Article

A Psychometric Analysis of the Brief Self-Control Scale

Assessment
1–18

DOI: 10.1177/1073191119890021
journals.sagepub.com/home/asm
SAGE

Patrick D. Manapat[1], Michael C. Edwards[1], David P. MacKinnon[1], Russell A. Poldrack[2], and Lisa A. Marsch[3]

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.

Keywords

factor analysis, item response theory, brief self-control scale, scale evaluation, psychometrics

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

[1]Arizona State University, Tempe, AZ, USA
[2]Stanford University, Stanford, CA, USA
[3]Geisel School of Medicine at Dartmouth College, Hanover, NH, USA

Corresponding Author:
Patrick D. Manapat, Department of Psychology, Arizona State University, 950 South McAllister Avenue, Tempe, AZ 85281, USA.
Email: pmanapat@asu.edu

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).

"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 ($\sqrt{1/\mathrm{INF}}$) 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.

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.

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.

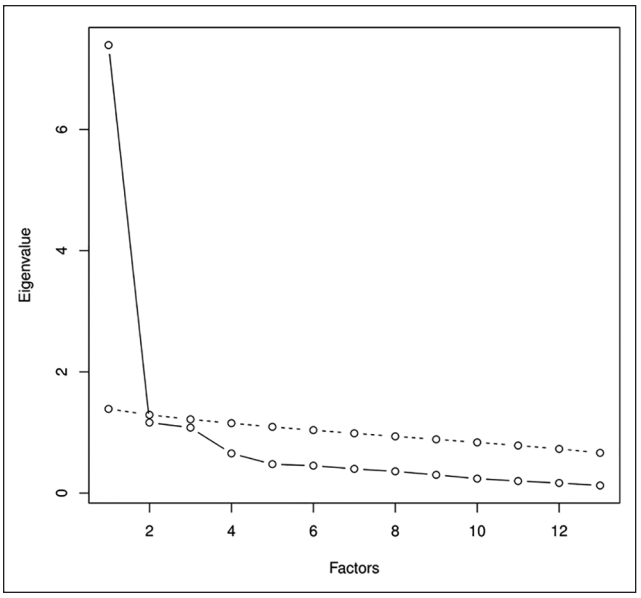

Figure 1. Solid line is scree plot from data (exploratory half of Sample 1; $n = 261$) and dotted line is random component from parallel analysis.

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].

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.

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

Table 5. Confirmatory Factor Analysis: Model Fit Summary.

Model	χ^2	*df*	TLI	CFI	SRMR	RMSEA	90% RMSEA 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]
Holdout 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 $p < .001$.

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

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.

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: $\sqrt{1/\text{INF}}$. 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.

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). Intercepts are c_1-c_4. The interfactor correlation was equal to .83.

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

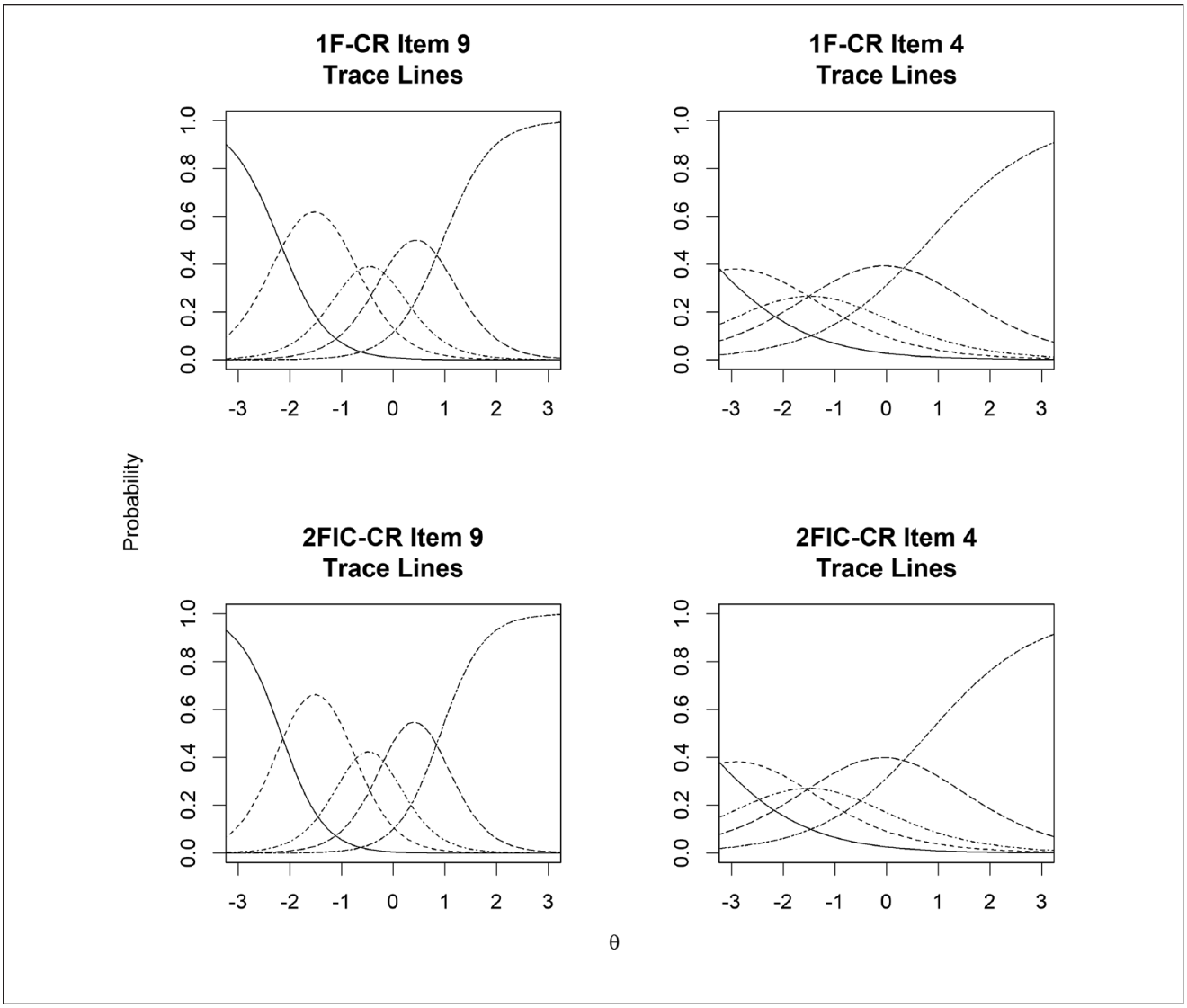

Figure 2. Trace lines for Items 9 and 4 from the one-factor (1F-CR) and two-factor (2FIC-CR) models.
Note. 1F-CR = 1F model with three residual covariances; 2FIC-CR = 2FIC model with three residual covariances.

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

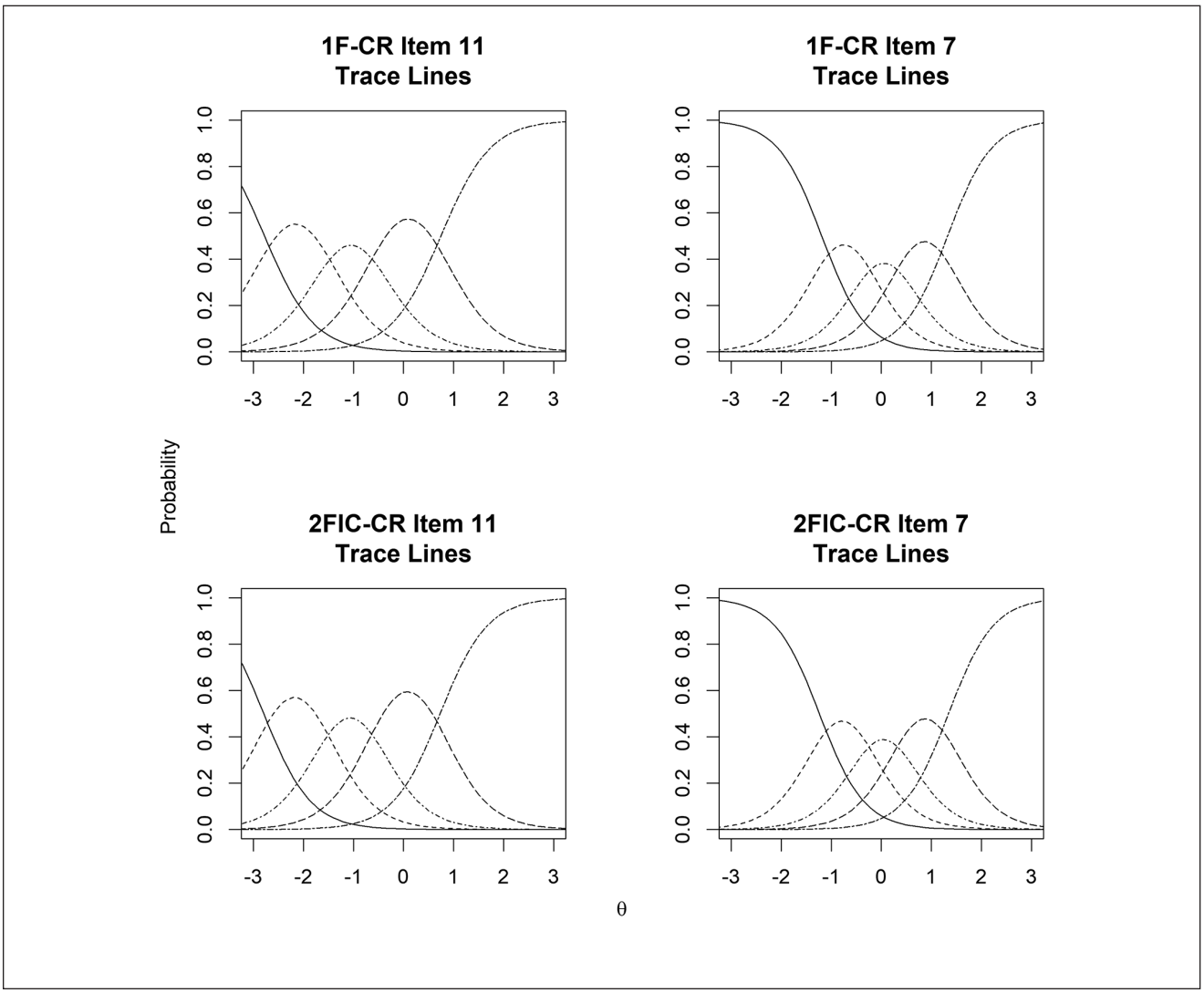

Figure 3. Trace lines for Items 11 and 7 from the one-factor (1F-CR) and two-factor (2FIC-CR) models.
Note. 1F-CR = 1F model with three residual covariances; 2FIC-CR = 2FIC model with three residual covariances.

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.

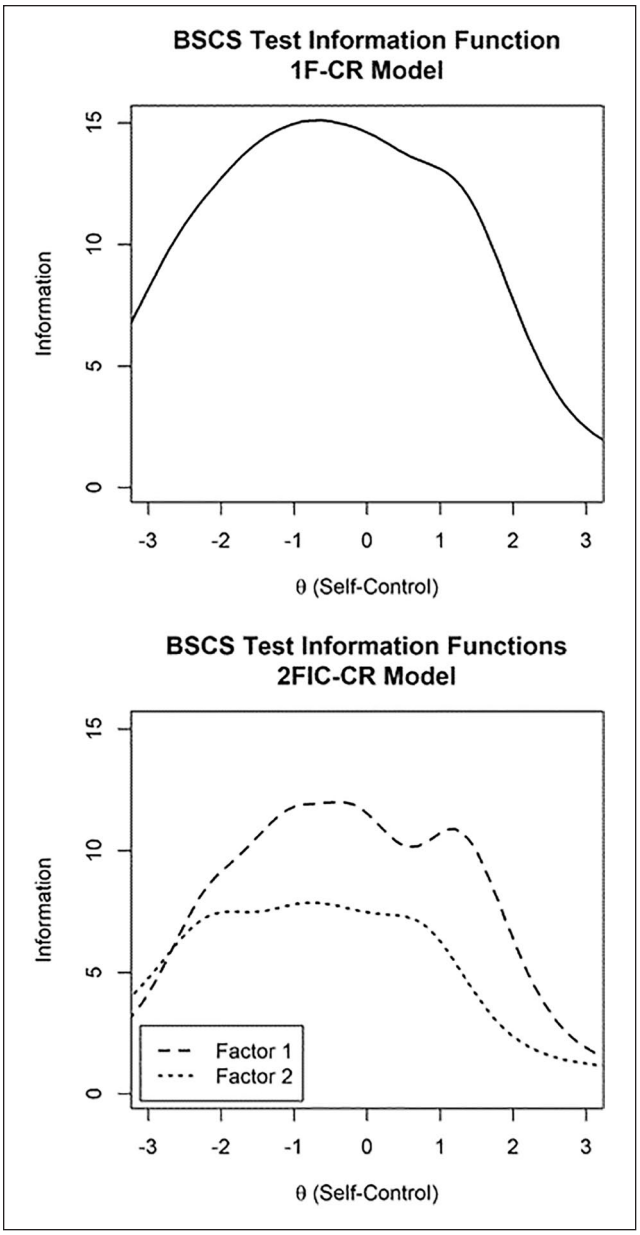

Figure 4. Information functions for the 13-item BSCS.
Note. BSCS = Brief Self-Control Scale; 1F-CR = 1F model with three residual covariances; 2FIC-CR = 2FIC model with three residual covariances. The top plot represents the information function for the 1F-CR model. The bottom plot represents the information 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.

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

Table 9. Regression of Validation Measures From Initial Data Collection Onto EAP Scores (Sample 1).

1F-CR EAP Scores						
Outcome	Predictor	N	Adj. R^2 or 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	1417.20	−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	1570.70	−0.12	0.04	−2.74*
2FIC-CR EAP Scores						
Outcome	Predictor	N	Adj. R^2 or 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	1415.60	−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	1568.60	−0.26	0.10	−2.63*
	Factor 2 EAP			0.14	0.10	1.44

Note: EAP = expected a posteriori; 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 adjusted R^2). *$p < .05$.

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.

Table 10. Regression of Validation Measures From Retest Onto EAP Scores (Sample 1).

1F-CR EAP Scores						
Outcome	Predictor	N	Adj. R^2 or 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	.52	−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	Adj. R^2 or AIC	*b*	*SE*	*t*-value
Frequency of Alcohol Use	Factor 1 EAP	150	.05	−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	.43	−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; 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 adjusted R^2). Retest measures collected a mean of 111 days after initial data collection. *$p < .05$.

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.

Acknowledgments

A special thanks to June Tangney and Jeffrey Stuewig for their valuable feedback on this article and for sharing their 2012 data set supported in part by the National Institute on Drug Abuse Grant No. R01DA14694.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was partially supported by the National Institute on Drug Abuse of the National Institutes of Health under Award Number UH2DA041713. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ORCID iD

Patrick D. Manapat https://orcid.org/0000-0003-1027-6652

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. doi:10.1111/j.1745-3992.2003.tb00136.x

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. doi:10.1177/014662168200600405

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. doi:10.1177/0049124192021002005

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. doi:10.1207/s15327752jpa8501_02

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. doi:10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307-335. doi:10.3102/1076998609353115

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. doi:10.1007/BF02291477

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334. doi:10.1007/BF02310555

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. doi:10.1016/j.paid.2011.01.015

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. doi:10.1037/a0032670

Edwards, M. C. (2009). An introduction to item response theory using the Need for Cognition Scale. *Social and Personality Psychology Compass*, *3*, 507-529. doi:10.1111/j.1751-9004.2009.00194.x

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. doi:10.1037/met0000121

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. doi:10.1016/j.brat.2017.09.014

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. doi:10.1073/pnas.1818430116

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. doi:10.1037/1082-989X.4.3.272

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. doi:10.1080/15560350802712371

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. doi:10.1037/1082-989X.9.4.466

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. doi:10.3102/0013189X17725519

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. doi:10.1080/10705519909540118

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. doi:10.1016/j.paid.2015.07.006

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84-99. doi:10.1037/1082-989X.4.1.84

Maloney, P. W., Grawitch, M. J., & Barber, L. K. (2012). The multifactor structure of the Brief Self-Control Scale: Discriminant validity of restraint and impulsivity. *Journal of Research in Personality*, *46*, 111-115. doi:10.1016/j.jrp.2011.10.001

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. doi:10.1111/j.2044-8317.1970.tb00432.x

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. doi:10.1111/emip.12092

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. doi:10.1037/pas0000003

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184-188. doi:10.1177/0963721414531598

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*, 768-774. doi:10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

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. doi:10.1177/0013164417690198

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. doi:10.1177/0146621697211002

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 Assessment*, *25*, 631-642. doi:10.1037/a0032161

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. doi:10.1007/s11336-008-9101-0

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. doi:10.1111/j.0022-3506.2004.00263.x

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. E. Millsap & A. Maydeu-Olivares (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. doi:10.1177/2167702618757690

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58-79. doi:10.1037/1082-989X.12.1.58

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