OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 33 (2018) 125–129

Archives of CLINICAL NEUROPSYCHOLOGY

Test–Retest Reliability of Common Behavioral Decision Making Tasks

Melissa T. Buelow*, Wesley R. Barnhart

Department of Psychology, The Ohio State University Newark, Newark, OH, USA

*Corresponding author at: Department of Psychology, The Ohio State University Newark, 1179 University Drive, Newark, OH 43055, USA. Tel.: +1-740-755-7808; fax: +1-740-366-5047.
E-mail address: buelow.11@osu.edu (M.T. Buelow).

Editorial Decision 3 April 2017; Accepted 5 April 2017

Abstract

Objective: To examine test–retest reliability of common behavioral decision making tasks.
Method: A total of 98 undergraduate students completed two administrations of the Iowa Gambling Task (IGT), Balloon Analogue Risk Task (BART), Columbia Card Task (CCT), and Game of Dice Task (GDT), three weeks apart.
Results: The BART, CCT, and GDT showed moderately strong correlations across time. On the IGT, no correlations were seen between Time 1 Trials 1–40 and Time 2 performance; however, weak correlations were observed between Time 1 Trials 41–100 and Time 2 performance. Paired-samples *t*-tests indicated participants were riskier at Time 1 than Time 2 on the IGT and GDT, but riskier at Time 2 on the BART.
Conclusions: The BART, CCT, and GDT showed moderate test–retest reliability, with the IGT showing weak reliability during the decision making under risk trials only. Implications for repeated test administration in clinical and non-clinical settings are discussed.

Keywords: Decision making; Reliability; Iowa Gambling Task; Balloon Analogue Risk Task; Columbia Card Task; Game of Dice Task

Psychologists are increasingly utilizing behavioral measures to assess decision making in clinical- and lab-based settings. Decision making is an executive function, linked with the prefrontal cortex (Bechara, 2007). Decision making impairments are shown in a variety of physical health, mental health, and neurological disorders (see Buelow & Suhr, 2009, for review), indicating the importance of understanding the statistical reliability and validity of decision making measures. Reliability is a necessary component in determining the validity of a neuropsychological task. Accurate assessment of a construct across time is important, as clinicians often conduct repeat assessments and researchers conduct pre/post-treatment evaluations tracking cognitive changes. Tasks with high test–retest reliability may provide evidence of a more stable versus fluctuating trait/characteristic, whereas tasks with low reliability may provide unstable/situational estimates. Executive functions are a relatively stable construct, as is decision making more specifically (Bechara, 2007), indicating the need for measurements of the construct to exhibit reliability across time.

The Iowa Gambling Task (IGT; Bechara, 2007) is a popular task used to assess decision making deficits in clinical populations when compared to healthy controls (see Buelow & Suhr, 2009, for review). Behavioral measures are often used with patient populations who may be relatively unaware of decision making deficits, which would in turn potentially limit validity of self-report questionnaire responses (Bechara, 2007). Previous research with the IGT indicates improved performance across repeated administrations in the same session (Lejuez et al., 2003), as well as over the course of varying time periods (one week to five years; Burdick, Braga, Gopin, & Malhotra, 2014; Waters-Wood, Xiao, Denburg, Hernandez, & Bechara, 2012). In specific studies of test–retest reliability, moderate to strong correlations are seen with IGTs administered several weeks (r = .35–.65; Xu, Korczykowski, Zhu, & Rao, 2013), months (r = .43–.47; Cardoso et al., 2010), or years apart (r = .19–.74; Tuvblad et al., 2013; Xiao et al., 2013). Others find no correlations on the IGT across time (four weeks; De Wilde, Bechara, Sabbe, Hulstijn, & Dom, 2013); however, in this particular study, IGT performance was examined before and after treatment for polysubstance addiction and the lack of correlation could be due to treatment effects.

doi:10.1093/arclin/acx038 Advance Access publication on 20 April 2017

In recent years, other behavioral decision making tasks have become popular, including the Balloon Analogue Risk Task (BART; Lejuez et al., 2002), Columbia Card Task (CCT; Figner, Mackinlay, Wilkening, & Weber, 2009), and Game of Dice Task (GDT; Brand et al., 2005) (see *Methods* for task descriptions). However, to date these tasks have primarily been used in lab versus clinical settings. Previous research suggests minimal relationships between performance on these tasks and the IGT (e.g., Lejuez et al., 2003), with a factor analysis suggesting each measures a separate type of decision making (Buelow & Blaine, 2015). Test–retest reliability, and the potential for practice effects, has not been extensively studied for these tasks. Several studies show moderate to high correlations between BART administrations over the course of several days (r = .79; Weafer, Baggott, & de Wit, 2013), weeks (r = .66–.78; White, Lejuez, & de Wit, 2008; Xu et al., 2013), and years (r = .33–.75; Collado, Felton, MacPherson, & Lejuez, 2014). However, no studies have examined test–retest reliability for the CCT or GDT. As decision making can be a viable outcome measure in studies of medication and other treatment effects (e.g., Abbate-Daga, Buzzichelli, Marzola, Amianto, & Fassino, 2012), it is important to know whether these tasks show practice effects from the first administration to follow-up. In addition, establishing test–retest reliability of measures used in lab and clinical settings is vital to ensuring a solid scientific basis for research and assessment of clinical outcomes. The aim of the present study is to examine relationships between decision making task performance on the IGT, BART, CCT, and GDT across two administrations.

Methods

Participants

A sample of 173 psychology students (61% female, M_{age} = 19.15 [SD_{age} = 2.85], M_{ed} = 12.15 [SD_{ed} = 0.45], 65.2% Caucasian) at a regional campus of a large Midwestern university completed the first task administration. Of these, 120 returned for a second administration (69.4% response rate). However, 22 were removed from analyses due to having scores on fewer than three decision making tasks at Time 2 (due to computer malfunction or time constraints). Thus, all analyses were conducted on the 98 participants who completed at least three decision making tasks at both Time 1 and Time 2 (55.9% female, M_{age} = 19.24 [SD_{age} = 3.29], M_{ed} = 12.13 [SD_{ed} = 0.45], 56.5% Caucasian). There were no differences in demographic or decision making variables between study drop-outs and study completers (ps > .08). In addition, there were no differences between the participants excluded at Time 2 and the study completers (ps > .20).

Materials and Procedure

The study was approved by the university's Institutional Review Board, and participants provided informed consent. Participants completed the standard IGT, BART, CCT, and GDT in a counterbalanced order (no order effects were found, ps > .10). Participants then returned to the lab three weeks later, on average, to complete a second administration of these tasks (M_{days} = 21.87, SD_{days} = 2.71). The three-week interval was determined based on retest intervals from previous studies (De Wilde et al., 2013; Weafer et al., 2013; White et al., 2008; Xu et al., 2013) and after taking into account time-based restrictions for participants on our campus. Many students transfer to a different campus to complete their degree, so having a retest interval that stretched into a second academic semester would limit recruitment efforts. Participants were debriefed at the end of the study.

The standard computerized IGT was administered (Bechara, 2007). Briefly, participants were instructed to maximize profit by selecting 100 times from one of four decks (A, B, C, D). Unbeknownst to the participants, and learned through trial feedback, two of the decks are disadvantageous (A, B), with high immediate gains but long-term negative outcomes, and two of the decks are advantageous (C, D), with lower immediate gains but long-term positive outcomes. Over the course of the task, participants should learn to avoid the disadvantageous decks. Previous research has shown IGT performance can be split into early trials, when decisions are made without knowledge about the risks/benefits of each deck (decision making under ambiguity), and later trials, when individuals have learned about the decks and instead make decisions under risk (Brand, Recknor, Grabenhorst & Bechara, 2007). For the present study, net scores ([A + B] – [C + D]) were calculated separately for the decision making under ambiguity (Trials 1–40) and decision making under risk (Trials 41–100) trials. More negative scores indicated riskier performance.

The standard BART was also administered (Lejuez et al., 2002). Participants are tasked with blowing up 30 balloons, one at a time. They earn five cents per pump but will lose that money if the balloon pops on a given trial. To save the money on any balloon, participants must stop pumping before the balloon pops and collect the money. The balloon is set to pop on a

random schedule, which cannot be learned by the participant. Performance was calculated based on the average pumps per balloon adjusted for only unexploded balloons (higher scores indicated riskier performance).

On the CCT (Figner et al., 2009), participants earn points by turning over a series of 32 cards. They are given information on each trial about the number of loss cards (1 or 3), the amount to be won on each card (10 or 30 points), and the amount to be lost if a loss card is chosen (250 or 750 points). Participants indicated the total number of cards to be turned over, and did not receive feedback on their selections until the end of the task (24 trials). The average number of cards selected was used as the outcome variable, with higher scores indicating riskier performance.

Finally, the GDT (Brand et al., 2005) was created to take into account some of the learning-based limitations of the IGT. Participants are given 18 trials in which to maximize profit by predicting the roll of a die, with information about relative risks/benefits of each decision outlined in the task instructions. They can predict a 1-number, 2-number, 3-number, or 4-number sequence, with the 1- and 2-number sequences constituting a riskier decision ($1,000 and $500 bets, respectively) than the 3- and 4-number sequences ($200 and $100 ,respectively). Although participants could see the monetary values of their bets, they were not directly told which bets were riskier/safer. For the present study, the percent disadvantageous (1,2) selections was calculated, with higher numbers indicating greater risk-taking.

Data Analysis

Correlations were calculated between performance on the four tasks at Time 1, as well as between performance on each task at Time 1 and corresponding Time 2 performance. In addition, paired-samples *t*-tests compared mean scores across time for each task.

Results

Table 1 includes the means, standard deviations, and correlations for the Time 1 variables. The only significant correlation ($p < .001$) to emerge among Time 1 variables was between the IGT ambiguity (Trials 1–40) and risk (Trials 41–100) trials. Table 2 provides the means and standard deviations for the Time 2 variables, as well as Time 1–Time 2 correlations for each task. On the IGT, Time 1 decision making under ambiguity (Trials 1–40) was not correlated with Time 2 performance, but Time 1 decision making under risk (Trials 41–100) was weakly correlated with all Time 2 IGT trials. BART performance at Time 1 was moderately to strongly correlated with BART performance at Time 2, and moderate correlations were also seen on the CCT and GDT.

Table 1. Correlations between Time 1 variables

Variable	*M* (*SD*)	1	2	3	4	5
1. IGT 1	−2.93 (10.64)	—				
2. IGT 2	4.26 (20.89)	0.51*	—			
3. BART	26.24 (13.73)	0.04	0.02	—		
4. CCT	11.82 (4.47)	0.04	0.01	0.12	—	
5. GDT	39.51 (27.40)	0.09	0.03	−0.12	0.05	—

Note: IGT = Iowa Gambling Task, Trials 1–40 (1) and Trials 41–100 (2); BART = Balloon Analogue Risk Task, average adjusted pumps; CCT = Columbia Card Task, average cards selected; GDT = Game of Dice Task, percent disadvantageous selections.
*$p \leq .001$.

Table 2. Correlations between Time 1 and Time 2 performance

Variable	*M* (*SD*) Time 2	IGT 1-1	IGT 1-2	BART 1	CCT 1	GDT 1
IGT 2-1	4.59 (13.98)	0.15	0.26**	0.15	0.06	−0.07
IGT 2-2	12.82 (25.60)	0.06	0.27**	0.05	0.10	−0.13
BART 2	30.24 (14.86)	0.05	0.14	0.69***	0.07	−0.08
CCT 2	12.92 (5.53)	0.14	0.24*	0.18	0.57***	0.12
GDT 2	29.48 (26.04)	0.07	−0.07	−0.24*	0.18	0.49***

Note: IGT = Iowa Gambling Task, Time 1 (1-) or Time 2 (2-), Trials 1–40 (1) and Trials 41–100 (2); BART = Balloon Analogue Risk Task, average adjusted pumps at Time 1 (1) or Time 2 (2); CCT = Columbia Card Task, average cards selected at Time 1 (1) or Time 2 (2); GDT = Game of Dice Task, percent disadvantageous selections at Time 1 (1) or Time 2 (2).
*$p \leq .05$; **$p \leq .01$; ***$p \leq .001$.

Paired-samples *t*-tests showed significant differences across time on all tasks except the CCT. On the IGT, both sets of trials (ambiguity: $t(93) = -4.50$, $p < .001$, Cohen's $d = 0.47$; risk: $t(93) = -2.98$, $p = .004$, $d = 0.31$) indicated performance was less risky at Time 2 than Time 1. Participants were also less risky on the GDT at Time 2 than Time 1, $t(93) = 3.82$, $p < .001$, $d = 0.37$. However, participants were riskier on the BART at Time 2 compared to Time 1, $t(95) = -3.69$, $p < .001$, $d = 0.36$. No differences were found on the CCT, $t(79) = -1.85$, $p = .068$, $d = 0.24$.

Discussion

The present study examined test–retest reliability of four common behavioral decision making tasks: IGT, BART, CCT, and GDT. Previous research suggested correlations of varying strengths across time on the IGT (Burdick et al., 2014; Cardoso et al., 2010; Lejuez et al., 2003; Tuvblad et al., 2013; Waters-Wood et al., 2012; Xiao et al., 2013; Xu et al., 2013) and BART (Collado et al., 2014; Weafer et al., 2013; White et al., 2008; Xu et al., 2013), with no examination, to our knowledge, of test–retest reliability on the CCT or GDT. We found minimal correlations between tasks at Time 1, providing further support that each task likely measures a separate component of decision making (Buelow & Blaine, 2015); however, additional factor analyses in larger samples are needed to confirm this finding. Performance on each task at Time 1 was significantly correlated with the corresponding task at Time 2, with correlations ranging from weak (IGT) to moderately strong (CCT, BART, GDT). Our Time 1–Time 2 IGT correlations were somewhat smaller than in previous research (Xu et al., 2013), but the BART finding was consistent with previous research (Weafer et al., 2013; White et al., 2008; Xu et al., 2013). We have also provided initial evidence of moderate test–retest reliability on the CCT and GDT.

Paired-samples *t*-tests indicated changes in performance on the IGT, BART, and GDT across time. Specifically, performance on the IGT and GDT was more advantageous at Time 2 compared to Time 1, which could be indicative of practice effects on these measures. The early, decision making under ambiguity trials of the IGT at Time 1 were not correlated with any IGT measurement at Time 2, consistent with previous research indicating it assesses a different component of decision making than the remaining trials (Brand et al., 2007). Once the participants learned information about each deck during these initial trials, this information was not "unlearned" and likely affected decisions on both the remaining Time 1 trials and all of the Time 2 trials. Using different terminology, more emotion-based decision making in the early trials shifted to more calculated, deliberate decision making in the later trials (and subsequent performance three weeks later; Brand et al., 2007). This evidence of practice effects reflects a continued concern with the IGT and its utility for repeated administration over time in clinical and non-clinical settings. This same concern likely exists for the GDT, which showed a similar pattern of strong time-based correlations and less risky performance at Time 2; however, additional research is needed to further examine and replicate our current findings with this task.

Performance on the BART was more disadvantageous at Time 2 than Time 1. It is uncertain why across-time performance differed on the BART versus IGT and GDT. The monetary amounts at risk on the BART are significantly lower than the IGT and GDT. In addition, the IGT and GDT risk the entire bank on each trial, whereas on the BART, banked money is never again at risk. Finally, the balloon could pop on any trial, whereas on the IGT and GDT participants can learn to avoid the riskiest decks/bets. These factors may have contributed to the increased riskiness on the BART at Time 2.

Although the CCT showed moderate correlations between Time 1 and Time 2 performance, no significant changes were seen across time on the paired-samples *t*-tests. It is possible the CCT might be less sensitive to practice effects and, with additional research, may show better test–retest reliability for repeated clinical- and research-based administrations. Additional research into the utility of the CCT to detect decision making impairments in patient populations is warranted, as to date most research has focused primarily on the IGT and to a lesser degree the BART.

The present study was conducted on a sample of healthy college student volunteers. Although consistent with much of the research to date utilizing these tasks, it is important to replicate these findings in both patient and non-patient samples across a variety of ages. In addition, participants were not screened for psychiatric or neurological conditions that could have affected performance on these tasks. As this was the first examination of test–retest reliability and potential practice effects on the CCT and GDT, it will be important to replicate these findings in a larger sample without any history of neurological or psychiatric conditions.

The present study replicated previous research examining test–retest reliability and practice effects on the IGT and BART while expanding this examination to the CCT and GDT. In general, evidence points towards lower reliability and greater practice effects on the IGT across a three-week time interval, with fewer practice effects and greater reliability on the CCT. Thus, with additional examination of performance stability in patient and non-patient samples, utilization of the CCT in future research studies may provide a more stable assessment of decision making processes.

Conflict of Interest

None declared.

References

Abbate-Daga, G., Buzzichelli, S., Marzola, E., Amianto, F., & Fassino, S. (2012). Effectiveness of cognitive remediation therapy (CRT) in anorexia nervosa: A case series. *Journal of Clinical and Experimental Neuropsychology*, *34*, 1009–1015. doi:10.1080/13803395.2012.704900.

Bechara, A. (2007). *Iowa gambling task professional manual*. Lutz, FL: Psychological Assessment Resources.

Brand, M., Fujiwara, E., Borsutzky, S., Kalbe, E., Kessler, J., & Markowitsch, H. J. (2005). Decision-making deficits of Korsakoff patients in a new gambling task with explicit rules: Associations with executive functions. *Neuropsychology*, *19*, 267–277. doi:10.1037/0894-4105.19.3.267.

Brand, M., Recknor, E. C., Grabenhorst, F., & Bechara, A. (2007). Decisions under ambiguity and decisions under risk: Correlations with executive functions and comparisons of two different gambling tasks with implicit and explicit rules. *Journal of Clinical and Experimental Neuropsychology*, *29*, 86–99. doi:10.1080/13803390500507196.

Buelow, M. T., & Blaine, A. L. (2015). The assessment of risky decision making: A factor analysis of performance on the Iowa gambling task, balloon analogue risk task, and Columbia card task. *Psychological Assessment*, *27*, 777–785. doi:10.1037/a0038622.

Buelow, M. T., & Suhr, J. A. (2009). Construct validity of the Iowa gambling task. *Neuropsychology Review*, *19*, 102–114. doi:10.1007/s11065-009-9083-4.

Burdick, K. E., Braga, R. J., Gopin, C. B., & Malhotra, A. K. (2014). Dopaminergic influences on emotional decision making in euthymic bipolar patients. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *39*, 274–282. doi:10.1038/npp.2013.177.

Cardoso, C. O., Carvalho, J. C. N., Cotrena, C., Bakos, D. G. S., Kristensen, C. H., & Fonseca, R. P. (2010). Reliability study of the neuropsychological test Iowa gambling task. *Journal Brasileiro de Psiquiatria*, *59*, 279–285.

Collado, A., Felton, J. W., MacPherson, L., & Lejuez, C. W. (2014). Longitudinal trajectories of sensation seeking, risk taking propensity, and impulsivity across early to middle adolescence. *Addictive Behaviors*, *39*, 1580–1588. doi:10.1016/j.addbeh.2014.01.024.

De Wilde, B., Bechara, A., Sabbe, B., Hulstijn, W., & Dom, G. (2013). Risky decision-making but not delay discounting improves during early inpatient treatment of polysubstance dependent alcoholics. *Frontiers in Addictive Disorders and Behavioral Dyscontrol*, *4*, 91. doi:10.3389/fpsyt.2013.00091.

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia card task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 709–730. doi:10.1037/a0014983.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, D. R., et al (2002). Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (BART). *Journal of Experimental Psychology: Applied*, *8*, 75–84. doi:10.1037//1076-898X.8.2.75.

Lejuez, C. W., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, W., et al (2003). The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology*, *11*, 26–33. doi:10.1037/1064-1297.11.1.26.

Tuvblad, C., Gao, Y., Wang, P., Raine, A., Botwick, T., & Baker, L. A. (2013). The genetic and environmental etiology of decision-making: A longitudinal twin study. *Journal of Adolescence*, *36*, 245–255. doi:10.1016/j.adolescence.2012.10.006.

Waters-Wood, S. M., Xiao, L., Denburg, N. L., Hernandez, M., & Bechara, A. (2012). Failure to learn from repeated mistakes: Persistent decision-making impairment as measured by the Iowa gambling task in patients with ventromedial prefrontal cortex lesions. *Journal of the International Neuropsychological Society*, *18*, 927–930. doi:10.1017/S135561771200063X.

Weafer, J., Baggott, M. J., & de Wit, H. (2013). Test-retest reliability of behavioral measures of impulsive choice, impulsive action, and inattention. *Experimental and Clinical Psychopharmacology*, *21*, 475–481. doi:10.1037/a0033659.

White, T. L., Lejuez, C. W., & de Wit, H. (2008). Test-retest characteristics of the balloon analogue risk task (BART). *Experimental and Clinical Psychopharmacology*, *16*, 565–570. doi:10.1037/a0014083.

Xiao, L., Wood, S. M. W., Denburg, N. L., Moreno, G. L., Hernandez, M., & Bechara, A. (2013). Is there a recovery of decision-making function after frontal lobe damage? A study using alternative versions of the Iowa gambling task. *Journal of Clinical and Experimental Neuropsychology*, *35*, 518–529. doi:10.1080/13803395.2013.789484.

Xu, S., Korczykowski, M., Zhu, S., & Rao, H. (2013). Risk-taking and impulsive behaviors: A comparative assessment of three tasks. *Social Behavior and Personality*, *41*, 477–486. doi:10.2224/sbp.2013.41.3.477.