Published in final edited form as:

Addict Behav. 2018 September ; 84: 126–130. doi:10.1016/j.addbeh.2018.04.005.

Methodological factors as a potential source of discordance between self-report and behavioral measures of impulsivity and related constructs

Jarrod M. Ellingson, PhD[1,2,*], **Marc N. Potenza, PhD, MD**[1,3,4,5,6], and **Godfrey D. Pearlson, MD**[1,3,7]

[1]Yale University School of Medicine, Department of Psychiatry, New Haven, CT, USA

[2]University of Colorado, Department of Psychology and Neuroscience, Boulder, CO, USA

[3]Yale University School of Medicine, Department of Neuroscience, New Haven, CT, USA

[4]Yale University School of Medicine, Yale Child Study Center, New Haven, CT, USA

[5]Yale University School of Medicine, the National Center on Addiction and Substance Abuse, New Haven, CT, USA

[6]Connecticut Mental Health Center, New Haven, CT, USA

[7]Olin Neuropsychiatry Center, Hartford Hospital, Hartford, CT, USA

Abstract

There is a consistent but poorly understood finding that self-report and behavioral measures of impulsivity are weakly correlated or uncorrelated. There are many possible explanations for this observation, including differences in how these instruments are administered and scored. The present study examined the utility of alternative scoring algorithms for self-report measures that aim to identify participants' peak impulsivity (or self-control), informed by estimates of item difficulty from Item Response Theory (IRT) analyses. College students were administered self-report questionnaires (Zuckerman Sensation Seeking Scale [ZSS], Barratt Impulsiveness Scale [BIS-11], behavioral measures related to risk-taking and impulsivity (Balloon Analog Risk Task [BART], Experiential Discounting Task [EDT]), and the substance use module of a clinical interview (past-six-month alcohol and marijuana use). IRT analyses were conducted on self-report measures to estimate item difficulty. Scoring algorithms ranked items by difficulty and scored items based on consecutive items endorsed or denied. A maximal scoring algorithm increased the concordance between the BIS-11 and BART (r = .08 vs. -.07), but there was no evidence of

[*]Correspondence regarding this article should be sent to: Jarrod Ellingson, PhD, Department of Psychology and Neuroscience, University of Colorado, Boulder, CO 80309, jarrod.ellingson@colorado.edu.

Conflict of Interest: The authors report no conflicts of interest with respect to the content of this manuscript.

Contributors: Author GDP designed and oversaw all data collection procedures. Author MNP was also involved in the study design and wrote portions of the manuscripts. Author D conducted the statistical analysis and wrote the first draft of the manuscript. All authors contributed to and have approved the final manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

increased incremental validity for substance-use. Findings suggest that methodological factors may help explain the poor concordance of self-report and behavioral measures of impulsivity, but the magnitude of these correlations remained quite small. Further, alternative scoring algorithms were correlated with substance use but did not explain any variance that was distinct from typical algorithms. Future directions are discussed for elucidating the discrepancy between self-report and behavioral measures of impulsivity-related constructs, such as using large self-report item pools to develop computer adaptive tests.

Keywords

Impulsivity; Measurement; Behavioral; Laboratory; Self-Report; Item Response Theory

Self-report and behavioral measures of impulsivity are often weakly correlated or uncorrelated in the empirical literature. Meta-analytic work has highlighted this problem, with a mean correlation between self-report and behavioral measures of impulsivity of just .10, regardless of whether measures assess the same or conceptually different facets of impulsivity (Cyders & Coskunpinar, 2012; Duckworth & Kern, 2011). That is, measures of ostensibly the same construct share approximately 1% of variation with each other. Although this finding has been found across samples and research groups, surprisingly little empirical work has investigated potential causes of this weak concordance. Understanding this measurement issue is particularly relevant to substance use and other addictive behaviors, for which impulsivity is among the most frequently studied risk factors. Further, this issue has broader relevance, as indicated by NIMH's efforts to identify and adequately assess transdiagnostic risk factors (e.g., cognitive control; Ellingson, Richmond-Rakerd, Statham, Martin, & Slutske, 2016) across multiple levels of measurement (e.g., self-report, behavior; Insel et al., 2010). While recognizing that many factors may drive the poor concordance among self-report and behavioral measures of impulsivity and related constructs, the current study focused on methodological differences related to how self-report and behavioral measures are administered and scored.

Self-report and behavioral measures differ in at least four important ways that may underlie their discordance (Willerman, Turner, & Peterson, 1976). First, self-report measures instruct participants to “be honest,” whereas behavioral measures instruct participants to “try hard” or “do (their) best.” Second, self-report measures may be monotonous, particularly for individuals high in impulsivity, resulting in varied motivation across participants. In contrast, behavioral measures may be more engaging, resulting in greater participant motivation. This feature may not hold for all behavioral measures as some may also be experienced as relatively monotonous (e.g., Go/No-Go or Stroop tasks). Third, all participants are administered the same items on self-report measures, but adaptive, computer-based behavioral measures may administer different item sets based on performance. Finally, self-report measures aim to identify participants' typical level on a given construct (e.g., typical impulsivity/self-control), and behavioral measures aim to identify participants' peak level on a given construct (e.g., maximal ability to exert self-control over impulses). These differences are not mutually exclusive. For example, behavioral measures may assess maximal ability, in part, because participants are instructed to “do their best.” The current study focused on methodological differences that may identify participants' typical versus maximal levels of impulsivity.

Traditionally, self-report measures of individual differences produce a score informed by all items (e.g., a sum or average), described as a *typical* score in that it assesses participants' typical functioning (Willerman et al., 1976). In contrast, behavioral measures identify a peak level, described as a *maximal* score in that it assesses participants' maximal functioning. Notable differences between typical and maximal measures have been highlighted. Willerman and colleagues (1976) used self-narratives (a short paragraph) to assess participants' anger susceptibility, including “what you *typically* would do and say in expressing your anger,” and “what you might be capable of doing and saying if you *maximally* expressed your anger.” Whereas the *typical* narrative was modestly correlated with a behavioral task of anger expression (*r*′s = -.03 – .32), the *maximal* narrative demonstrated moderate to strong correlations (*r*'s = .48 – .59). The nature of self-report used by Willerman and colleagues differs from self-report questionnaires commonly used to assess individual differences, but these findings suggest that assessing self-report by using maximal ability may increase the concordance with behavioral measures, relative to typical ability.

There have been limited published follow-up studies to these findings. In examining cognitive functioning, some have speculated that fluid intelligence, including constructs related to impulsivity (e.g., working memory, processing speed, reasoning), is better assessed by *maximal* than *typical* ability (see Ackerman & Kanfer, 2004 for a review). In relation to substance use, however, self-report measures tend to demonstrate stronger associations. For example, self-report measures of self-control are associated with (main effects) of alcohol use and physical activity, in models where behavioral measures are unrelated to these outcomes (Allom, Panetta, Mullan, & Hagger, 2016; Thush et al., 2008). One should note, that some findings more strongly link behavioral rather than self-report measures of impulsivity to substance use (e.g., cessation outcomes in treatment-seeking adolescent tobacco smokers (Krishnan-Sarin et al., 2007).

The current study applied a typical scoring algorithm (i.e., sum score) and two maximal scoring algorithms to self-report measures of impulsivity, in a sample that was also administered behavioral measures of impulsivity-related constructs. Thus, the concordance among self-report and behavioral measures of impulsivity was compared across algorithms. Item Response Theory (IRT) analyses were used to estimate item difficulty, from which results were incorporated into the maximal scoring algorithms. We hypothesized that the maximal scoring algorithms would increase the concordance of self-report and behavioral measures. Finally, we examined whether these alternative scoring algorithms are more strongly associated with substance use measures (alcohol and marijuana), relative to traditional algorithms.

Methods

Participants

Participants were 1,038 college students attending one of two universities (one public, one private) (*M* age = 18.4 years; 49.7% female; 84.6% European ancestry). All participants were administered self-report measures of impulsivity and substance use. A subset was administered behavioral measures to investigate alcohol and neurocognitive functioning, through the Brain and Alcohol Research in College Students study (BARCS; Dager et al., 2013). All data collection was approved by the Institutional Review Boards at both universities.

Measures

Self-Reported Impulsivity—Participants completed the 40-item Zuckerman Sensation Seeking Scale (ZSS; M. Zuckerman, Kolin, Price, & Zoob, 1964) and 30-item Barratt Impulsiveness Scale, Version 11 (BIS-11; Patton, Stanford, & Barratt, 1995). The ZSS is thought to assess general reward sensitivity, using forced-choice items between one of two preferences (e.g., "I like 'wild' uninhibited parties" vs. "I prefer quiet parties with good conversation"). The BIS-11 assesses impulsivity more broadly, using a 5-point scale of item endorsement, from *never/rarely* to *almost always/always*.

Behavioral Impulsivity—Participants completed two behavioral measures, the Balloon Analog Risk Task (BART; assessing behavioral risk-taking) and the Experiential Discounting Task (EDT; assessing behavioral intertemporal choice). The BART is a computer-based task that assesses risk-taking in the context of obtaining increasing reward within a trial, while the risk of losing the reward concurrently increases probability within that trial. Thus, the BART involves "actual risky behavior for which, similar to real-world situations, riskiness is rewarded up until a point at which further riskiness results in poorer outcomes" (Lejuez et al., 2002 p. 76). The total number of pumps was used as a measure of performance on the task. In contrast, the EDT assesses delay-discounting (Reynolds & Schiffbauer, 2004), the tendency for an individual to forego a small but immediate reward for a larger but more distal and probabilistic reward.

Substance Use—Alcohol and marijuana use were assessed using quantity measures from the substance use disorders module of the Structured Clinical Interview for DSM-IV-TR (First, Spitzer, Gibbon, & Williams, 2002). Alcohol use was measured as the number of drinks consumed within the last six months. Marijuana use was measured as the number of joints consumed in the last six months. Therefore, both measures assessed past-six-month consumption.

Analytic Procedures

IRT Analyses—IRT analyses were conducted in R using the mirt package (Chalmers, 2012). The check.monotonicity function from the mokken package in R was used to examine the monotonicity assumption of IRT analyses (Van der Ark, 2007). Three items from the ZSS and three items from the BIS-11 violated the assumption of monotonicity, and 37 items from the ZSS and 27 items from the BIS-11 were retained for subsequent analyses.

Further, Horn's analyses (or parallel analyses) suggested that both the ZSS and BIS-11 are multidimensional (Horn, 1965).

Multidimensional IRT models were conducted using a four-factor structure for the ZSS and a six-factor structure for the BIS-11, consistent with empirical and theoretical work on these scales (Reise, Moore, Sabb, Brown, & London, 2013; Marvin Zuckerman, 1971). Two-parameter IRT models estimated item difficulty and discrimination. Based on the item response format, a dichotomous response was modeled for the ZSS, and a polytomous response was modeled for the BIS-11.

Scoring Algorithms—Two alternative scoring algorithms for self-report measures (ZSS and BIS-11) were examined, informed by item difficulty estimates from IRT analyses; one algorithm was based items in ascending order of difficulty and one based on items in descending order of difficulty. Item difficulty provides an estimate of the latent trait value at which 50% of the sample will endorse the item. For example, assuming a standardized latent variable, an item on the ZSS with a difficulty of 1.96 (i.e., corresponding to a *z*-score of 1.96) suggests that participants who endorse the item fall, on average, at the 95th percentile of the underlying trait. Given the polytomous structure of the BIS-11, there were multiple difficulties for each item, and the item difficulty at the threshold corresponding to an endorsement of at least *occasionally* (i.e., 2 on a 0-4 scale) was used to inform the scoring algorithms.

First, a scoring algorithm was adapted from the Wechsler Adult Intelligence Scale, wherein items are ordered by ascending difficulty (WAIS; Wechsler, 1981). Items are administered and scored until three consecutive items are answered incorrectly. This *WAIS algorithm* was applied to the BIS-11 and ZSS, with a denial of impulsivity counting as an “incorrect” response. Endorsements were scored until three consecutive items were denied.

Second, a *maximal algorithm* was implemented, wherein items were ordered by descending difficulty. Once two consecutive items were endorsed, the algorithm multiplied the mean score of all prior items by the number of remaining items. For example, if the two most difficult items on the ZSS were endorsed, a score of 37 was generated (i.e., [[1+1]/2]*37).

A typical scoring algorithm (i.e., a sum score) was also applied to the BIS-11 and ZSS. Correlations were moderate to strong between the typical and alternative algorithms but tended to be higher for the WAIS algorithm ($r = .83 – .84$) than the maximal algorithm ($r = .55 – .58$).

Correlation Analyses—All subsequent analyses were conducted in Mplus (Muthén & Muthén, 1998). Participants' university was included as a cluster-level variable in all analyses, which adjusts for potentially biased standard errors due to non-independent observations in cluster sampling (i.e., students from distinct universities) (Rebollo, do Moor, Dolan, & Boomsma, 2006). Participant age, gender, and race/ethnicity were included as covariates. Correlation analyses investigated associations between these algorithms and performance on the BART and EDT. Differences between correlations were evaluated using chi-square difference tests, between models in which correlations were freely estimated (i.e., typical self-report and behavioral measure, maximal self-report and behavioral measures) and constrained to be equal. A significant decrement of fit in the constrained model would suggest a statistically significant difference between the use of the typical and alternative algorithm.

Incremental Validity—Cholesky models were conducted to decompose the proportion variation in substance-use outcomes into the typical scoring algorithm, and any remaining variation attributable to the alternative-scoring algorithm. That is, after accounting for the typical scoring algorithm, the proportion of variation in substance use attributable to the maximal and WAIS scoring algorithms was estimated. If statistically significant, this would suggest that the alternative algorithm accounts for variance not captured by the typical algorithm.

Results

Correlation Analyses

Raw, unadjusted correlations are shown in Table 1, to provide a benchmark for associations among self-report impulsivity measures using typical scoring algorithms (ZSS, BIS-11), behavioral impulsivity measures (EDT, BART), and substance use (alcohol, marijuana). Self-report measures were moderately correlated with each other ($r = .46$), and behavioral measures were weakly correlated ($r = .15$). Consistent with prior studies, self-report and behavioral impulsivity measures were weakly correlated (*r*s = -.07 – .09). Further, substance use measures appeared to be more strongly correlated with self-reported impulsivity (*r*s = . 23 – .44) than behavioral measures of impulsivity (*r*s = -.10 – .17).

Correlations between behavioral measures of impulsivity and self-report scores derived from *typical, maximal,* and the traditional *WAIS* scoring algorithms are displayed in Table 2. Consistent with prior work, there were weak correlations between behavioral and self-report measures of impulsivity-related constructs using the typical scoring algorithm. In fact, using typical scoring algorithms for self-report measures, only two of the four correlations reached statistical significance. The ZSS was positively correlated with the EDT ($r = .09$), and the BIS-11 was *negatively* correlated with the BART ($r = -.07$). The *WAIS algorithm* yielded similar correlations as the *typical algorithm,* and these estimates were not significantly different from each other in constrained models. The *maximal algorithm,* however, yielded a correlation for the BIS-11 and BART that was significantly different than the typical algorithm ($r = .08$) and in the expected direction ($\Delta\chi^2_{(1)} = 4.13$, $p < .05$). Thus, the *maximal scoring* algorithm increased the concordance between the BIS-11 and BART and yielded similar correlations as the *typical algorithm* in all other models.

Incremental Validity

Table 3 displays the proportion of variation explained by self-reported impulsivity, across the scoring algorithms, in behavioral measures of impulsivity and substance use. Consistent with prior studies, the *typical algorithm* for self-report measures explained approximately 1% of the variance in behavioral measures of impulsivity. Notably, the *maximal algorithm* for BIS-11 explained an additional 2% of the variation in the BART that was not accounted for by the *typical algorithm*. The *WAIS algorithm* failed to account for a significant proportion of variation in the EDT or BART, after accounting for the *typical algorithm*; however, the *WAIS algorithm* for the ZSS explained 1%-2% of variation in these measures (*p*s = .07 – .37. Altogether, over 95% of the variation in EDT and BART performance was unexplained by self-report measures, regardless of the scoring algorithm.

For substance use, the *typical algorithm* of the ZSS accounted for a large and statistically significant proportion of variation in substance use (10% - 17% [SEs = 1.9% - 2.3%], *p*s < .001). Similarly, the BIS-11 accounted for a substantial proportion of variation in marijuana use (4.3% [SE = 2.1%], *p* = .04), and a large but statistically nonsignificant proportion in alcohol use (7.4% [SE = 5.7%], *p* = .19). After accounting for the *typical algorithm*, a negligible proportion was accounted for by either the *maximal* or *WAIS* algorithms (<1%). Thus, the *maximal* and *WAIS algorithms* did not explain any variation in substance use outcomes not already accounted for by the typical algorithm.

Discussion

The current study suggests that methodological factors may underlie some of the discrepancies between self-report and behavioral measures of impulsivity-related assessments. In particular, a scoring algorithm that identifies an individual's maximal, or peak, levels of impulsivity or impulse control may improve the concordance between self-report and behavioral measures, although the incremental proportion of the variance accounted for by this approach is small, or in most cases, not statistically significant. It has been suggested that behavioral measures reflect peak, rather than typical, functioning, and it is possible that these algorithms increased concordance by assessing peak functioning (i.e., impulse control) rather than typical functioning (Ackerman & Kanfer, 2004). Notably, associations with substance-use outcomes were not improved using the alternative scoring algorithms of self-report measures.

Further researcher is needed to explore the many possible explanations for the discordance between self-report and behavioral measures of impulsivity-related constructs. While a maximal scoring algorithm yielded some benefit in the current study, there may be greater advantages if self-report measures of impulsivity-related constructs are developed with such methodological factors (e.g., see Willerman et al., 1976). There is already a vast literature on the utility of IRT models for developing measures that cover a wide spectrum of item difficulty, which could be implemented to more adequately assess impulsivity in ways that may correspond with laboratory tasks (Mellenbergh, 1994; Smith & McCarthy, 1995; Smith, McCarthy, & Anderson, 2000; Steinberg & Thissen, 1995; Thissen & Steinberg, 1988). For example, impulsivity scales developed from comprehensive items pools and informed by IRT analyses may further increase the concordance between behavioral and self-report measures. Statistical packages have recently been developed to aid such efforts, including the administration of computer adaptive tests (Magis & Barrada, 2017). Further, it may be worth investigating how behavioral measures can be made more like self-report measures.

These findings may suggest that self-report and behavioral measures assess different aspects of the same construct, such as trait-like differences (via self-report) and state-like processes (via behavioral measures) (Allom et al., 2016); however, others have conceptualized self-report measures as subjective and explicit, whereas behavioral measures are objective and implicit (Dislich, Zinkernagel, Ortner, & Schmitt, 2015; Ortner & Proyer, 2014). As noted by Willerman and colleagues (1976), self-report measures typically assess stable, trait-like constructs, and behavioral measures may be more state-like. Analytic approaches incorporating multiple time points of data, such as state-trait models, may be used to investigate this possibility. Additionally, the domains assessed with different instruments (impulsiveness, sensation-seeking, risk-taking, delay-discounting, as assessed in the current study) may vary. Such limitations warrant further investigation in future studies. An additional limitation in the current study involves the multiple comparisons conducted; as such, the study should be considered exploratory. That being said, the largely null findings suggest that the methodological considerations explored do not account for a substantial amount of the variation in the relationships between the self-report and behavioral measures.

Distinguishing self-report and behavioral measures of impulsivity conceptually (e.g., explaining state vs. trait risk), as well as empirically (i.e., weakly correlated), may also provide some benefits to understanding substance use. For example, models that incorporate both types of measurement methods, as weakly correlated risk factors, may explain distinct and meaningful variance in substance use, compared to models that include highly concordant risk factors that are less conceptually distinct. This potential benefit does not negate, however, the utility of understanding why different measures impulsivity are discordant.

Summary

The current findings suggest that methodological factors related to instrument administration and scoring may in part contribute to the poor concordance among self-report and behavioral measures of impulsivity, but that these contributions at most contribute only a small amount of the variance. As such, other factors clearly underlie this discordance and warrant further investigation. Further, alternative scoring algorithms were correlated with substance use but did not explain any variance that was distinct from typical algorithms. Thus, other possibilities should be explored.

Acknowledgments

Role of Funding Sources: This work was supported by NIAAA (R01 AA016599), the National Center for Responsible Gaming, and the National Center on Addiction and Substance Abuse. The funding agencies did not provide input or comment on the content of the manuscript, and the content of the manuscript reflects the contributions and thoughts of the authors and do not necessarily reflect the views of the funding agencies.

Disclosures: Dr. Potenza has consulted for Ironwood, Lundbeck, Boehringer Ingelheim, Shire, INSYS, RiverMend Health, Opiant/Lakelight Therapeutics and Jazz Pharmaceuticals; has received research support from the National Institutes of Health, Veteran's Administration, Mohegan Sun Casino, the National Center for Responsible Gaming, Pfizer, and Psyadon Pharmaceuticals; has participated in surveys, mailings or telephone consultations related to drug addiction, impulse control disorders or other health topics; has consulted for gambling and legal entities on issues related to impulse control disorders; provides clinical care in the Connecticut Department of Mental Health and Addiction Services Problem Gambling Services Program; has performed grant reviews for the National Institutes of Health and other agencies; has guest-edited journal sections; has given academic lectures in grand rounds, CME events and other clinical or scientific venues; and has generated books or book chapters for publishers of mental health texts.

References

Ackerman, PL., Kanfer, R. Cognitive, affective, and conative aspects of adult intellect within a typical and maximal performance framework. In: Dai, DY., Sternberg, RJ., editors. Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2004. p. 119-141.

Allom V, Panetta G, Mullan B, Hagger MS. Self-report and behavioural approaches to the measurement of self-control: Are we assessing the same construct? Personality and Individual Differences. 2016; 90:137–142.

Chalmers RP. mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software. 2012; 48(6):1–29.

Cyders MA, Coskunpinar A. The relationship between self-report and lab task conceptualizations of impulsivity. Journal of Research in Personality. 2012; 46(1):121–124.

Dager A, Anderson BM, Stevens MC, Pulido C, Rosen R, Jiantonio-Kelly RE, et al. Pearlson GD. Influence of Alcohol Use and Family History of Alcoholism on Neural Response to Alcohol Cues in College Drinkers. Alcoholism, Clinical and Experimental Research. 2013; 37(Suppl 1):E161–E171. https://doi.org/10.1111/j.1530-0277.2012.01879.x.

Dislich FX, Zinkernagel A, Ortner TM, Schmitt M. Convergence of direct, indirect, and objective risk-taking measures in gambling. Zeitschrift Für Psychologie/Journal of Psychology. 2015

Duckworth AL, Kern ML. A meta-analysis of the convergent validity of self-control measures. Journal of Research in Personality. 2011; 45(3):259–268. https://doi.org/10.1016/j.jrp.2011.02.004. [PubMed: 21643479]

Ellingson JM, Richmond-Rakerd LS, Statham DJ, Martin NG, Slutske WS. Most of the genetic covariation between major depressive and alcohol use disorders is explained by trait measures of negative emotionality and behavioral control. Psychological Medicine. 2016; 46(14):2919–2930. https://doi.org/10.1017/S0033291716001525. [PubMed: 27460396]

First, MB., Spitzer, RL., Gibbon, M., Williams, JBW. Structured Clinical Interview for DSM-IV-TR Axis I Disorders--Research Version, Non-Patient Edition (SCID-I/NP, 11/2002 Revision). New York: Biometrics Research Department, New York State Psychiatric Institute; 2002.

Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965; 30:179–185. [PubMed: 14306381]

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Wang P. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. The American Journal of Psychiatry. 2010; 167(7):748–751. https://doi.org/10.1176/appi.ajp.2010.09091379. [PubMed: 20595427]

Krishnan-Sarin S, Reynolds B, Duhig AM, Smith A, Liss T, McFetridge A, et al. Potenza MN. Behavioral impulsivity predicts treatment outcome in a smoking cessation program for adolescent smokers. Drug and Alcohol Dependence. 2007; 88:79–82. [PubMed: 17049754]

Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, Stuart GL, et al. Brown RA. Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). Journal of Experimental Psychology Applied. 2002; 8(2):75–84. [PubMed: 12075692]

Magis D, Barrada JR. Computerized adaptive testing with R: Recent updates of the package catR. Journal of Statistical Software. 2017; 76(1):1–19.

Mellenbergh GJ. Generalized linear item response theory. Psychological Bulletin. 1994; 115(2):300–307. https://doi.org/10.1037/0033-2909.115.2.300.

Muthén, LK., Muthén, BO. Mplus User's Guide. 7th. Los Angeles, CA: Muthén & Muthén; 1998.

Ortner TM, Proyer RT. Objective personality tests. Manuscript Submitted for Publication. 2014

Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt Impulsiveness Scale. Journal of Clinical Psychology. 1995; 51(6):768–774. [PubMed: 8778124]

Rebollo I, do Moor MH, Dolan CV, Boomsma DI. Phenotypic factor analysis of family data: Correction of the bias due to dependency. Twin Research and Human Genetics. 2006; 9(3):367–376. https://doi.org/10.1375/twin.9.3.367. [PubMed: 16790147]

Reise SP, Moore TM, Sabb FW, Brown AK, London ED. The Barratt Impulsiveness Scale–11: Reassessment of its structure in a community sample. Psychological Assessment. 2013; 25(2): 631–642. [PubMed: 23544402]

Reynolds B, Schiffbauer R. Measuring state changes in human delay discounting: an experiential discounting task. Behavioural Processes. 2004; 67(3):343–356. https://doi.org/10.1016/j.beproc.2004.06.003. [PubMed: 15518985]

Smith GT, McCarthy DM. Methodological considerations in the refinement of clinical assessment instruments. Psychological Assessment. 1995; 7(3):300–308.

Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychological Assessment. 2000; 12(1):102. [PubMed: 10752369]

Steinberg L, Thissen D. Item response theory in personality research. Personality Research, Methods, and Theory: A Festschrift Honoring Donald W Fiske. 1995:161–181.

Thissen D, Steinberg L. Data analysis using item response theory. Psychological Bulletin. 1988; 104(3):385–395. https://doi.org/10.1037/0033-2909.104.3.385.

Thush C, Wiers RW, Ames SL, Grenard JL, Sussman S, Stacy AW. Interactions between implicit and explicit cognition and working memory capacity in the prediction of alcohol use in at-risk adolescents. Drug and Alcohol Dependence. 2008; 94(1–3):116–124. [PubMed: 18155856]

Van der Ark LA. Mokken scale analysis in R. Journal of Statistical Software. 2007; 20(11):1–19.

Wechsler, D. Wechsler adult intelligence scale-revised. Psychological Corporation; 1981.

Willerman L, Turner RG, Peterson M. A comparison of the predictive validity of typical and maximal personality measures. Journal of Research in Personality. 1976; 10(4):482–492. https://doi.org/10.1016/0092-6566(76)90063-5.

Zuckerman M, Kolin EA, Price L, Zoob I. Development of a sensation-seeking scale. Journal of Consulting Psychology. 1964; 28(6):477–482. [PubMed: 14242306]

Zuckerman, Marvin. Dimensions of sensation seeking. Journal of Consulting and Clinical Psychology. 1971; 36(1):45–52.

Highlights

- Self-report and behavioral measures of impulsivity are only weakly concordant.
- We examined whether a scoring algorithm of peak impulsivity improves concordance.
- Assessing peak, instead of typical, impulsivity improved concordance for some measures but explained only a small proportion of variance.

Table 1

Correlation estimates among self-report and behavioral measures of impulsivity-related assessments and substance use.

		ZSS	BIS-11	EDT	BART	Alcohol Use
	M (SD)	**Pearson Correlation Estimates (Standard Errors)**				
ZSS	19.99 (6.21)					
BIS-11	65.05 (11.31)	.46 (.03)				
EDT	0.65 (0.13)	.07 (.02)	.03 (.01)			
BART	27.12 (9.65)	.09 (.05)	-.07 (.01)	.15 (.01)		
Alcohol Use Quantity	21.22 (31.40)	.44 (.04)	.30 (.08)	.02 (.01)	-.10 (.02)	
Marijuana Use Quantity	2.66 (6.00)	.35 (.05)	.23 (.04)	.17 (.06)	.01 (.03)	.41 (.03)

Note. Substance use quantity was the total number of drinks for alcohol and joints smoked for marijuana in the last six months. Abbreviations: ZSS: Zuckerman Sensation Seeking Scale; BIS-11: Barratt Impulsiveness Scale-Version 11; EDT: Experiential Discounting Task; BART: Balloon Analog Risk Task performance, modeled by total pumps.

Table 2

Correlations estimates for the Experiential Discounting Task and Balloon Analog Risk Task with the Barratt Impulsiveness Scale-Version 11 and Zuckerman Sensation Seeking Scale using Typical, Maximal, and WAIS scoring algorithms.

Behavioral Measure	**Typical**	**Maximal**	**WAIS**
	Pearson Correlation Estimates (Standard Errors)		
	BIS-11		
EDT	.02 (.02)	.06 (.05)	-.01 (.02)
BART ^{M}	-.07 (.03) **	.08 (.01) ***	-.08 (.01) ***
	ZSS		
EDT	.09 (.02) ***	.12 (.01) ***	.03 (.01) *
BART	.08 (.05)	.08 (.05)	.02 (.03)

Note.

^{M}statistically significant difference, with maximal scoring algorithm demonstrating greater concordance with behavioral measure than typical algorithm.

****p* < .001,

***p* < .01,

**p* < .05.

Abbreviations: ZSS: Zuckerman Sensation Seeking Scale; BIS-11: Barratt Impulsiveness Scale-Version 11; EDT: Experiential Discounting Task; BART: Balloon Analog Risk Task performance, modeled by total pumps.

Table 3

Estimates of the proportion of variation explained by typical and alternative scoring algorithms in behavioral measures of impulsivity and substance use.

	Proportion of Variance Explained by:	**Remaining Proportion of Variance Explained by:**	
Outcome	**Typical Algorithm**	**Maximal Algorithm**	**WAIS Algorithm**
	BIS-11		
Behavioral Measures			
EDT	.00 (.00)	.00 (.01)	.00 (.00)
BART	.01 (.00) ***	.02 (.00) ***	.00 (.00)
Past Six -Month Substance Use			
Alcohol Quantity	.07 (.06)	.00 (.00)	.00 (.00)
Marijuana Quantity	.04 (.02) *	.00 (.00)	.00 (.00)
	ZSS		
Behavioral Measures			
EDT	.01 (.00) *	.01 (.01)	.01 (.01)
BART	.01 (.01)	.00 (.00)	.02 (.01)
Past Six -Month Substance Use			
Alcohol Quantity	.17 (.02) ***	.00 (.01)	.00 (.00)
Marijuana Quantity	.10 (.02) ***	.00 (.00)	.00 (.00)

Note.

$^{***}p < .001$,

$^{**}p < .01$,

$^{*}p < .05$.

Standard errors are listed in parentheses. Estimates listed as .00 are less than .01. Alcohol quantity was the total number of drinks in the last six months. Marijuana quantity was the total number of joints smoked in the last six months. Abbreviations: BART: Balloon Analog Risk Task; BIS-11: Barratt Impulsiveness Scale-Version 11; EDT: Experiential Discounting Task; ZSS: Zuckerman Sensation Seeking Scale