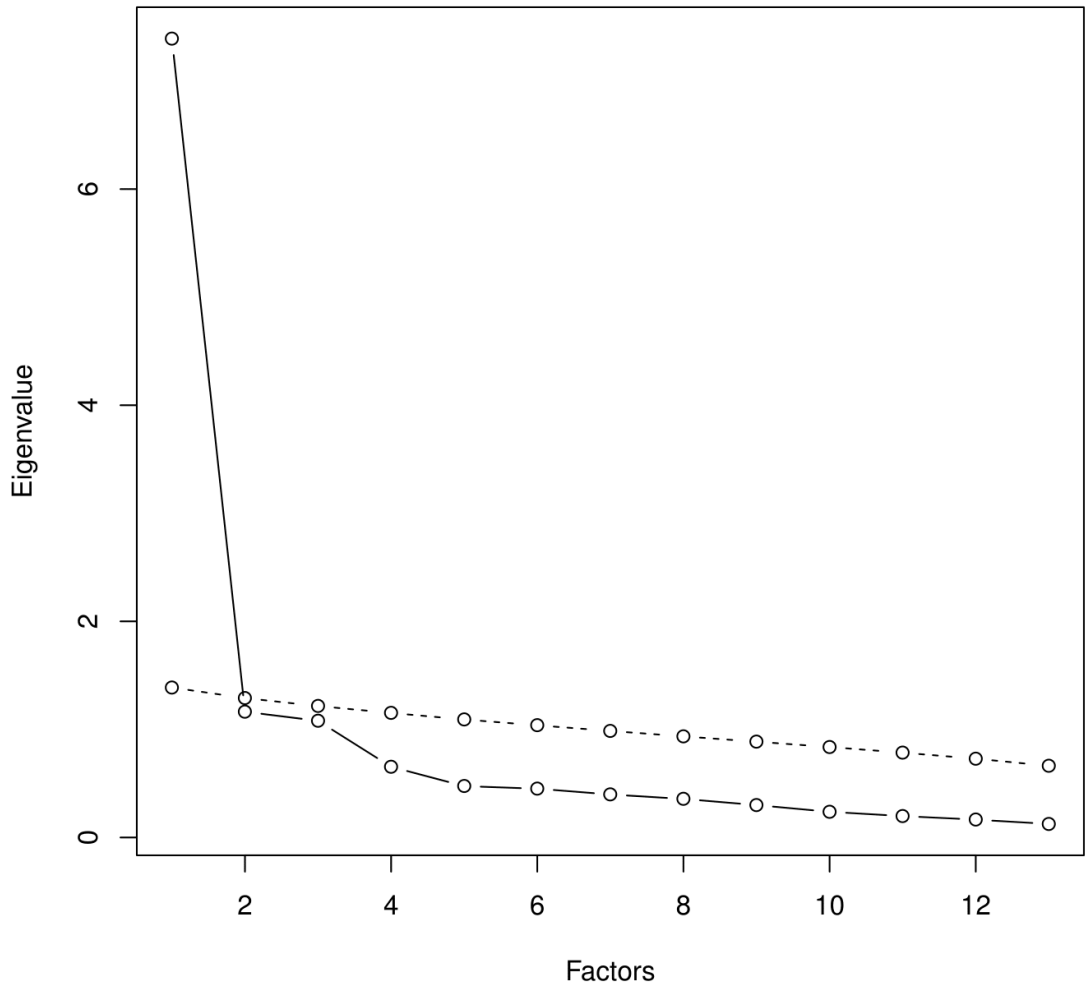
Venables N. C., Foell J., Yancey J. R., Kane M. J., Engle R. W., & Patrick C. J. (2018). Quantifying inhibitory control as externalizing proneness: A cross-domain model. *Clinical Psychological Science*, 6, 561-580.

Wirth R. J., & Edwards M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58-79.

Zhang Z., & Yuan K.-H. (2015). coefficientalpha: Robust coefficient alpha and omega with missing and non-normal data (R package version 0.5) [Computer software]. Retrieved from https://CRAN.R-project.org/package=coefficientalpha

Figure 1

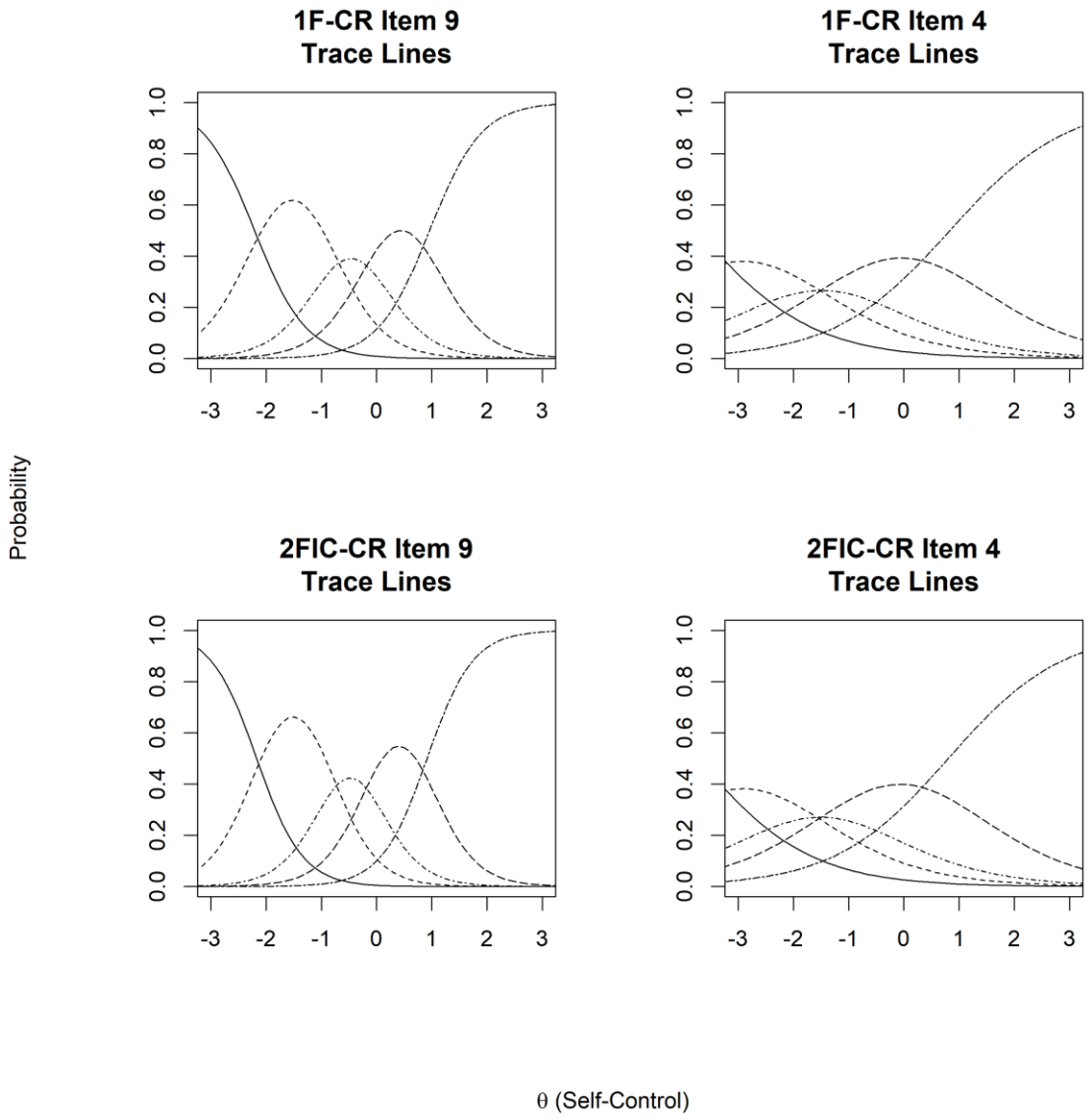
Parallel Analysis



Note. Solid line is scree plot from data (exploratory half of Sample 1; n = 261) and dotted line is random component from parallel analysis.

Figure 2

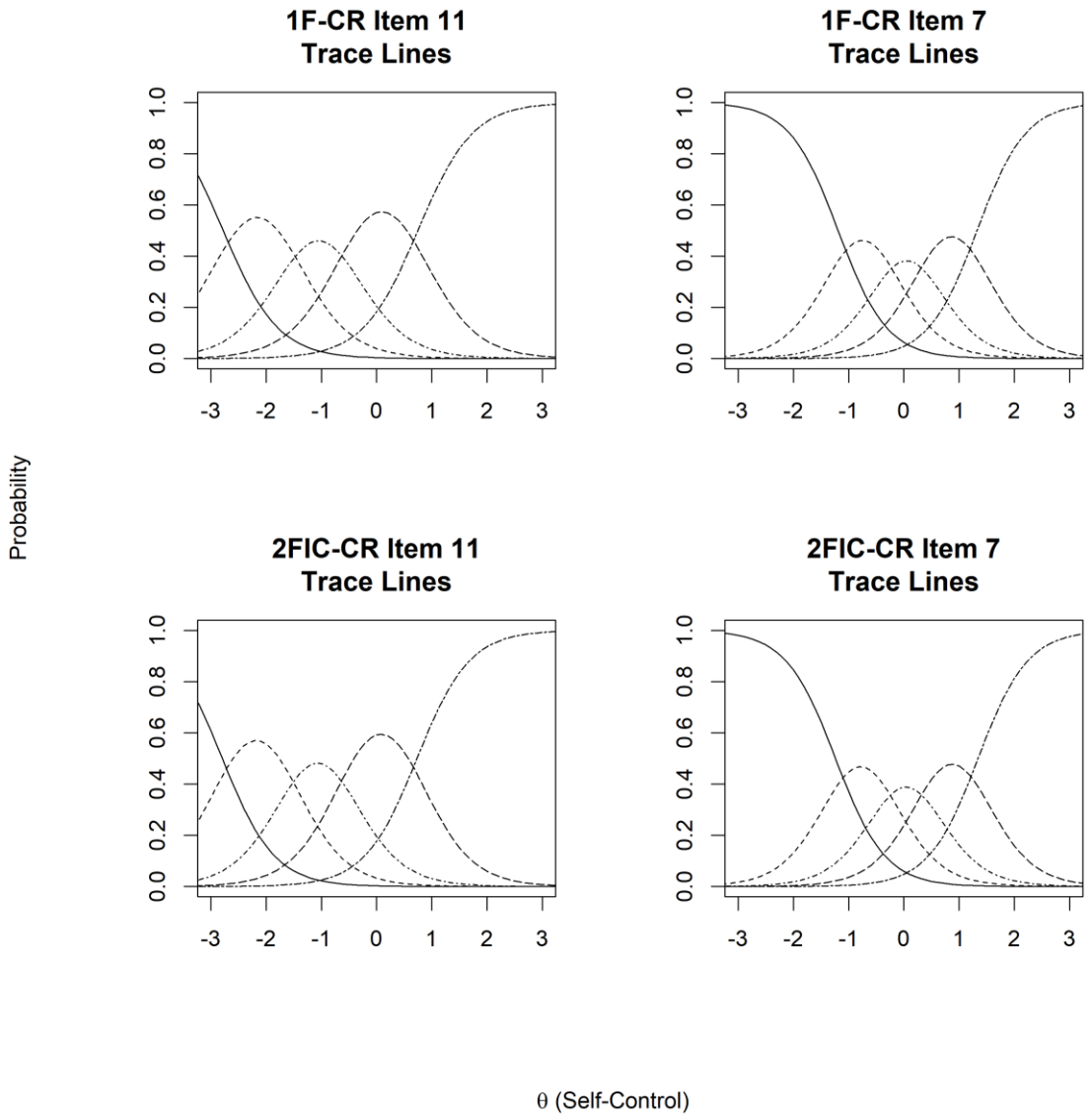
Trace Lines for Items 9 and 4 From the One-Factor (1F-CR) and Two-Factor (2FIC-CR) Models



Note. 1F-CR = 1F model with three residual covariances; 2FIC-CR = 2FIC model with three residual covariances.

Figure 3

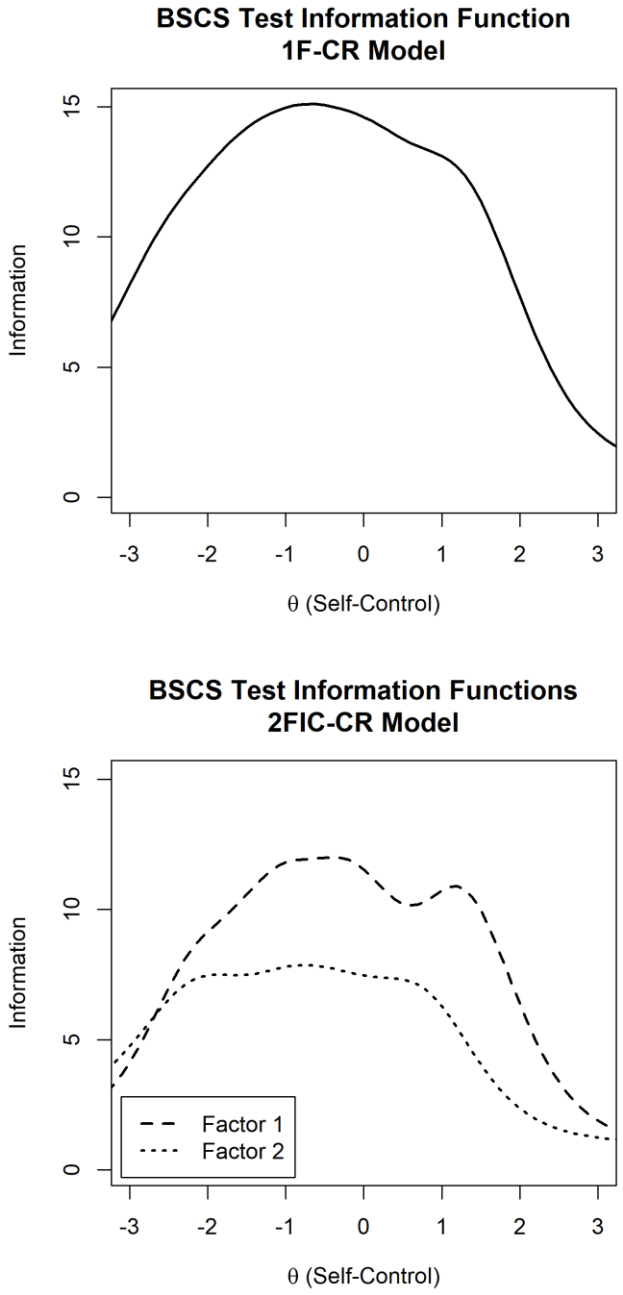
Trace Lines for Items 11 and 7 From the One-Factor (1F-CR) and Two-Factor (2FIC-CR) Models



Note. 1F-CR = 1F model with three residual covariances; 2FIC-CR = 2FIC model with three residual covariances.

Figure 4

Information Functions for the 13-Item BSCS



Note. BSCS = Brief Self-Control Scale; 1F-CR = 1F model with three residual covariances; 2FIC-CR = 2FIC model with three residual covariances. The top plot

represents the information function for the 1F-CR model. The bottom plot represents the information functions for the 2FIC-CR model. The long-dashed line represents information for Factor 1 of the 2FIC-CR model and the short-dashed line represents information for Factor 2 of the 2FIC-CR model.

Table 1

The 13-Item Brief Self-Control Scale (BSCS) Developed by Tangney et al. (2004)

#	Item	(+/-)
1	I am good at resisting temptation.	+
2	I have a hard time breaking bad habits.	-
3	I am lazy.	-
4	I say inappropriate things.	-
5	I do certain things that are bad for me, if they are fun.	-
6	I refuse things that are bad for me.	+
7	I wish I had more self-discipline.	-
8	People would say that I have iron self-discipline.	+
9	Pleasure and fun sometimes keep me from getting work done.	-
10	I have trouble concentrating.	-
11	I am able to work effectively toward long-term goals.	+
12	Sometimes I can't stop myself from doing something, even if I know it is wrong.	-
13	I often act without thinking through all the alternatives.	-

Note. Positively phrased items indicated by (+) and negatively phrased items indicated by (-). Rating scale ranging from 1 (*not at all like me*) to 5 (*very much like me*).

Table 2

Methods for Factor Analyzing the BSCS

Study	Method (Specifications)
Ferrari et al. (2009)	EFA (ML estimation; orthogonal varimax rotation)
De Ridder et al. (2011)	CFA (estimator not specified)
Maloney et al. (2012)	EFA (estimator not specified; oblique direct oblimin rotation) → CFA (ML estimation)
Morean et al. (2014)	CFA (robust ML estimation) → EFA (robust ML estimation; oblique varimax rotation)
Current study	EFA (OLS estimation; direct quartimin rotation) → CFA (DWLS/WLSMV estimation)

Note. BSCS = Brief Self-Control Scale; EFA = exploratory factor analysis; CFA = confirmatory factor analysis; ML = maximum likelihood; OLS = ordinary least squares; DWLS = diagonally weighted least squares; WLSMV = weighted least squares mean and variance adjusted. Items were dropped from De Ridder et al. (2011), Maloney et al. (2012), and Morean et al. (2014).

Table 3

Polychoric Correlation Matrix for the 13-Item BSCS (Sample 1; n = 522)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	.73	1											
3	.45	.58	1										
4	.33	.31	.41	1									
5	.53	.53	.40	.48	1								
6	.71	.63	.41	.36	.69	1							
7	.61	.68	.52	.27	.47	.54	1						
8	.74	.68	.50	.27	.49	.64	.63	1					
9	.51	.54	.61	.34	.58	.48	.58	.47	1				
10	.42	.55	.61	.35	.39	.37	.50	.40	.62	1			
11	.59	.61	.59	.27	.39	.51	.53	.53	.55	.59	1		
12	.58	.57	.45	.39	.65	.57	.59	.51	.58	.42	.45	1	
13	.53	.47	.44	.45	.56	.49	.50	.45	.62	.47	.47	.64	1

Note. BSCS = Brief Self-Control Scale.

Table 4

Exploratory Factor Analysis: Factor Loadings, Communalities, and Factor Correlations for the One- and Two-Factor Models of the 13-Item BSCS (Exploratory Half of Sample 1; n = 261)

	One-factor solution		Two-factor solution		
Item	Factor 1	Comm.	Factor 1	Factor 2	Comm.
1	.81	.66	.93		.80
2	.83	.69	.62		.69
3	.62	.39		.75	.52
4	.53	.29	–	–	.29
5	.76	.58	.64		.60
6	.74	.55	.89		.70
7	.76	.59	.43	.40	.58
8	.75	.57	.80		.65
9	.77	.60		.80	.72
10	.62	.39		.81	.57
11	.73	.53		.56	.56
12	.78	.60	.53		.60
13	.70	.49		.48	.49
Factor			Factor correlations		
1	1		1		
2			.68	1	

Note. BSCS = Brief Self-Control Scale; Comm. = communalities. For two-factor solution, an oblique direct quartimin rotation was used, factor loadings less than an absolute value of .40 were omitted, and emdashes (–) indicate that an item failed to clearly load onto either of the factors.

Table 5

Confirmatory Factor Analysis: Model Fit Summary

Model	χ^2	df	TLI	CFI	SRMR	RMSEA	90% CI
Exploratory half of sample 1 (n = 261)							
1F	306.95	65	.980	.983	.082	.120	[.106, .133]
2FCL	219.49	62	.986	.989	.068	.099	[.085, .113]
2FIC	231.26	64	.986	.989	.070	.100	[.087, .114]
1F-CR	242.40	62	.984	.988	.075	.106	[.092, .120]
2FCL-CR	159.54	59	.991	.993	.061	.081	[.066, .096]
2FIC-CR	169.97	61	.990	.993	.063	.083	[.068, .098]
Hold-out half of sample 1 (n = 261)							
1F-CR	187.41	62	.987	.989	.064	.088	[.074, .103]
2FIC-CR	140.29	61	.992	.993	.056	.071	[.055, .086]
Complete sample 1 (N = 522)							
1F-CR	374.96	62	.985	.988	.065	.098	[.089, .108]
2FIC-CR	254.09	61	.990	.992	.054	.078	[.068, .088]
Validation sample (Sample 2; N = 285 after listwise deletion)							
1F-CR	164.48	62	.981	.985	.065	.076	[.062, .091]
2FIC-CR	146.89	61	.984	.987	.061	.070	[.056, .085]

Note. TLI = Tucker-Lewis index; df = degrees of freedom; CFI = comparative fit index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CI = confidence interval; 1F = one-factor model; 2FCL = two-factor model with cross-loadings; 2FIC = two-factor model with independent clustering; 1F-CR = one-factor model with three residual covariances; 2FCL-CR = two-factor model with cross-loadings and three residual covariances; 2FIC-CR = two-factor model with independent clustering and three residual covariances.

Table 6

Confirmatory Factor Analysis: Factor Loadings and Factor Correlations for the Final One- and Two-Factor Models of the 13-Item BSCS (Complete Sample 1; N = 522)

	1F-CR	2FIC-CR	
Item	Factor 1	Factor 1	Factor 2
1	.81 (.02)	.83 (.02)	
2	.84 (.02)	.85 (.01)	
3	.70 (.02)		.74 (.02)
4	.48 (.04)		.51 (.04)
5	.71 (.02)	.73 (.02)	
6	.75 (.02)	.77 (.02)	
7	.76 (.02)	.78 (.02)	
8	.75 (.02)	.76 (.02)	
9	.76 (.02)		.81 (.02)
10	.68 (.03)		.72 (.03)
11	.72 (.03)		.76 (.03)
12	.74 (.02)	.75 (.02)	
13	.68 (.03)		.72 (.03)
Factor	Factor correlations		
1		1	
2		.85 (.02)	1

Note. BSCS = Brief Self-Control Scale; 1F-CR = one-factor model with three residual covariances; 2FIC-CR = two-factor model with independent clustering and three residual covariances. Standard errors are in parentheses. All factor loadings, factor correlations, and residual covariances were significant at $p < .001$.

Table 7

Multidimensional Graded Response Model Parameter Estimates for 1F-CR Model of the 13-Item BSCS (Sample 1; N = 522)

Slopes								
Item	a_1	*SE*	r_1	*SE*	r_2	*SE*	r_3	*SE*
1	3.05	0.18	1.44	0.09		----		----
2	2.77	0.18		----		----		----
3	1.79	0.13		----		----		----
4	0.95	0.10		----		----		----
5	2.25	0.15		----	1.40	0.09		----
6	2.58	0.16		----	1.40	0.09		----
7	2.26	0.15		----		----		----
8	2.59	0.15	1.44	0.09		----		----
9	2.13	0.14		----		----		----
10	1.69	0.13		----		----		----
11	2.01	0.15		----		----		----
12	2.49	0.16		----		----	1.13	0.09
13	2.09	0.15		----		----	1.13	0.09
Intercepts								
Item	c_1	*SE*	c_2	*SE*	c_3	*SE*	c_4	*SE*
1	7.06	0.43	3.34	0.21	0.65	0.15	-4.07	0.24
2	4.31	0.27	1.73	0.15	-0.24	0.13	-3.56	0.23
3	4.08	0.26	2.37	0.16	0.78	0.11	-1.22	0.12
4	3.56	0.24	1.96	0.13	0.87	0.10	-0.79	0.10
5	5.41	0.32	2.35	0.16	0.28	0.13	-2.99	0.19
6	5.29	0.30	2.12	0.16	-0.56	0.14	-4.46	0.26
7	2.69	0.17	0.69	0.12	-0.92	0.12	-2.99	0.19
8	3.05	0.18	0.32	0.14	-1.92	0.16	-4.75	0.27
9	4.70	0.30	1.81	0.14	0.16	0.11	-2.04	0.15
10	3.97	0.25	2.05	0.14	0.88	0.11	-1.18	0.12
11	5.58	0.40	3.10	0.19	1.11	0.12	-1.50	0.13
12	5.89	0.36	3.06	0.19	1.35	0.14	-1.95	0.16
13	5.97	0.40	3.61	0.21	2.22	0.16	-0.88	0.13

Note. BSCS = Brief Self-Control Scale; *SE* = standard error; 1F-CR = one-factor (a_1) model with three residual covariances (r_1-r_3). Intercepts are c_1-c_4.

Table 8

Multidimensional Graded Response Model Parameter Estimates for 2FIC-CR Model of the 13-Item BSCS (Sample 1; N = 522)

Slopes										
Item	a_1	*SE*	a_2	*SE*	r_1	*SE*	r_2	*SE*	r_3	*SE*
1	3.17	0.21		----	1.13	0.12		----		----
2	3.02	0.21		----		----		----		----
3		----	2.08	0.16		----		----		----
4		----	0.97	0.10		----		----		----
5	2.14	0.14		----		----	1.37	0.11		----
6	2.73	0.17		----		----	1.37	0.11		----
7	2.23	0.15		----		----		----		----
8	2.59	0.17		----	1.13	0.12		----		----
9		----	2.42	0.18		----		----		----
10		----	2.07	0.15		----		----		----
11		----	2.11	0.16		----		----		----
12	2.47	0.17		----		----		----	1.34	0.10
13		----	2.15	0.17		----		----	1.34	0.10
Intercepts										
Item	c_1	*SE*	c_2	*SE*	c_3	*SE*	c_4	*SE*		
1	7.24	0.46	3.42	0.23	0.70	0.15	-4.07	0.27		
2	4.73	0.31	1.94	0.18	-0.22	0.14	-3.83	0.27		
3	4.55	0.29	2.67	0.18	0.91	0.12	-1.32	0.13		
4	3.63	0.24	2.02	0.14	0.91	0.10	-0.78	0.10		
5	5.38	0.32	2.35	0.16	0.30	0.13	-2.92	0.19		
6	5.58	0.32	2.27	0.17	-0.54	0.14	-4.64	0.28		
7	2.76	0.18	0.73	0.12	-0.91	0.13	-2.99	0.20		
8	3.03	0.19	0.35	0.14	-1.84	0.16	-4.61	0.27		
9	5.25	0.35	2.06	0.17	0.25	0.12	-2.21	0.17		
10	4.53	0.29	2.38	0.17	1.04	0.13	-1.31	0.13		
11	5.89	0.43	3.30	0.22	1.20	0.13	-1.54	0.14		
12	6.22	0.39	3.24	0.20	1.46	0.16	-2.00	0.17		
13	6.43	0.43	3.94	0.23	2.43	0.17	-0.90	0.14		

Note. BSCS = Brief Self-Control Scale; *SE* = standard error; 2FIC-CR = two-factor (a_1-a_2) model with independent clustering and three residual covariances (r_1-r_3). The interfactor correlation was equal to .83.

Table 9

Regression of Validation Measures From Initial Data Collection Onto EAP Scores (Sample 1)

1F-CR EAP Scores						
Outcome	Predictor	N	R^2/AIC	*b*	*SE*	*t*-value
Frequency of Alcohol Use	Single-Factor EAP	522	.01	-0.10	0.05	-1.86
# of Alcoholic Drinks Per Day	Single-Factor EAP	522	1417.2	-0.13	0.04	-3.56*
Cognitive Impulsivity	Single-Factor EAP	522	.51	-6.29	0.27	-23.12*
Behavioral Impulsivity	Single-Factor EAP	522	.42	-4.43	0.23	-19.41*
# of Lifetime Arrests	Single-Factor EAP	521	970.93	-0.29	0.07	-3.85*
# of Lifetime Traffic Accidents	Single-Factor EAP	521	1570.7	-0.12	0.04	-2.74*
2FIC-CR EAP Scores						
Outcome	Predictor	N	R^2/AIC	*b*	*SE*	*t*-value
Frequency of Alcohol Use	Factor 1 EAP	522	.02	-0.40	0.13	-3.19*
	Factor 2 EAP			0.32	0.13	2.49*
# of Alcoholic Drinks Per Day	Factor 1 EAP	522	1415.6	-0.24	0.09	-2.85*
	Factor 2 EAP			0.12	0.09	1.35
Cognitive Impulsivity	Factor 1 EAP	522	.55	0.39	0.62	0.62
	Factor 2 EAP			-6.93	0.63	-10.96*
Behavioral Impulsivity	Factor 1 EAP	522	.44	-0.43	0.54	-0.81
	Factor 2 EAP			-4.13	0.55	-7.52*
# of Lifetime Arrests	Factor 1 EAP	521	972.86	-0.21	0.17	-1.25
	Factor 2 EAP			-0.08	0.17	-0.44
# of Lifetime Traffic Accidents	Factor 1 EAP	521	1568.6	-0.26	0.10	-2.63*
	Factor 2 EAP			0.14	0.10	1.44

Note. EAP = expected a posteriori score; *SE* = standard error; 1F-CR = one-factor model with three residual covariances; 2FIC-CR = two-factor model with independent clustering

and three residual covariances. Poisson regression used for all count outcomes (AIC reported for Poisson regression instead of R^2). *$p < .05$.

Table 10

Regression of Validation Measures from Retest Onto EAP Scores (Sample 1)

1F-CR EAP Scores						
Outcome	Predictor	N	R^2/AIC	*b*	*SE*	*t*-value
Frequency of Alcohol Use	Single-Factor EAP	150	.03	-0.24	0.10	-2.30*
# of Alcoholic Drinks Per Day	Single-Factor EAP	150	399.24	-0.14	0.07	-1.85
Cognitive Impulsivity	Single-Factor EAP	150	.53	-6.44	0.50	-12.85*
Behavioral Impulsivity	Single-Factor EAP	150	.43	-4.46	0.42	-10.59*
# of Lifetime Arrests	Single-Factor EAP	150	281.05	-0.38	0.16	-2.43*
# of Lifetime Traffic Accidents	Single-Factor EAP	150	453.99	-0.29	0.09	-3.25*
2FIC-CR EAP Scores						
Outcome	Predictor	N	R^2/AIC	*b*	*SE*	*t*-value
Frequency of Alcohol Use	Factor 1 EAP	150	.06	-0.66	0.27	-2.47*
	Factor 2 EAP			0.44	0.27	1.62
# of Alcoholic Drinks Per Day	Factor 1 EAP	150	401.19	-0.03	0.18	-0.19
	Factor 2 EAP			-0.10	0.18	-0.56
Cognitive Impulsivity	Factor 1 EAP	150	.55	-0.56	1.26	-0.44
	Factor 2 EAP			-6.06	1.28	-4.75*
Behavioral Impulsivity	Factor 1 EAP	150	.44	-1.40	1.08	-1.29
	Factor 2 EAP			-3.13	1.10	-2.84*
# of Lifetime Arrests	Factor 1 EAP	150	280.24	-0.75	0.35	-2.20*
	Factor 2 EAP			0.39	0.35	1.09
# of Lifetime Traffic Accidents	Factor 1 EAP	150	451.68	-0.58	0.20	-2.83*
	Factor 2 EAP			0.30	0.21	1.43

Note. EAP = expected a posteriori score; *SE* = standard error; 1F-CR = one-factor model with three residual covariances; 2FIC-CR = two-factor model with independent clustering and three residual covariances. Poisson regression used for all count outcomes (AIC

reported for Poisson regression instead of R^2). Retest measures collected a mean of 111 days after initial data collection. *$p < .05$.

CHAPTER 3

A REVISED AND EXPANDED TAXONOMY FOR UNDERSTANDING HETEROGENEITY IN RESEARCH AND REPORTING PRACTICES

Abstract

Concerns about replication failures can be partially recast as concerns about excessive heterogeneity in research results. Although this heterogeneity is an inherent part of science (e.g., sampling variability; studying different conditions), not all heterogeneity results from unavoidable sources. In particular, the flexibility researchers have when designing studies and analyzing data adds additional heterogeneity. This flexibility has been the topic of considerable discussion in the last decade. Ideas, and corresponding phrases, have been introduced to help unpack researcher behaviors, including researcher degrees of freedom and questionable research practices. Using these concepts and phrases, methodological and substantive researchers have considered how researchers' choices impact statistical conclusions and reduce clarity in the research literature. While progress has been made, inconsistent, vague, and overlapping use of the terminology surrounding these choices has made it difficult to have clear conversations about the most pressing issues. Further refinement of the language conveying the underlying concepts can catalyze further progress. We propose a revised, expanded taxonomy for assessing research and reporting practices. In addition, we redefine several crucial terms in a way that reduces overlap and enhances conceptual clarity, with particular focus on distinguishing practices along two lines: research versus reporting practices and choices involving multiple empirically supported options versus choices known to be subpar. We illustrate the effectiveness of these changes using conceptual and simulated demonstrations, and we discuss how this taxonomy can be valuable to substantive researchers by helping to navigate this flexibility and to methodological researchers by motivating research toward areas of greatest need.

Introduction

Replication is a fundamental aspect of scientific progress. A successful replication study provides supportive evidence for research findings, which bolsters trust and enables future research to systematically build upon original studies. However, until recently, replication studies have been limited in psychology (Makel et al., 2012), which has in turn limited the development of fundamental definitional and methodological aspects of replicability. In the spirit of a renewed interest in replication and increased scrutiny of methodological flexibility (e.g., Shrout & Rodgers, 2018), we detail a novel taxonomy for assessing research and reporting practices that add heterogeneity to research literatures and make it difficult to understand whether and when an effect has replicated. The goals of this taxonomy are to: (a) help researchers understand the different sources of heterogeneity across studies, (b) provide a structure for organizing and evaluating research and reporting practices that contribute to avoidable heterogeneity, (c) help researchers navigate the complex series of decisions made among acceptable and unacceptable practices during the research process, and (d) drive research toward domains where more methodological guidance on these decisions is needed.

We intend for this taxonomy to be widely applicable to flexibility in multiple methodological areas, including measurement, research design, statistical analysis, and reporting. We note here that research (mal)practices also include nonmethodological aspects such as improper credit in writing, sloppy conduct of research, failure to publish relevant results, authorship disputes, poor mentoring, failure to share data, and improper recordkeeping. However, we limit our current focus to methodological topics, which comprise a broad, critical category of research practice for which: (a) there is room for

improvement and (b) conversation among methodologists can promote such improvement. We first review research on the replication crisis and research practices before describing the taxonomy. Following a detailed description of our taxonomy, we show its organizational advantages with conceptual examples and a simulated demonstration. We end with a discussion of where our taxonomy fits into the broader literature on replication, research practices, and metascience.

The Crisis of Confidence

In the past decade, replication has received increased attention due to heightened awareness of replication failures and heavy scrutiny of research practices in general (see Fanelli, 2009; Ioannidis, 2005; Lindsay, 2015; Martinson et al., 2005). Unfortunately, replication success rates across a wide range of scientific fields continue to be underwhelming, which has cast doubt on the trustworthiness of published findings (e.g., clinical trials, epidemiology, molecular research; Ioannidis, 2005). For example, a survey of 1,576 researchers published in *Nature* found that 70% failed to replicate another researcher's work at least once and over 50% failed to replicate their own work at least once (Baker, 2016). Other bleak findings were shown in an effort by the pharmaceutical company Bayer to replicate medication trials (Mullard, 2011). Without scientific reform and a better understanding of replication, high failure rates such as these will continue to jeopardize the credibility of and support for science (Martinson et al., 2005; Pittinsky, 2015; Rutjens et al., 2018).

Amid the replication crisis, the field of psychology received a great deal of criticism. Parallel to the story for science in general, the low replication rates observed in psychology potentially limited confidence in findings across the field (Pashler &

Wagenmakers, 2012; Wingen et al., 2020). For example, high-profile failed replications of studies assessing ESP and priming were reported (e.g., Doyen et al., 2012; Galak et al., 2012), and only three out of 14 replication attempts were successful in a special issue of the journal *Social Psychology* (Nosek & Lakens, 2014). Responses to the crisis received even more momentum following a large-scale study published in *Science* that reported a replication success rate of 36% based on agreement in statistical significance for a sample of published work in top-tier psychology journals (Open Science Collaboration, 2015). And more recently, only about half of the replication attempts in the Many Labs 2 project agreed with the original study in both statistical significance and direction (Klein et al., 2018).

As reports of replication failures grew, discussions about replication began to shift from standard reporting of replication results to metascience-focused commentaries about the nuances behind high failure rates and what should and should not be expected of replication studies. Topics of these commentaries included the role of statistical power (Anderson & Maxwell, 2017; Hedges & Schauer, 2019; Maxwell et al., 2015), overreliance on statistical significance, both generally and in measuring replication (Anderson, 2020; Anderson & Maxwell, 2016; McShane, Gal, et al., 2019; Patil et al., 2016; Schauer & Hedges, 2021; Verhagen & Wagenmakers, 2014), and the failure to consider different sources of heterogeneity among effects (Kenny & Judd, 2019; McShane & Böckenholt, 2014), in addition to more procedural aspects of replication research (Brandt et al., 2014). These writings prompted a more complete and effective treatment of replication (Shrout & Rodgers, 2018) as well as specific

suggestions for procedural, statistical, and editorial improvements (see McNutt, 2014; Nosek et al., 2015; Wicherts et al., 2016).

One area of confusion when interpreting replication studies relates to heterogeneity across results and how to determine when two or more effects are "similar enough" to be considered successful replications, especially when there are multiple sources of heterogeneity potentially obscuring similarities across results. This idea is in keeping with Serlin and Lapsley's (1993) good-enough principle. For example, Many Labs 2 reported significant levels of heterogeneity across multisite replications of 40% of the 28 original effects studied, though little of this could be attributed to differences in samples, settings, or cultures (Klein et al., 2018). Importantly for our purposes, there is a distinction to be made between unavoidable sources of heterogeneity and (potentially) avoidable sources of heterogeneity. Unavoidable heterogeneity across replications is reasonable to expect, should be accounted for (Kenny & Judd, 2019; McShane & Böckenholt, 2014), and perhaps even embraced (McShane, Tackett, et al., 2019). Avoidable heterogeneity, however, has the potential to weaken trust in study results and stall progress (Silberzahn et al., 2018). Efforts to diminish sources of avoidable heterogeneity when possible would help with obtaining more accurate impressions of a series of research results while minimizing extraneous heterogeneity across replications. Despite an emerging literature on heterogeneity in the research literature, there remains an opportunity to address one source of such avoidable heterogeneity: flexibility in research practices, a topic we focus on and turn to next.

Our goal for the current article is to address the data analytic gray area in research practices (e.g., *p*-hacking, questionable research practices, researcher degrees of

freedom), which contributes to heterogeneity in methods and can yield avoidable heterogeneity in results. When analyzing data, researchers are required to sift through, evaluate, and choose among numerous, often subjective methodological options. Different choices, regardless of whether choices are defensible or not, can produce different results (e.g., Gelman & Loken, 2014; Klein et al., 2018; Silberzahn et al., 2018). One particularly problematic example of methodological flexibility occurs when presumed publication criteria (e.g., statistical significance) are prioritized over options that adhere to best practices. This methodological flexibility interacts with pressures to publish and increases the risk of false-positive findings (Simmons et al., 2011). Moreover, explicit *p*-hacking has been shown to interact with publication bias and yield biased literatures, as evidenced by meta-analysis, although studies differ in what the extent of the bias is deemed to be (Friese & Frankenbach, 2020; Head et al., 2015).[1] Although there has been a great deal of recent interest in flexible research practices (e.g., Anderson, 2020; Gelman & Loken, 2014; John et al., 2012; Lindsay, 2015; Martinson et al., 2005; Simmons et al., 2011; Wicherts et al., 2016), there is a need to refine the extant language surrounding methodological flexibility to help researchers navigate the complexities in choosing among analytic strategies and to clarify areas where more research and methodological guidance are needed.

[1] The term "*p*-hacking" has been used in various ways in the literature. We use *p*-hacking to refer specifically to selecting only options that move the *p*-value toward or under the significance level, checking for significance at each step. Our forthcoming taxonomy will cover a broader array of research practices, thus using more general terms, but *p*-hacking in this sense can be thought of as a specific type of questionable research practice under our taxonomy.

Methodological Flexibility

In the metascience literature, researcher degrees of freedom (RDFs; e.g., Simmons et al., 2011) and questionable research practices (QRPs; e.g., John et al., 2012) have emerged as common terms to describe flexibility in methodological practices that can have intended or unintended consequences on replication. The terms RDF and QRP were coined independently and not necessarily meant to be directly compared or contrasted. Not surprisingly, given that both the Simmons et al. (2011) and John et al. (2012) articles referred to methodological flexibility and the overlap among some of the specific practices described, the terms RDFs and QRPs are often used interchangeably. Because of this blurring of terms, there is an opportunity to gain more utility when describing or categorizing RDFs and QRPs by clarifying their definitions and purposefully distinguishing between them. We draw from the existing literature on research practices to construct concise and novel definitions for both RDFs and QRPs. The goal of elaborating on previously established definitions is to allow for clearer methodological guidance, thus reducing ambiguity.

In the research practices literature, RDFs have been described as choices among acceptable methodological alternatives (Simmons et al., 2011). Essentially, RDFs come into play when a researcher is required to make a choice from a set of justifiable options. Wicherts et al. (2016) provide examples including choices about which variables to include as covariates, how to handle violations of statistical assumptions, and which statistical model best suits a research question. Often, there is an honest "ambiguity in how best to make these decisions" (Simmons et al., 2011, p. 1359). QRPs, on the other hand, commonly describe choices that tend to stray from best practices. This corresponds

with definitions penned by the National Academy of Sciences, National Academy of Engineering, and Institute of Medicine (1992): QRPs are "actions that violate traditional values of the research enterprise and that may be detrimental to the research process" (p. 28). John et al. (2012) give examples of QRPs such as reporting *p*-values greater than .05 as less than .05 and excluding data after looking at the results. Although exploitation of either RDFs and QRPs can increase the risk of false-positives (John et al., 2012; Simmons et al., 2011), we posit that RDFs represent acceptable practices whereas QRPs represent unacceptable practices.

Modified Definitions: The Basis for a New Taxonomy

Building off this qualitative difference that considers QRPs as substandard compared with RDFs, we propose more precise, modified definitions that distinguish between RDFs and QRPs. We propose that RDFs refer to research practices selected from a set of viable/supported research practices. In contrast, we propose that QRPs refer to research practices that are known to be wrong or suboptimal. Along with these definitions, we consider three caveats to be of note.

Caveat 1: Although the appropriate category depends on the weight of evidence for or against the practice in question, whether a practice is an RDF or a QRP is often context dependent. Caveat 2: The appropriate category for a practice is not set in stone. If new research provides clearer evidence against certain decisions, practices once considered RDFs can be later considered QRPs. Caveat 3: The difference between RDFs and QRPs does not solely depend on the severity of impact on results. In some cases, RDFs may yield greater heterogeneity than QRPs because methodological research has not yet clearly identified what options are optimal in such situations. Moreover, the RDF-

QRP distinction also rests on whether enough information has been acquired to make defensible decisions in the first place and failing to do this would be categorized as a QRP. For example, failing to check assumptions could have minimal impact if the assumptions are unknowingly satisfied, but doing this leaves a researcher unable to properly address violations of these assumptions. In other words, acceptable practice is contingent on both acquiring all necessary/relevant information and taking a defensible action (see Table 1).

These new definitions can be useful for determining whether practices are acceptable (RDFs, which still may have unintended consequences) or should be clearly avoided (QRPs). We provide three examples below to highlight aspects of our modified definitions before building the taxonomy in more detail.

As a first example, Wicherts et al. (2016) lists choices among different statistical software programs and estimation methods as RDFs. This coincides with recent literature investigating the level of heterogeneity resulting from use of ostensibly the same procedure across different software packages (e.g., McCoach et al., 2018; Oster & Hilbe, 2008; Wang & Johnson, 2019). For more straightforward cases, different choice of software and/or estimation method may have negligible to no impact on results. For example, for a multiple linear regression with complete data, using SPSS with ordinary least squares (OLS) estimation yields similar (if not identical) results to using SAS PROC CALIS with full information maximum likelihood (FIML) estimation. These options would be RDFs by our proposed definitions. Even in areas where differences among software packages have been shown (e.g., different algorithmic

implementations; Wicherts et al., 2016), choosing among procedures with approximately equal supporting evidence would remain an RDF (see Caveat 3 above).

However, different software and/or estimation method choices may be better reflected as QRPs for more complex, but common cases where certain choices have been shown to be inferior (see Caveat 1 above). For example, if missing data are present on predictor variables, the same multiple regression analysis performed in SPSS (OLS estimation) and SAS PROC CALIS (FIML estimation) now would yield differing results (e.g., Allison, 2012). Unless a missing data procedure (e.g., multiple imputation) is employed, in this latter scenario, estimating with OLS could yield substantial bias and incorrect results of significance tests. This would be considered a QRP (i.e., suboptimal choice) by our definition.

Second, consider diagnostics for influential observations when conducting a regression analysis. There is strong consensus in the methodological literature that influential observations have the potential to drastically alter the outcome of an analysis, which is why diagnostics are universally recommended (Agresti & Franklin, 2007; Cohen et al., 2003). Because conducting diagnostic analyses is optimal in all situations (at least within the regression framework), we refer to such a practice as the single best option (SBO). However, the choice of diagnostic approach among a set of valid alternatives would be considered an RDF by our definition. For example, when considering whether to use DFFITS or Cook's D, the evidence in the methodological literature supports both options but does not clearly favor one over the other. Both global measures of influence are closely related and likely to yield similar results. However, as noted in Caveat 3 above, RDFs may not always yield similar results. Failing to run diagnostics before an

analysis, on the other hand, would be considered a QRP by our definition given the weight of evidence for the detrimental effects of ignoring influential observations.

Third, there are different ways to address potential violations of the sphericity assumption when conducting repeated measures ANOVA. Suppose a researcher infers that sphericity may be violated (e.g., via Mauchly's test or previous knowledge of the variables involved). The researcher could ignore the evidence, rely on output from the standard mixed model (univariate) approach, and risk Type-I error inflation (e.g., Hertzog & Rovine, 1985). Clearly, this is discouraged and would be classified as a QRP. Alternatively, the researcher could respond by reporting the results associated with a degrees of freedom adjustment such as Geisser-Greenhouse (Geisser & Greenhouse, 1958). Although none of these adjustments reflect a perfect solution, using one of these could justifiably represent an RDF. Finally, there is a third potential response: The researcher could use a multivariate or multilevel approach to more appropriately model the covariance structure (Maxwell et al., 2018), an RDF which comes close to an SBO. Of course, additional research on any of these topics could result in practices once considered RDFs moving to the QRP category (see Caveat 2 above).

A New Taxonomy for Assessing Research and Reporting Practices

We have suggested modified definitions for RDFs and QRPs that capitalize on a useful distinction. In the subsequent sections, we develop a new taxonomy for the assessment of research and reporting practices that is based on this RDF/QRP distinction and expands upon it in several constructive ways. This new taxonomy is based on the Steneck (2006) framework for defining research practices, which consists of three categories. On one end, there is responsible conduct of research which represents ideal

behavior. On the other end, there is fabrication, falsification, and plagiarism which represent worst behavior. The last category, QRPs, falls in between ideal and worst behavior. We begin by separating research practices from reporting practices and then apply this split to the Steneck (2006) framework. With this revised base, we then extend the framework to build a more precise and comprehensive taxonomy for assessing research and reporting practices (see Figure 1).

Separating Research Practices From Reporting Practices

The highest level of our proposed taxonomy separates research practices from reporting practices. The impetus for this separation comes from a debate found primarily in the literature on questionable practices. Some researchers have characterized QRPs broadly as any practice that is questionable in nature (John et al., 2012; Sijtsma et al., 2016; Steneck, 2006). For example, John et al. (2012) listed both research (e.g., collecting additional data after seeing if results were significant or not) and reporting practices (e.g., failing to report all dependent measures) as QRPs. Other researchers distinguish between research and reporting practices but restrict the focus to reporting. For example, Wigboldus and Dotsch (2016) argued that what makes a research practice questionable is not the practice itself but how the practice is reported and suggested modifying QRPs to mean questionable reporting practices instead of questionable research practices. These authors added that research practices for analyzing data cannot be deemed questionable.

While we agree that thorough and streamlined reporting practices are crucial for enabling critical peer review, maintaining a complete record of progress in a research area, and decreasing barriers to reproducibility and replication, we see value in

maintaining a distinction between research and reporting and allowing for practices in both arms to be deemed questionable if evidence shows them to be suboptimal. Reporting practices, which often involve issues of transparency, are the pipeline through which information about research practices flow from the researcher to the reader. A complete and effective description of the choices made for an analysis (reporting practices) is different from the actual choices that were employed (research practices), and questionable decisions can be made in both arenas.

Consider two researchers who plan to conduct the same statistical analysis with samples drawn from the same population. Upon inspection of the data, both researchers find strong evidence for a violation of a statistical assumption. Suppose that Researcher A does nothing to address the violation and runs the standard analysis. They do, however, clearly report ignoring the violation. Now suppose that Researcher B switches from running the standard analysis to running an analysis that is more robust to violations of statistical assumptions. However, Researcher B fails to report the switch. Researcher A's study could lead to incorrect inferences drawn from the focal results of the study, although readers/replicators would have been alerted to the assumption violation, whereas Researcher B arrives at more statistically valid inferences but makes it difficult for other researchers to appropriately replicate the study or understand why replications are producing different results.

Because each scenario causes different problems that call for different solutions, we see value in separating research practices from reporting practices. This is reflected in our proposed taxonomy, which is comprised of a framework for assessing research practices as well as a separate framework for assessing reporting practices. We

reorganize the Steneck (2006) framework to align with this research/reporting split by detaching responsible conduct of research and QRPs from fabrication, falsification, and plagiarism. Responsible conduct of research and QRPs become categories under the framework for research practices whereas fabrication and falsification become categories under the framework for reporting practices. Plagiarism is excluded from the taxonomy (see our rationale in the A New Framework for Reporting Practices section).

A New Framework for Research Practices

The first part of our new taxonomy focuses on research practices (left side of Figure 1). Under this framework, practices are first categorized as either responsible or questionable. A practice is considered responsible conduct of research if it meets the highest standards of research ethics/moral principles and research integrity/professional standards (i.e., best practices; Steneck, 2006). Conversely, a practice constitutes a QRP if it lacks empirical support or if the evidence weighs against its use. Under this framework, responsible conduct of research and QRPs are mutually exclusive, where practices for a specific research decision, by definition, are one or the other. If a practice falls under responsible conduct of research, then it may be classified as one of two subcategories: SBO or RDF. As previously defined, the SBO is an option the evidence suggests is optimal in all situations (leaving no need to make a choice). Alternatively, RDFs reflect a set of empirically supported practices where there is still a reasonable level of ambiguity in which option to prefer.

We build this new taxonomy to acknowledge the real-world complexity of the scientific literature. When considering a particular methodological option, one must examine the existing empirical/theoretical literature supporting that method (or, even

better, a convenient summary of various states of evidence among competing options). If the evidentiary literature weighs heavily against the option in question, then such an option is considered a QRP and should not be used. If an option is not a QRP, then it is classified as responsible conduct of research. More commonly, responsible conduct will require choosing among multiple viable options that are valid in at least similar circumstances (e.g., choosing between DFFITS or Cook's D as in our previous example). Recommendations for RDFs may vary by situation and sometimes it may not be clear why recommendations vary. In some circumstances, responsible conduct may not require a choice because there is an SBO. If a methodological option is the SBO (e.g., checking for influential observations before a regression analysis), this option should always be used and anything else is a QRP.

Of course, the burden should not fall solely upon substantive researchers to weed through the methodological literature and try to discern the valid options for a particular situation. Ideally, specialists and methodologists should strive to provide recommendations that are clear and based on what is known currently. In situations where this is not possible, classifying options according to our proposed framework for research practices can help the applied user choose a reasonable approach and avoid getting lost in a technical debate. This process of sorting methodological options into the SBO, RDFs, or QRPs also helps with setting out a research agenda for quantitative researchers. For example, areas in which the only available options are faulty or unacceptable, or where clear guidelines are lacking for how to choose among various options are deserving of prioritized attention from methodologists. We provide more prescriptive advice on this in the Discussion section.

A New Framework for Reporting Practices

The second part of our new taxonomy focuses on reporting practices (right side of Figure 1). Under this framework, practices are first differentiated on the basis of fraud, which is defined as the "intentional perversion of truth" and the "act of deceiving or misrepresenting" (Merriam-Webster, n.d.). Although there are more specific uses of the term fraud (e.g., legal, economic), we use fraud in its more general form as it relates to the reporting of data analyses. Readers may be aware that the term "research misconduct" is often used to describe instances of fraud in the research integrity literature. However, we argue that fraud is a fundamentally distinct subtype of research misconduct, the latter of which encompasses a broader array of issues, including fraud and QRPs (which we discuss; shown as the right side of each arm in Figure 1) and other aspects of wrongdoing such as harassment (which is beyond our current focus). Fraud, as used presently, more precisely describes actions taken with the intent of deceiving or misrepresenting.

We submit that fraud encompasses fabrication and falsification from the Steneck (2006) framework. According to the National Academy of Sciences, National Academy of Engineering, and Institute of Medicine (1992), fabrication is defined as "making up data or results" and falsification is defined as "changing data or results" (p. 27). However, for fabrication and falsification to qualify as fraud, these definitions must be contextualized to more clearly justify their inclusion under the reporting arm of our taxonomy and to better reflect the intent behind them. For example, it is reasonable for methodologists to generate (i.e., make up) data for simulation studies and for researchers, in some cases, to change data for an analysis such as using multiple imputation or removing influential observations. These examples represent research, rather than

reporting practices, and as long as these decisions are reported, there is no reporting malpractice, and most notably, no attempt to deceive.

To constitute fraud, there must be an attempt to purposefully mask a QRP. Therefore, fraud is always an issue of reporting as it obscures the truth, which removes the ability of a reader/replicator/reviewer to accurately evaluate the research practices implemented in a study. We thus revise the definitions of fabrication and falsification to be more consistent with the definition of fraud, which will allow for more precision in assessing whether or not reporting is fraud. We propose that fabrication be defined as: Making up data or results and failing to disclose doing so with the intent of deceiving or misrepresenting. In addition, we propose that falsification be defined as: Changing data or results and failing to disclose doing so with the intent of deceiving or misrepresenting.

Ultimately, fraud is a reporting issue, but will always be in reference to a QRP, and both pieces are included in the revised definitions. Because QRPs and fraudulent reporting are related issues (e.g., both fall under the broader umbrella of misconduct), it may be an easy mistake to euphemize instances of fraud as "questionable." However, we must stress that these issues are separate. By our definition, a QRP is a suboptimal research practice. Even if considered bad science or misconduct, a QRP is not fraud unless there is an intent to deceive or mask the QRP (i.e., intentionally obscuring what was done). We argue that what makes a research practice fraud is the deliberate distortion of information, which impedes a fair assessment of research practices. Our position is that fraud is more about deceptive reporting as opposed to the research practice itself. Although both fraud and QRPs represent misconduct, better distinguishing these terms will improve how the research community understands and addresses these issues.

As just noted, in our framework, the definition of fraud includes elements of both research (the initial QRP) and reporting (the deception) practices. Importantly, honest reporting does not necessarily solve issues with the underlying QRP, although it would remove the element of fraud. Consider a case where a researcher simulates data and adds it to empirical data before performing an analysis. If the researcher does not disclose that step, then fraud (in the form of fabrication) has been committed. On the other hand, suppose the researcher reports the fabrication of data and provides an entirely unacceptable explanation for the addition. This is not fraud—it was honestly reported—but remains a QRP as the research practice taken is (obviously) questionable.

Plagiarism, the last piece of the Steneck (2006) framework, also constitutes fraud. It is defined as "using the ideas or words of another person without giving appropriate credit" (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 1992, p. 27), which is another form of misrepresentation/deception. However, we do not include plagiarism in our taxonomy. All of the research and reporting practices organized by our taxonomy have the ability to influence quantitative results (e.g., augmenting heterogeneity across effects; leading to incorrect conclusions regarding data [Type-I/II errors]), and plagiarism does not have consequences of this specific nature. As put by Steneck (2006), "Plagiarism has no necessary impact on the reliability of the research record. Results are results, whether or not the person reporting them deserves credit for their discovery" (p. 62). Although we do not include it in our proposed taxonomy because of the lack of potential impact on results, this is not meant to suggest it is acceptable.

While fraud always involves an issue of reporting, reporting issues are not always fraudulent, particularly when the intent is not to deceive (e.g., unknowingly leaving out information). If reporting does not qualify as fraudulent, then reporting is either complete and correct, incomplete, incorrect, or incomplete and incorrect. As an example, we draw from the Flake and Fried (2020) study that focuses on improving reporting practices for psychological measurement. Consider a researcher who reports the use of a psychological scale in their study. For complete and correct reporting, the researcher should clearly and correctly report details pertaining to what the construct is, why/how the measure was selected, how scores were calculated, if the scale was modified (and if so, why/how), and if the measure is new (Flake & Fried, 2020). As examples, reporting would be incomplete if the researcher failed to report dropping items from the scale, incorrect if the researcher reported the wrong scoring method, and incomplete and incorrect if the researcher did both things simultaneously.

Importantly, these cases of incomplete and/or incorrect reporting signal an opportunity for the methodological literature to provide additional clarity in terms of what to report and how to report it. One hurdle is that some journals have strict word limits, which pushes analytical details to supplemental materials or functionally encourages their outright omission. While replicators do have access to supplemental materials, researchers may consult these less often and get comfortable with reporting fewer methodological details. Although not a one-size-fits-all solution, “methods checklists” could provide helpful and standardized guidance on reporting. In addition to the reporting standards set forth by Flake and Fried (2020) for psychological measurement practices, reporting standards have been recommended for other areas

including structural equation modeling (Boomsma et al., 2012), quantitative research in psychology (Appelbaum et al., 2018; Cooper, 2020), and more broadly, transparency and openness in science (Nosek et al., 2015).

Summary of New Taxonomy

To summarize, our recommended taxonomy first separates research practices from reporting practices. The left side of Figure 1 displays the framework for research practices, which are classified as either responsible conduct of research (acceptable) or as QRPs (unacceptable). Under responsible conduct of research, there is either the SBO (clearly optimal in all situations) or RDFs (multiple viable options). SBOs should always be used, QRPs should never be used, and any option from a set of RDFs is acceptable. The right side of Figure 1 displays the framework for reporting practices—how research practices are communicated to the reader. Reporting practices are classified as either honest or fraudulent. The goal is a truthful representation (i.e., honest) with a complete and correct account of research practices. Alternatively, accounts may be honest but incomplete, incorrect, or both incomplete and incorrect. If reporting is fraudulent, then it is either fabrication or falsification.

Research and reporting practices have the potential to impact quantitative conclusions, and consequently replication, in different ways, which is why we emphasize classifying them separately. This higher order distinction along with the rest of the taxonomy we have built out based on the Steneck (2006) framework is designed to incorporate more fine-grained details of research and reporting practices. This allows for more effective evaluations of practices, less ambiguity about best practices, and ultimately, less avoidable heterogeneity in results.

Conceptual Examples

To demonstrate our taxonomy in practice, Table 2 includes numbered examples of research and reporting practices with the original classification based on the article each practice is drawn from and a new classification based on our proposed taxonomy. Sometimes our new classifications agree with the previous classifications, while other times our classifications add nuance or replace the initial classifications. We note that many of the practices we classify as RDFs or QRPs also have a reporting element to them, when their use is not properly disclosed, but we chose to limit the reporting classification to those practices which are primarily an issue of reporting. An example of a classification that stayed the same was "choosing sample size" (Practice 1 in Table 2; Simmons et al., 2011, p. 1360). This practice remained an RDF because there is more than one viable approach to selecting sample size (see Maxwell et al., 2008 for a review), which follows our proposed definition. Of course, not providing any justification for the use of a particular sample size for a particular goal would be a QRP, given that this does not provide the researcher the evidence needed to make an informed decision. A classification that changed under the new taxonomy was for "withholding details of methodology or results in papers or proposals" (Practice 13 in Table 2; Martinson et al., 2005, p. 737). Although questionable (and justifiably classified as a QRP initially), the issue with this practice is more precisely about incomplete reporting because of the information omitted.

The new taxonomy adds nuance to the practices classified as conditional RDFs/QRPs (Practices 8 through 11 in Table 2). For example, "choosing among dependent variables" (Practice 8 in Table 2; Simmons et al., 2011) was initially classified

as an RDF. Certainly, when a researcher has access to different versions of a dependent variable, selecting among these is an RDF if data analysis is not conducted and the particular version of the construct is selected based on theory or measurement quality. However, if the decision to use a particular dependent variable is made based on statistical performance, this qualifies as a QRP. Consistent with this, Gelman and Loken (2014) note that researchers often use the version of the dependent variable that worked best, out of a post hoc rationale that this variable was the most appropriate.

Another example of a conditional RDF/QRP is "choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing)" (Practice 10 in Table 2; Wicherts et al., 2016, p. 3). When conducting multiple hypothesis tests (assuming the researcher correctly reports that this was done), there are often several different approaches that are methodologically justified depending on the researcher's goals (e.g., Maxwell et al., 2018), such as the more conservative Bonferroni adjustment (Dunn, 1961) versus a false discovery rate adjustment that sacrifices some Type-I error control to improve power (Benjamini & Hochberg, 1995). Choosing among these viable alternatives would be an RDF. However, failing to use any adjustment or using an adjustment known to provide poor Type-I error control (e.g., protected testing approaches in many situations; Levin et al., 1994) would qualify as a QRP.

Another advantage of using this new taxonomy is exhibited in Practices 3 through 5 from Table 2, which all describe the same practice: optional stopping rules for data collection. Known to inflate Type-I error rates (Lindsay, 2015; Wicherts et al., 2016), this practice is typically dependent on significance testing. Data collection either stops if a statistical test yields significance or continues until significance is obtained. Despite

consistent descriptions of optional stopping across multiple articles, the practice had a different classification in each of the following three: a QRP in John et al. (2012), *p*-hacking in Lindsay (2015), and an RDF in Wicherts et al. (2016). Our proposed taxonomy resolves this inconsistency and makes it clear that, given the evidence of inflating Type-I error, optional stopping is a QRP and should be avoided.[2]

Simulated Demonstration

To further illustrate its utility, we use our proposed taxonomy to guide methodological decisions based on evidence from a simulated data example. We demonstrate the importance of checking statistical assumptions before conducting a statistical analysis and use this information to promote best practices. Our example focuses on the homogeneity of variance assumption, violations of which can result in inflated (or reduced) Type-I error rates (Maxwell et al., 2018). We now briefly describe our methodology as well as the quantitative impact of this assumption violation on heterogeneity and Type-I error rates. We follow this by connecting these results to our taxonomy and identifying where RDFs and QRPs can emerge within the series of decisions associated with this one, seemingly simple step in an analysis. We aim for this demonstration to do three things: (1) remind readers of the consequences of assumption violations on the statistical validity of results, (2) illustrate our taxonomy with a common series of decisions that researchers encounter, and (3) emphasize the complexity of how QRPs and RDFs present themselves in practice.

[2] Note that some newer methods, such as sequential analysis, allow for additional data collection after seeing results, though stopping criteria are carefully defined a priori and the process is very controlled (Kelley et al., 2018; Lakens, 2014).

To demonstrate the impact of violating homogeneity of variance in the context of an independent samples *t*-test, data for two groups with equal means ($\mu_1 = 0$ and $\mu_2 = 0$) were generated in R (R Core Team, 2020) to correspond to three conditions. The first condition represented homogeneity of variance (control). Samples of equal size ($n_1 = 64$ and $n_2 = 64$) were drawn from normal distributions with equal variances ($\sigma_1^2 = 1$ and $\sigma_2^2 = 1$). The choice of sample size reflects the necessary sample size for 80% power assuming a medium effect size. The second condition represented a mild situation of heterogeneity of variance with equal sample sizes. Specifically, group sample sizes were the same as the control condition ($n_1 = 64$ and $n_2 = 64$) but group variances were set to differ by a ratio of nine ($\sigma_1^2 = 1$ and $\sigma_2^2 = 9$), which amounts to a more mild degree of heterogeneity of variance when sample sizes are equal (Keppel, 1991). Despite generating the same ratio of variances as the second condition ($\sigma_1^2 = 1$ and $\sigma_2^2 = 9$), the third condition represented a more severe situation, based on criteria from Blanca et al. (2018), because sample sizes were set to be unequal ($n_1 = 96$, $n_2 = 32$; ratio of 3:1). For each condition, 10,000 data sets were generated.

Following data generation, a pooled independent *t*-test was conducted to test the difference in means for significance on each of the 10,000 data sets. For each replication, we recorded the sample estimate of the difference in means and whether this difference was statistically significant (two-tailed). The sample descriptive statistics for the estimated mean difference are displayed in Table 3. On average, the estimate of the mean difference was unbiased across all conditions. However, the average standard deviation (*SD*) of the mean difference changed dramatically with respect to condition: When compared with the control condition, the standard deviation more than doubled under

Condition 2 (mild heterogeneity of variance and equal sample sizes) and more than tripled under Condition 3 (mild heterogeneity of variance and unequal sample sizes), emphasizing an increase in avoidable heterogeneity. Table 4 displays the influence of this assumption violation on Type-I error. With mild heterogeneity of variance and equal sample sizes, there was a negligible increase to .051. With mild heterogeneity of variance and unequal sample sizes, however, the Type-I error rate was .210, four times the nominal rate.

With the evidence provided by this simulation, we now walk through how our proposed taxonomy can be used to guide methodological decisions. The first choice to consider is whether the data should be assessed for potential violations of assumptions (e.g., homogeneity of variance). Because of the potential Type-I error inflation of up to .210 based on violation of only one statistical assumption, the results of this simulation stress the importance of checking assumptions before a statistical analysis.

Regarding our taxonomy, this research practice of checking statistical assumptions is the SBO because the only alternative is not doing so. This alternative would be a QRP because if a researcher does not assess statistical assumptions, the researcher does not have the necessary evidence with which to make an informed decision as to whether the selected model and statistical test are appropriate for the data. As shown in our example, failing to check for assumption violations can increase the risk of Type-I errors, which make it difficult to evaluate replication findings and interpret results across studies. This is in addition to the added avoidable heterogeneity that such violations add to the literature, evidenced by the two- and three-fold increases in the standard deviations among mean differences produced by the 10,000 otherwise identical

replications. Of course, results are not always impacted by failing to check assumptions, in the lucky situations in which all assumptions are satisfied, but as previously mentioned and shown in Table 1, QRPs should not be solely dependent on the potential impact on results but also dependent on having sufficient information to make defensible decisions. By our definition, QRPs include practices that lead to suboptimal results as well as practices that lead to ignorance about whether results are optimal or not.

Once a researcher selects the SBO of checking assumptions, a set of RDFs is presented in terms of how the assumptions are evaluated. For heterogeneity of variance, there are various guidelines regarding the ratio of group variances (e.g., Blanca et al., 2018; Keppel, 1991) as well as more formal tests (e.g., Levene's test), in addition to visual inspection (e.g., residual plots). For many situations, there is no firmly agreed upon dividing line for what degree of heterogeneity is deemed problematic, and visual inspection can be subject to individual differences. Though methodologists usually suggest using several methods to arrive at a careful conclusion as to the potential assumption violation, the varied acceptable options suggest classification as RDFs.

Once a researcher has arrived at a method for checking the assumption, the evidence may suggest, for example, that variances are generally homogeneous for an independent samples *t*-test. If so, the researcher may proceed with the standard analysis as usual. However, if the evidence suggests mild heterogeneity of variance, then we arrive at an additional tier of RDFs: ignore the violation or address it. Importantly, whichever of the two options is chosen, the researcher has sufficient evidence to support that results will be minimally affected. These differing responses are RDFs (viable/supported research practices) because the methodological literature has noted that

while the assumptions are technically violated, ignoring the mild violation results in negligible consequences. This was replicated in our demonstration, with Type-I error increasing almost imperceptibly when compared with the control condition with no assumption violation. Alternatively, if the researcher chooses to address the mild violation, then we are met with yet another tier of RDFs. For the specific case of a *t*-test, Maxwell et al. (2018) suggest Welch's W instead of the pooled *t*-test, but other separate variance tests are available more generally, as well as sandwich-type standard errors, of which there are several options (Cai & Hayes, 2008; Hayes & Cai, 2007). Although not all responses are equivalent, the methodological literature has not yet arrived at an SBO, particularly for more complex designs, and thus, most of these options would be considered RDFs.

Finally, if the evidence from checking assumptions suggests a mild violation of homogeneity of variance when sample sizes are fairly unequal (or a more severe degree of heterogeneity of variance), then ignoring the violation becomes a QRP because the methodological literature is clear as to the nonignorable impact to Type-I error. The SBO, in this case, is to address the violation. However, depending on context and the evidence in favor for each option, the choice of a particular approach to dealing with the violation again becomes an RDF, as it did in the mild violation scenario. We hope this demonstration has not only illustrated our taxonomy but also emphasized the complex web of decisions researchers must make throughout data analysis, which is compounded when one considers that this example only examined one small piece of data analysis. We recognize that there are a great deal of decisions researchers must make, and we hope that

organizing the RDF-QRP landscape can help streamline this decision-making process or identify where more methodological focus is needed.

Discussion

In this article, we developed a revised and expanded taxonomy for assessing and classifying research practices as well as reporting practices. Under the research practices framework, if a practice constitutes responsible conduct of research, it is either the SBO if optimal in all situations or an RDF if one of multiple viable options. Opposite to responsible conduct of research are QRPs, which are suboptimal practices and a form of misconduct. For the reporting practices framework, there is fraudulent, which includes fabrication and falsification. If reporting is honest, then it is either complete and correct or incomplete and/or incorrect.

The first goal of the new taxonomy was to help researchers understand the different sources of heterogeneity across studies. As the field moved away from narrow definitions of replication (i.e., focusing on statistical significance only or on exact agreement among effect sizes), it became reasonable to expect results to vary across replications. Some sources of this heterogeneity are unavoidable, such as sampling error (more common with smaller sample size studies) or even the nature of an effect changing over time. Heterogeneity can also be unavoidable when driven by sources that are more explicit but hard to identify or control (Kenny & Judd, 2019). As evidence of some of this heterogeneity, replication results varied across studies with a high degree of control including preregistered replications (Klein et al., 2018) and analyses using the same dataset (Silberzahn et al., 2018).

There are also avoidable sources of heterogeneity, which can be partly attributed to methodological flexibility. QRPs are a more obvious source of heterogeneity as such practices are suboptimal according to our definition. Our demonstration showed that failing to test for even a single statistical assumption could lead to substantial increases in avoidable heterogeneity. It is important to note, though, that the severity of QRPs in practice is likely to have a built-in ceiling. Reviewers with quantitative expertise, regardless of primary area of study, can potentially notice and correct cases of improper methodology thus preventing more serious or obvious QRPs from getting published. If studies using QRPs are published, this can point out which areas need best practices to be more carefully delineated and where more stringent guidelines are needed to discourage suboptimal practice. RDFs are also a source of heterogeneity, but these may currently be less malleable because such practices are perceived as viable and acceptable. Despite this, selection among different alternatives may be adding avoidable heterogeneity to results and research should attempt to identify areas where such flexibility can be reduced or streamlined.

The second goal of the new taxonomy was to provide a structure for organizing and evaluating research and reporting practices that contribute to avoidable heterogeneity. It is important to develop a clear lexicon when describing a topic area that has important implications for research literatures. The classifications (and tiers of classification) provided by our proposed taxonomy help reduce the ambiguity that surrounds methodological flexibility. Our taxonomy is also adaptive to changing guidelines and open to debate among methodologists and researchers. If additional research leads to more optimal choices or reveals a current accepted practice to be

suboptimal, a practice once classified as an RDF may be reclassified as a QRP or become a conditional RDF/QRP if it is viable in some situations but suboptimal in others. Having a clear organizational structure makes it easier to have productive discussions around these topics.

The third goal of the new taxonomy was to help researchers navigate the complex series of decisions made among acceptable and unacceptable practices throughout the research process. This complexity was illustrated most clearly in our simulated demonstration, where a seemingly simple task of checking a single statistical assumption resulted in a series of diverging paths among SBOs, QRPs, and RDFs. If a researcher knows which practices fall under responsible conduct of research (SBOs or RDFs) or which fall under QRPs, methodological decisions become more straightforward.

The fourth goal of the new taxonomy was to drive research toward domains where more methodological guidance is needed. SBOs should be the aspirational goal as there is no ambiguity when a single option is optimal in all situations. Therefore, RDFs are a potential call for more methodological research. Additional methodological work could provide evidence suggesting that an RDF is a QRP. This would help narrow the options confronting researchers. Alternatively, a previously hidden SBO could rise to the top and outperform the other methods that were thought to be RDFs. Again, this would narrow down unnecessary options. In situations where RDFs end up being truly equivalent, then the specific choice among them does not matter. If this is the case, it could be argued that all but one option should be eliminated to streamline decision making.

Although we have provided a revised and expanded taxonomy, there are still limitations. First, many practices do not fit neatly into one of the taxonomy classifications. As we have shown, classifications for conditional RDFs/QRPs change from one situation to the next, which can complicate the decision-making process. Moreover, classifications may change over time, as new research clarifies the consequences of different choices. Additionally, as previously mentioned, some research practices contain elements of reporting practices and vice versa, which can also blur how to best classify a practice. Lastly, other researchers may disagree with us or one another regarding one or more of our classifications or may prefer an expanded definition that includes nonmethodological research practices. Our taxonomy was not developed to address all facets of misconduct, nor to resolve all ambiguity associated with methodological flexibility, but rather to add greater linguistic clarity and provide a productive framework in which to situate these important discussions.

In the future, we hope to see the use of our taxonomy extended and tailored to fit specialty areas. We feel psychometrics is an especially fruitful area where heterogeneity in results has not yet received much attention in the replication literature. Many studies depend on the soundness of psychological measures, which stresses the importance of understanding what heterogeneity means for psychometric studies. Additionally, our proposed taxonomy could be useful for guiding sensitivity analyses (McShane & Böckenholt, 2014) to examine the consequences of different sources of unavoidable/avoidable heterogeneity. Knowing the sources of heterogeneity that impact results the most would further our aims to provide clearer guidance and reduce methodological flexibility. Finally, as we mentioned for our fourth goal of the new

taxonomy, future research should aspire toward SBOs, which leave no room for ambiguity. Under our taxonomy, RDFs inject the most ambiguity into a methodological decision. Therefore, future methodological work would benefit from sifting out SBOs from a set RDFs, revealing QRPs disguised as RDFs, and eliminating any unnecessary options. This would help simplify the many choices researchers are required to navigate through.

Our proposed taxonomy is broadly applicable to replication issues across a variety of methodological areas. Throughout this article, we have demonstrated the use of this taxonomy to better understand heterogeneity related to design, analysis, and reporting. Researchers may wonder where this taxonomy sits among the numerous solutions that have been proposed to aid replication and open science (e.g., preregistration, registered reports, direct replications, sample size planning, sensitivity analyses, reporting standards, hold-out sampling, reliable measures, sequential testing, statistical controls, crowdsourcing data, etc.). We do not view this taxonomy as a "solution," but rather view it as a structural aid to help understand the problem and gain greater clarity on what facets of the problem are being addressed by various proposed solutions. We believe that the literature on replication and research practices has been hampered by a lack of clear terminology and without some consistency and an overarching framework, researchers will continue to use the same words to mean different things, or use different words to mean the same thing, both of which stall progress.

We hope that this taxonomy facilitates a more productive discussion of these issues for both methodologists and applied researchers, which we hope will make further improvements in identifying methodological areas for improvement and contribute to the

greater discussion on replicability. Science is at its best when it is robust, compelling, and impactful. We hope this taxonomy provides a useful framework to examine and improve the robustness of our science.

References

Agresti, A., & Franklin, C. (2007). *Statistics: The art and science of learning from data*. Pearson Prentice Hall.

Allison, P. D. (2012). Handling missing data by maximum likelihood. *(Keynote Presentation at the SAS Global Forum, April 23, 2012, Orlando, Florida)*. https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf

Anderson, S. F. (2020). Misinterpreting p: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychological Methods*, *25*(5), 596–609. https://doi.org/10.1037/met0000248

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12. https://doi.org/10.1037/met0000051

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305–324. https://doi.org/10.1080/00273171.2017.1289361

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, *50*(3), 937–962. https://doi.org/10.3758/s13428-017-0918-2

Boomsma, A., Hoyle, R. H., & Panter, A. T. (2012). The structural equation modeling research report. In R. H. Hoyle (Ed.), *Handbook of Structrual Equation Modeling* (pp. 341–358). The Guilford Press.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*(1), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Cai, L., & Hayes, A. F. (2008). A new test of linear hypotheses in OLS regression under heteroscedasticity of unknown form. *Journal of Educational and Behavioral Statistics*, *33*(1), 21–40. https://doi.org/10.3102/1076998607302628

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates, Publishers.

Cooper, H. (2020). *Reporting quantitative research in psychology: How to meet APA Style Journal Article Reporting Standards* (2nd ed.). American Psychological Association. https://doi.org/10.1037/0000178-000

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), 1–7. https://doi.org/10.1371/journal.pone.0029081

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245920952393

Friese, M., & Frankenbach, J. (2020). p-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. https://doi.org/10.1037/met0000246

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*(6), 933–948. https://doi.org/10.1037/a0029709

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics*, *29*(3), 885–891. https://doi.org/10.1214/aoms/1177706545

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465. https://doi.org/10.1511/2014.111.460

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*(4), 709–722. https://doi.org/10.3758/BF03192961

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), 1–15. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, *24*(5), 557–570. https://doi.org/10.1037/met0000189

Hertzog, C., & Rovine, M. (1985). Repeated-measures analysis of variance in developmental research : Selected issues. *Child Development*, *56*(4), 787–809. https://doi.org/10.2307/1130092

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Chance*, *18*(4), 40–47. https://doi.org/10.1080/09332480.2005.10722754

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. https://doi.org/10.1037/met0000127

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Prentice-Hall, Inc.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin*, *115*(1), 153–159. https://doi.org/10.1037/0033-2909.115.1.153

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Martinson, B. C., Anderson, M. S., & De Vries, R. (2005). Scientists behaving badly. *Nature*, *435*(7043), 737–738. https://doi.org/10.1038/435737a

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis?: What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A. J., & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, *43*(5), 594–627. https://doi.org/10.3102/1076998618776348

McNutt, M. (2014). Reproducibility. *Science*, *343*(6168), 229. https://doi.org/10.1126/science.1250475

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612–625. https://doi.org/10.1177/1745691614548513

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *American Statistician, 73*(sup1), 99–105. https://doi.org/10.1080/00031305.2018.1505655

Merriam-Webster. (n.d.). Fraud. Retrieved February 23, 2021, from https://www.merriam-webster.com/dictionary/fraud

Mullard, A. (2011). Reliability of "new drug target" claims called into question. *Nature Reviews Drug Discovery*, *10*(9), 643–644. https://doi.org/10.1038/nrd3545

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (1992). *Responsible science: Ensuring the integrity of the research process: Volume I*. The National Academies Press. https://doi.org/10.17226/2091

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., … Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oster, R. A., & Hilbe, J. M. (2008). An examination of statistical software packages for parametric and nonparametric data analyses using exact methods. *The American Statistician, 62*(1), 74–84. https://doi.org/10.1198/000313008X268955

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Pittinsky, T. L. (2015). America's crisis of faith in science. *Science, 348*(6234), 511–512. https://doi.org/10.1126/science.348.6234.511-a

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rutjens, B. T., Heine, S. J., Sutton, R. M., & Van Harreveld, F. (2018). Attitudes towards science. In *Advances in Experimental Social Psychology* (1st ed., Vol. 57). Elsevier Inc. https://doi.org/10.1016/bs.aesp.2017.08.001

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127–139. https://doi.org/10.1037/met0000302

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Lawrence Erlbaum Associates, Inc.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Sijtsma, K., Veldkamp, C. L. S., & Wicherts, J. M. (2016). Improving the conduct and reporting of statistical analysis in psychology. *Psychometrika*, *81*(1), 33–38. https://doi.org/10.1007/s11336-015-9444-2

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics*, *12*(1), 53–74. https://doi.org/10.1007/PL00022268

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. https://doi.org/10.1037/a0036731

Wang, J., & Johnson, D. E. (2019). An examination of discrepancies in multiple imputation procedures between SAS® and SPSS®. *The American Statistician*, *73*(1), 80–88. https://doi.org/10.1080/00031305.2018.1437078

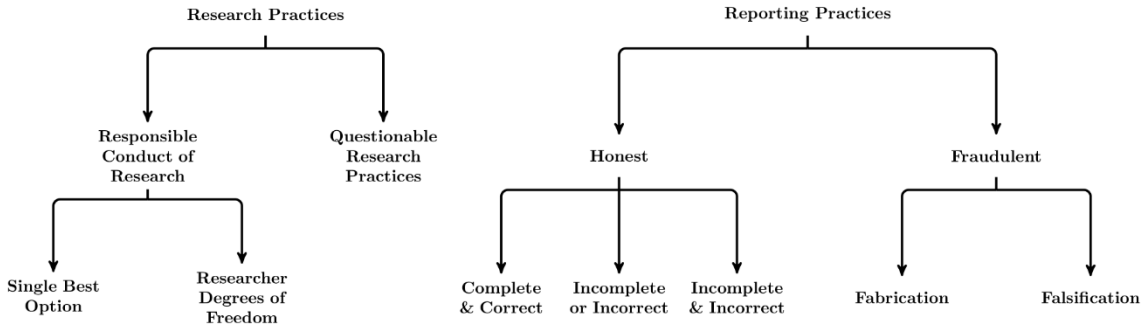
Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(1832), 1–12. https://doi.org/10.3389/fpsyg.2016.01832

Wigboldus, D. H. J., & Dotsch, R. (2016). Encourage playing with data and discourage questionable reporting practices. *Psychometrika*, *81*(1), 27–32. https://doi.org/10.1007/s11336-015-9445-1

Wingen, T., Berkessel, J. B., & Englich, B. (2020). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science*, *11*(4), 454–463. https://doi.org/10.1177/1948550619877412

Figure 5

New Taxonomy for Assessing Research and Reporting Practices



Note. Framework for assessing research practices on left and framework for assessing reporting practices on right.

Table 11

Conditions for Acceptable and Unacceptable Research Practices

Condition	All the necessary/relevant information *has* been acquired	All the necessary/relevant information *has not* been acquired
The action taken *was* defensible	RDF or SBO	QRP
The action taken *was not* defensible	QRP	QRP

Note. RDF = researcher degree of freedom; SBO = single best option; QRP = questionable research practice.

Table 12

Reclassification of Research and Reporting Practices from the Literature

#	Practice	Original classification	New classification
1	Choosing sample size (Simmons et al., 2011)	RDF	RDF
2	Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators (Wicherts et al., 2016)	RDF – design	RDF
3	Deciding whether to collect more data after looking to see whether the results were significant (John et al., 2012)	QRP	QRP
4	Using the optional-stopping strategy during data collection (Lindsay, 2015)	*p*-hacking	QRP
5	Determining the data collection stopping rule on the basis of desired results or intermediate significance testing (Wicherts et al., 2016)	RDF – collection	QRP
6	Dropping subjects, observations, measures, or conditions that yielded inconvenient data (Lindsay, 2015)	*p*-hacking	QRP
7	Using inappropriate statistical or other methods of measurement to enhance the significance of research findings (National Academy of Sciences et al., 1992)	QRP	QRP
8	Choosing among dependent variables (Simmons et al., 2011)	RDF	Conditional RDF/QRP
9	Conducting explorative research without any hypothesis (Wicherts et al., 2016)	RDF – hypothesis	Conditional RDF/QRP
10	Choosing inference criteria (e.g., Bayes factors, alpha level, sidedness of the test, corrections for multiple testing) (Wicherts et al., 2016)	RDF – analyses	Conditional RDF/QRP
11	Deciding how to deal with violations of statistical assumptions in an *ad hoc* manner (Wicherts et al., 2016)	RDF – analyses	Conditional RDF/QRP
12	Specifying pre-processing of data (e.g., cleaning, normalization, smoothing, motion correction) in an ad hoc manner (Wicherts et al., 2016)	RDF – analyses	Reporting – complete
13	Withholding details of methodology or results in papers or proposals (Martinson et al., 2005)	QRP	Reporting – incomplete
14	Falsifying or ‘cooking’ research data (Martinson et al., 2005)	QRP	Reporting – fraudulent (falsification)

Note. Original classification = classification based on cited article. New classification = classification based on new taxonomy for research and reporting practices; RDF = researcher degree of freedom; QRP = questionable research practice.

Table 13

Sample Descriptives for Mean Difference When the Homogeneity of Variance Assumption is Violated

Condition	Mean difference	SD of mean difference
No violation (control)	< 0.001	0.175
Mild violation; equal *n*	< 0.001	0.395
Mild violation; unequal *n*	0.010	0.546

Note. True mean difference is zero; Sample descriptives averaged across 10,000 replications; *n* = sample size.

Table 14

Effect of Violating the Homogeneity of Variance Assumption on Type-I Error

Condition	Type-I error
No violation (control)	0.046
Mild violation; equal *n*	0.051
Mild violation; unequal *n*	0.210

Note. *n* = sample size.

CHAPTER 4

CONCLUSION

With the exception of a few studies (e.g., Flake et al., 2017, 2022; Flake & Fried, 2020; Hussey & Hughes, 2020; Shaw et al., 2020), the critical role that psychometrics plays in replication has been largely overlooked. This dissertation works toward integrating psychometrics into the broader area of replication. Specifically, I focused on exploring heterogeneity in factor analytic results. Chapter 2 provides an example of heterogeneity in results for studies factor analyzing the same scale (BSCS). Upon further examination of the empirical scale development literature, I found heterogeneity in factor analytic results for two additional psychological scales (CES-D and NEO-PI-R). These findings indicated that potential replication issues in factor analysis may be widespread, which warranted further investigation. There was a need to better understand the heterogeneity being observed in results for psychological scales to effectively assess factor analytic replication. Excessive heterogeneity can blur impressions across a series of results, so minimizing sources of avoidable heterogeneity would help prevent heterogeneity from becoming excessive.

A source of (potentially) avoidable heterogeneity that I focused my attention on was methodological flexibility. I looked to the replication literature outside of psychometrics in search of a framework that could guide evaluations of methodological flexibility in factor analysis. Unfortunately, the language used to talk about research practices (e.g., RDFs, QRPs) was inconsistent, vague, and overlapping. A coherent framework was needed to continue the exploration of heterogeneity in factor analytic results. In Chapter 3, a new taxonomy is developed to assess research practices. As part

of the taxonomy, key terms such as RDFs and QRPs are redefined and distinguished to allow for improved conceptual clarity. The new taxonomy facilitates more effective assessments of research practices, reduces unnecessary methodological flexibility, and minimizes avoidable heterogeneity. It is broadly applicable to many methodological areas and has been used to evaluate heterogeneity across factor analytic results.

The first extension of the work in this dissertation involved an examination of heterogeneity across factor analyses in the empirical literature. In a review of 31 factor analytic studies, the taxonomy from Chapter 3 was used to evaluate how flexibility in research practices may be contributing to potentially avoidable heterogeneity in factor analytic results. RDFs (acceptable practices) were distinguished from QRPs (unacceptable practices) with the goal of reducing unnecessary methodological flexibility in factor analysis. Another extension of the work in this dissertation examined heterogeneity across factor analyses within a simulation-based framework. The simulation was designed to assess the impact of different methodological choices on heterogeneity to identify which practices might be injecting avoidable heterogeneity into results. For data features, sample size, number of factors, strength of factor correlations, and strength of factor loadings had a sizable impact on factor analysis results. For methodological options, choice of rotation was relatively influential.

There are many potential future directions at the intersection of psychometrics and replication. First, there is an opportunity for future work to develop formal definitions of replication in factor analysis. A more precise definition would allow for better assessments of replication in the literature and clearer goals for studies attempting to replicate a factor analysis. It may be useful to draw from the terminology used in the

measurement invariance literature: configural, metric, strong, and strict (see Meredith, 1993). These levels of invariance could be adapted into levels of replication. Second, there are other sources of heterogeneity not addressed in this dissertation that could have a substantial influence on factor analytic results (e.g., sampling variability). Lastly, additional simulation studies could be conducted that directly measure heterogeneity in factor analytic results or assess the impact of different sources.

Although it has been generally overlooked, psychometric replication has an important role to play in fostering a more replicable scientific literature. The replicability of a psychometric study strengthens confidence in the quality of a measure for a particular context and, more importantly, the use and interpretation of scores derived from that measure (i.e., validity). I hope this work encourages methodologists to consider the intersection between psychometrics and replication in further, more nuanced ways, and encourages researchers to be sensitive to the impact of research practices on replication when designing and conducting factor analytic studies. This work serves to promote the role of psychometrics in conferring a cohesive, cumulative body of scientific knowledge.

REFERENCES

Ackerman T. A., Gierl M. J., & Walker C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51.

Agresti, A., & Franklin, C. (2007). *Statistics: The art and science of learning from data*. Pearson Prentice Hall.

Allison, P. D. (2012). Handling missing data by maximum likelihood. *(Keynote Presentation at the SAS Global Forum, April 23, 2012, Orlando, Florida)*. https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf

Anderson, S. F. (2020). Misinterpreting p: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychological Methods, 25*(5), 596–609. https://doi.org/10.1037/met0000248

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*(1), 1–12. https://doi.org/10.1037/met0000051

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research, 52*(3), 305–324. https://doi.org/10.1080/00273171.2017.1289361

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Babinski L., Hartsough C., & Lambert N. (1999). Childhood conduct problems, hyperactivity-impulsivity, and inattention as predictors of adult criminal activity. *Journal of Child Psychology and Psychiatry*, 40, 347-355.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, *50*(3), 937–962. https://doi.org/10.3758/s13428-017-0918-2

Bock R. D., & Mislevy R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psycho-logical Measurement*, 6, 431-444.

Boomsma, A., Hoyle, R. H., & Panter, A. T. (2012). The structural equation modeling research report. In R. H. Hoyle (Ed.), *Handbook of Structrual Equation Modeling* (pp. 341–358). The Guilford Press.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*(1), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Brennan G. M., & Baskin-Sommers A. R. (2018). Brain-behavior relationships in externalizing: P3 amplitude reduction reflects deficient inhibitory control. *Behavioural Brain Research*, 337, 70-79.

Browne M. W., & Cudeck R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21, 230-258.

Browne M. W., Cudeck R., Tateneni K., & Mels G. (2010). CEFA: Comprehensive exploratory factor analysis, version 3.04 [Computer software and manual]. Retrieved from https://psychology.osu.edu/dr-browne-software

Byrne B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment*, 85, 17-32.

Cai L. (2017). *flexMIRT version 3.51: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

Cai L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33-57.

Cai L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35, 307-335.

Cai, L., & Hayes, A. F. (2008). A new test of linear hypotheses in OLS regression under heteroscedasticity of unknown form. *Journal of Educational and Behavioral Statistics*, *33*(1), 21–40. https://doi.org/10.3102/1076998607302628

Chen W., & Thissen D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.

Christoffersson A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40, 5-32.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates, Publishers.

Cooper, H. (2020). *Reporting quantitative research in psychology: How to meet APA Style Journal Article Reporting Standards* (2nd ed.). American Psychological Association. https://doi.org/10.1037/0000178-000

Costa Jr., P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI–R) and NEO five-factor inventory (NEO–FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

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

Čubranić-Dobrodolac M., Lipovac K., Čičević S., & Antić B. (2017). A model for traffic accidents prediction based on driver personality traits assessment. *Promet (Zagreb)*, 29, 631-642.

De Ayala R. J. (2008). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Ridder D. T. D., De Boer B. J., Lugtig P., Bakker A. B., & Van Hooft E. A. J. (2011). Not doing bad things is not equivalent to doing the right thing: Distinguishing between inhibitory and initiatory self-control. *Personality and Individual Differences*, 50, 1006-1011.

DeVellis R. F. (2012). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

DeWalt D. A., Thissen D., Stucky B. D., Langer M. M., Morgan DeWitt E., Irwin D. E., . . . Varni J. W. (2013). PROMIS Pediatric Peer Relationships Scale: Development of a peer relationships item bank as part of social health measurement. *Health Psychology*, 32, 1093-1103.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), 1–7. https://doi.org/10.1371/journal.pone.0029081

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Edwards M. C. (2009). An introduction to item response theory using the Need for Cognition Scale. *Social and Personality Psychology Compass*, 3, 507-529.

Edwards M. C., Houts C. R., & Cai L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, 23, 138-149.

Eisenberg I. W., Bissett P. G., Canning J. R., Dallery J., Enkavi A. Z., Whitfield-Gabrieli S., . . . Poldrack R. A. (2018). Applying novel technologies and methods to inform the ontology of self-regulation. *Behaviour Research and Therapy*, 101, 46-57.

Embretson S. E., & Reise S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Enkavi A. Z., Eisenberg I. W., Bissett P. G., Mazza G. L., MacKinnon D. P., Marsch L. A., & Poldrack R. A. (2019). A large-scale analysis of test-retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 5472-5477.

Fabrigar L. R., Wegener D. T., MacCallum R. C., & Strahan E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Ferrari J. R., Stevens E. B., & Jason L. A. (2009). The role of self-regulation in abstinence maintenance: Effects of communal living on self-regulation. *Journal of Groups in Addiction & Recovery*, 4, 32-41.

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*. https://doi.org/10.1037/amp0001006

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flora D. B., & Curran P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.

Follmer D. J., Sperling R. A., & Suen H. K. (2017). The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher*, 46, 329-334.

Friese, M., & Frankenbach, J. (2020). p-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. https://doi.org/10.1037/met0000246

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*(6), 933–948. https://doi.org/10.1037/a0029709

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics*, *29*(3), 885–891. https://doi.org/10.1214/aoms/1177706545

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465. https://doi.org/10.1511/2014.111.460

Hagger, M. S., Zhang, C.-Q., Kangro, E.-M., Ries, F., Wang, J. C. K., Heritage, B., & Chan, D. K. C. (2018). Trait self-control and self-discipline: Structure, validity, and invariance across national groups. *Current Psychology*. https://doi.org/10.1007/s12144-018-0021-6

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*(4), 709–722. https://doi.org/10.3758/BF03192961

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), 1–15. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. V., & Schauer, J. M. (2019a). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, *44*(5), 543–570. https://doi.org/10.3102/1076998619852953

Hedges, L. V., & Schauer, J. M. (2019b). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, *24*(5), 557–570. https://doi.org/10.1037/met0000189

Hertzog, C., & Rovine, M. (1985). Repeated-measures analysis of variance in developmental research : Selected issues. *Child Development*, *56*(4), 787–809. https://doi.org/10.2307/1130092

Horn J. L. (1965). A rationale and technique for estimating the number of factors in factor analysis. *Psychometrika*, 30, 179-185.

Hu L., & Bentler P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, 251524591988290. https://doi.org/10.1177/2515245919882903

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Chance*, *18*(4), 40–47. https://doi.org/10.1080/09332480.2005.10722754

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. https://doi.org/10.1037/met0000127

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Prentice-Hall, Inc.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin, 115*(1), 153–159. https://doi.org/10.1037/0033-2909.115.1.153

Lindner C., Nagy G., & Retelsdorf J. (2015). The dimensionality of the Brief Self-Control Scale: An evaluation of unidimensional and multidimensional applications. *Personality and Individual Differences*, 86, 465-473.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

MacCallum R. C., Widaman K. F., Zhang S., & Hong S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-99.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Maloney P. W., Grawitch M. J., & Barber L. K. (2012). The multi-factor structure of the Brief Self-Control Scale: Discriminant validity of restraint and impulsivity. *Journal of Research in Personality*, 46, 111-115.

Manapat, P. D., Anderson, S. F., & Edwards, M. C. (2022). A revised and expanded taxonomy for understanding heterogeneity in research and reporting practices. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000488

Manapat, P. D., Edwards M. C., MacKinnon, D. P., Poldrack, R. A., & Marsch, L. A. (2021). A psychometric analysis of the Brief Self-Control Scale. *Assessment, 28*(2), 395-412. https://doi.org/10.1177/1073191119890021

Martinson, B. C., Anderson, M. S., & De Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737–738. https://doi.org/10.1038/435737a

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis?: What does “failure to replicate” really mean? *American Psychologist, 70*(6), 487–498. https://doi.org/10.1037/a0039400

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A. J., & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, *43*(5), 594–627. https://doi.org/10.3102/1076998618776348

McDonald R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1-21.

McNeish D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412-433.

McNutt, M. (2014). Reproducibility. *Science*, *343*(6168), 229. https://doi.org/10.1126/science.1250475

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612–625. https://doi.org/10.1177/1745691614548513

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *American Statistician*, *73*(sup1), 99–105. https://doi.org/10.1080/00031305.2018.1505655

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Merriam-Webster. (n.d.). Fraud. Retrieved February 23, 2021, from https://www.merriam-webster.com/dictionary/fraud

Monroe S., & Cai L. (2015). Examining the reliability of student growth percentiles using multidimensional IRT. *Educational Measurement: Issues and Practice*, 34(4), 21-30.

Morean M. E., Demartini K. S., Leeman R. F., Pearlson G. D., Anticevic A., Krishnan-Sarin S., . . . O'Malley S. S. (2014). Psychometrically improved, abbreviated versions of three classic measures of impulsivity and self-control. *Psychological Assessment*, 26, 1003-1020.

Mullard, A. (2011). Reliability of "new drug target" claims called into question. *Nature Reviews Drug Discovery*, *10*(9), 643–644. https://doi.org/10.1038/nrd3545

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (1992). *Responsible science: Ensuring the integrity of the research process: Volume I*. The National Academies Press. https://doi.org/10.17226/2091

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., … Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oster, R. A., & Hilbe, J. M. (2008). An examination of statistical software packages for parametric and nonparametric data analyses using exact methods. *The American Statistician, 62*(1), 74–84. https://doi.org/10.1198/000313008X268955

Paolacci G., & Chandler J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23, 184-188.

Pargent, F., Hilbert, S., Eichhorn, K., & Bühner, M. (2019). Can't make it better nor worse: An empirical study about the effectiveness of general rules of item construction on psychometric properties. *European Journal of Psychological Assessment, 35*(6), 891–899. https://doi.org/10.1027/1015-5759/a000471

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Patton J. H., Stanford M. S., & Barratt E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, 51, 768-774.

Pittinsky, T. L. (2015). America's crisis of faith in science. *Science, 348*(6234), 511–512. https://doi.org/10.1126/science.348.6234.511-a

Preston K. S. J., Gottfried A. W., Park J. J., Manapat P. D., Gottfried A. E., & Oliver P. H. (2018). Simultaneous linking of cross-informant and longitudinal data involving positive family relationships. *Educational and Psychological Measurement*, 78, 409-429.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reckase M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.

Reise S. P., Moore T. M., Sabb F. W., Brown A. K., & London E. D. (2013). The Barratt Impulsiveness Scale-11: Reassessment of its structure in a community sample. *Psychological Assess-ment*, 25, 631-642.

Rosseel Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/

Rutjens, B. T., Heine, S. J., Sutton, R. M., & Van Harreveld, F. (2018). Attitudes towards science. In *Advances in Experimental Social Psychology* (1st ed., Vol. 57). Elsevier Inc. https://doi.org/10.1016/bs.aesp.2017.08.001

Samejima F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127–139. https://doi.org/10.1037/met0000302

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Lawrence Erlbaum Associates, Inc.

Sharma L., Markon K. E., & Clark L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, 140, 374-408.

Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from Many Labs 2. *Canadian Psychology*, *61*(4), 289–298. https://doi.org/10.1037/cap0000220

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

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

Sijtsma, K., Veldkamp, C. L. S., & Wicherts, J. M. (2016). Improving the conduct and reporting of statistical analysis in psychology. *Psychometrika, 81*(1), 33–38. https://doi.org/10.1007/s11336-015-9444-2

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics, 12*(1), 53–74. https://doi.org/10.1007/PL00022268

Tangney J. P., Baumeister R. F., & Boone A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72, 271-324.

Thissen D., & Steinberg L. (2009). Item response theory. In Millsap R. E., Maydeu-Olivares A. (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 148-177). London, England: Sage.

Thissen D., & Wainer H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.
Thurstone L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Venables N. C., Foell J., Yancey J. R., Kane M. J., Engle R. W., & Patrick C. J. (2018). Quantifying inhibitory control as externalizing proneness: A cross-domain model. *Clinical Psychological Science*, 6, 561-580.

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. https://doi.org/10.1037/a0036731

Wang, J., & Johnson, D. E. (2019). An examination of discrepancies in multiple imputation procedures between SAS® and SPSS®. *The American Statistician*, *73*(1), 80–88. https://doi.org/10.1080/00031305.2018.1437078

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(1832), 1–12. https://doi.org/10.3389/fpsyg.2016.01832

Wigboldus, D. H. J., & Dotsch, R. (2016). Encourage playing with data and discourage questionable reporting practices. *Psychometrika*, *81*(1), 27–32. https://doi.org/10.1007/s11336-015-9445-1

Wingen, T., Berkessel, J. B., & Englich, B. (2020). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science*, *11*(4), 454–463. https://doi.org/10.1177/1948550619877412

Wirth R. J., & Edwards M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58-79.

Zhang Z., & Yuan K.-H. (2015). coefficientalpha: Robust coefficient alpha and omega with missing and non-normal data (R package version 0.5) [Computer software]. Retrieved from https://CRAN.R-project.org/package=coefficientalpha

APPENDIX A

STATEMENTS OF PERMISSION FOR PUBLISHED WORKS

1. Samantha F. Anderson and Michael C. Edwards have granted permission for use of the following published work in this dissertation:

Manapat, P. D., Anderson, S. F., & Edwards, M. C. (2022). A revised and expanded taxonomy for understanding heterogeneity in research and reporting practices. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000488

2. Michael C. Edwards, David P. MacKinnon, Russell A. Poldrack, and Lisa A. Marsch have granted permission for use of the following published work in this dissertation:

Manapat, P. D., Edwards M. C., MacKinnon, D. P., Poldrack, R. A., & Marsch, L. A. (2021). A psychometric analysis of the Brief Self-Control Scale. *Assessment*, *28*(2), 395-412. https://doi.org/10.1177/1073191119890021