Article

The Performance of College Students on the Iowa Gambling Task: Differences Between Scoring Approaches

Assessment
2022, Vol. 29(6) 1190–1203

DOI: 10.1177/10731911211004741
journals.sagepub.com/home/asm
SAGE

Wesley R. Barnhart[1*] and Melissa T. Buelow[2*]

Abstract

The Iowa Gambling Task (IGT) is one of the most common behavioral decision-making tasks used in clinical and research settings. Less-than-expected performance among healthy adults generates concerns about the validity of this task, and it is possible the particular scoring approach utilized could impact interpretation. We examined how performance patterns changed across several scoring approaches, utilizing a large, college student sample, both with (n = 406) and without (n = 1,547) a self-reported history of psychiatric or other diagnosis. Higher net scores were seen when participants selected decks with a low loss frequency than decks with high long-term outcomes; however, participants overall underperformed the IGT normative data sample. Receiver operating characteristic curves examining multiple scoring approaches revealed no threshold of impaired performance that both maximized sensitivity and minimized false positive rate on the IGT. Scoring approach matters in the determination of impaired decision making via the IGT in adults.

Keywords

Iowa Gambling Task, IGT, decision making, assessment

The Iowa Gambling Task (IGT) is one of the most commonly used behavioral risky decision-making tasks across research and clinical settings. First published in 1994 (Bechara et al., 1994), multiple IGT iterations are currently used in research, with each version varying on factors such as the magnitude of immediate reward, number of decks, and progressive nature of successive losses. The IGT was standardized for clinical practice with the publication of the Psychological Assessment Resources (PAR) version (Bechara, 2007). With this came an attempt to standardize scoring, as the manual presents normative data to aid interpretation. However, concerns remain about what constitutes "normal" performance on the IGT if healthy control and undergraduate student participants are falling below this criterion across different samples. These scoring-related concerns are consistent with continued discussion of the evidence for, or against, the construct validity and reliability of the IGT as a behavioral measure of risky decision making.

Description of the IGT

On the IGT, participants select 100 cards, one at a time, from one of four decks: A, B, C, or D. On each trial, participants win some money but might also lose some money (Bechara et al., 1994; Bechara, 2007), resulting in immediate net gains or net losses (Figure 1). The object of the game is to win, or avoid losing, as much money as possible. Unknown to participants at the task start, but learned through trial-and-error feedback, two decks have long-term positive outcomes (Decks C and D, advantageous decks) and two have long-term negative outcomes (Decks A and B, disadvantageous decks; Bechara et al., 1994). In particular, Decks A and B feature larger immediate rewards than Decks C and D. But these immediate gains are offset by larger losses such that continued selections from Decks A and B lead to long-term losses whereas continued selections from Decks C and D lead to long-term gains. By avoiding the disadvantageous decks in favor of the advantageous decks (both so termed due to their long-term outcomes), participants employ the best decision-making strategy and win money on the task.

Multiple studies examined the utility of the IGT to detect decision-making impairments among patient populations. Previous studies suggest several clinical populations for

[1]Bowling Green State University, Bowling Green, OH, USA
[2]The Ohio State University, Newark, OH, USA

*co-first authorship

Corresponding Author:
Wesley R. Barnhart, Department of Psychology, Bowling Green State University, 822 East Merry Avenue, Bowling Green, OH 43403, USA.
Email: wrbarnh@bgsu.edu

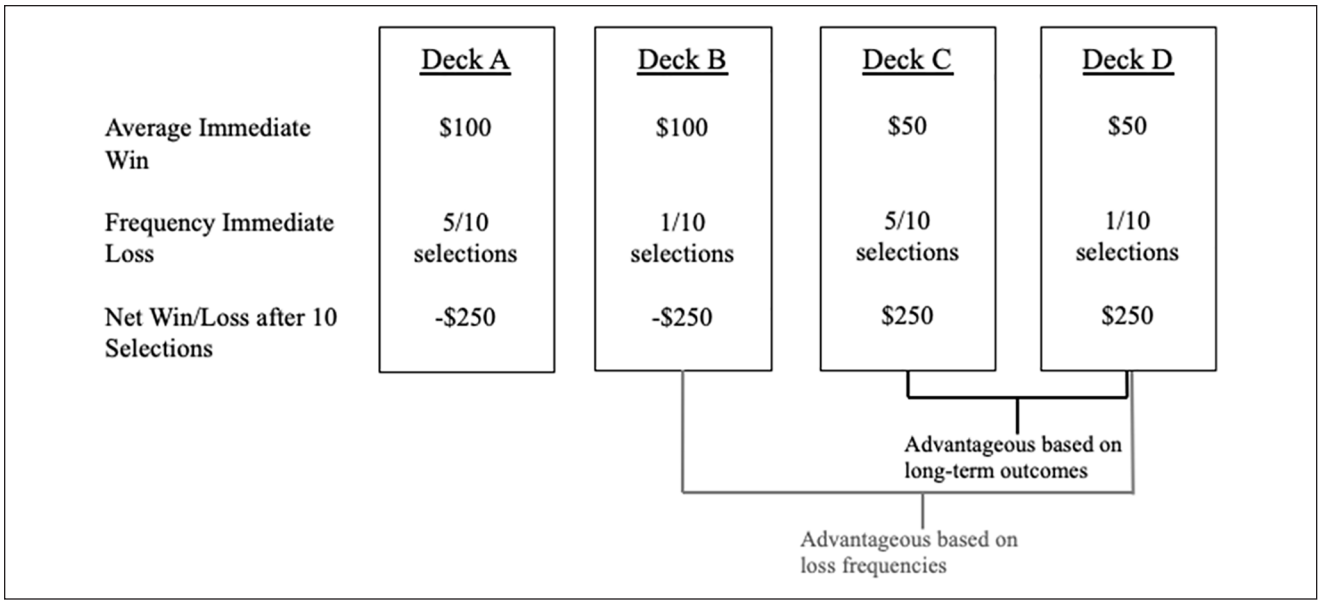

Figure 1. Differences between Decks on the IGT.
Note. IGT = Iowa Gambling Task.

which there are consistent IGT impairments, including traumatic brain injury (TBI) and lesions to the frontal lobe or amygdala (e.g., Buelow & Suhr, 2009). Other psychiatric diagnoses, including depression, anxiety, and Attention Deficit/Hyperactivity Disorder (ADHD), have a mixed relationship with the IGT, in that some studies suggest disorder-related impairments, whereas others do not find impairments (see Buelow, 2020, for a review). More specifically, relatively consistent findings of impaired IGT performance are seen in pathological gambling and the eating disorders, whereas inconsistent findings (e.g., some find worse IGT performance and others find no difference versus those without psychiatric diagnoses) as a function of generalized anxiety disorder, obsessive compulsive disorder, major depressive disorder, bipolar disorder, ADHD, and schizophrenia. Variations in IGT performance across substance use disorder studies appear linked, in part, to the particular substance of use, recency of last use, and length of history of use.

Scoring the IGT

Multiple scoring approaches exist for assessing IGT performance. Originally, the task was scored in several ways (Bechara et al., 1994): (1) number of selections from each deck, (2) number of advantageous (C + D) and number of disadvantageous (A + B) selections separately, and (3) subtracting disadvantageous (A + B) from advantageous (C + D) selections [net score] across all 100 trials. Subsequent studies examined long-term losses (disadvantageous decks A + B) compared with long-term gains (advantageous decks C + D). Utilizing this approach, healthy controls selected from the advantageous decks more than disadvantageous decks at a high rate (.60-.70 proportion of good deck selections; Bechara et al., 1994; Bechara et al., 1998; Bechara et al., 1999; Blair et al., 2001; North & O'Carroll, 2001; Overman et al., 2004; Premkumar et al., 2008; Shurman et al., 2005; Tomb et al., 2002; see Steingroever et al., 2013, for review).

Even in its infancy, there was no consistent manner to score IGT performance. The first attempt at creating a scoring guideline was that more than 50 selections from disadvantageous decks evidenced impaired performance (Bechara et al., 1998; Bechara et al., 1999). The next general guideline suggested total net scores ([C + D] − [A + B] across 100 trials) falling under 10 indicated impaired decision making (Bechara et al., 2001). Since then, researchers focused on two types of decision making assessed on the IGT. The early IGT trials are decision making under ambiguity, given little is known about the relative risks and benefits of each deck (Brand et al., 2007). The later 60 (Brand et al., 2007) or 40 (Ko et al., 2010; Noël et al., 2007) trials are instead considered decision making under risk, as decisions are made based on the relative risks and benefits of each deck. Thus, continued selections from "disadvantageous" decks over "advantageous" decks, even as these risks become apparent, constitute risky decision making on the task.

Despite multiple studies indicating decision-making changes as the task progresses, no new threshold value was created for impaired decision making. Researchers continued to report total net scores and performance by 20- or 25-card blocks of trials (e.g., Buelow & Barnhart, 2018a; Overman & Pierce, 2013). More recently, research focused on differences within the advantageous decks and the

disadvantageous decks (Figure 1). Decks C and D, although termed advantageous based on their long-term outcomes, differ in their frequency of losses (Deck C: losses on 5/10 selections; Deck D: losses on 1/10 selections). The same occurs for Decks A and B (Deck A: losses on 5/10 selections; Deck B: losses on 1/10 selections). It is possible that differences occur in the pattern of individual deck selections based on these factors (see below).

Normative Data for IGT Interpretation

The publication of the PAR version of the IGT allowed a preferred scoring approach to be presented to clinicians to aid interpretation. Researchers could also utilize the normative data and suggested scoring approach. Per the PAR manual, IGT scores can be used to "obtain information that supports a diagnosis" (Bechara, 2007, p. 7), and decision making is an executive function that can elucidate understanding of current frontal lobe function in a neuropsychological assessment, thus having a standard to interpret IGT scores is warranted. The PAR manual (Bechara, 2007) suggests scoring the IGT both as the total net score ([C+D] − [A+B]) and by 20-card block of trials. The normative sample included 932 adults, aggregated across multiple sites (Bechara, 2007). Adults with no psychological diagnosis were matched to the U.S. population in terms of age (18-95 years), gender (45.3% male, 54.7% female), and educational background (3-22 years). A separate U.S. Census-matched sample ($n = 264$) was included, based on a demographically-corrected sample (i.e., 49.2% males, 50.8% females; 18-90 years of age; 3-20 years of education; Bechara, 2007). Normative data are presented for total scores and by 20-card blocks of trials, with these guidelines: (1) T-scores (T) > 44 indicate good/nonimpaired decision making, (2) $T < 40$ indicates impaired decision making, and (3) $40 \leq T \leq 44$ indicate below average (or less than expected) decision making (Bechara, 2007). Only a percentile range is provided for individual deck selections. The PAR manual data (Bechara, 2007) indicate that 12.1% to 23.1% of the healthy control sample evidenced a net score falling in the impaired range (14.3% failure rate predicted by a normal distribution), with an additional 10.7% to 15.5% falling in the below average range (14.4% failure rate predicted by the normal distribution).

It is unclear whether intact decision making is consistently labelled as such across the PAR and non-PAR scoring approaches, in part because the normative data focus on one scoring approach. Learning that Deck B is a bad deck—or that Deck C is a good deck—takes more time to learn than learning that Deck D is a good deck (Buelow et al., 2013; Steingroever et al., 2013). It is possible that different scoring approaches, one labelling Decks C and D as advantageous and one labelling Decks B and D as advantageous, could lead to opposite classifications of performance. Moreover, these two decision-making strategies may lead to difficulties interpreting research on trait-level characteristics such as personality or state-level characteristics such as current mood.

A related concern is that "typical" performance on the IGT changed over time. In early studies, healthy controls selected a high number of advantageous cards (e.g., .60-.70). More recently, their averages decreased, more often falling in the .40 to .55 range (see Steingroever et al., 2013, for review). One possible reason for this difference comes from studies assessing individuals deck selections. Participants may minimize the frequency of losses, preferring Decks B and D, rather than maximize long-term outcomes, preferring Decks C and D (Ahn et al., 2008; Buelow & Brunell, 2020; Buelow et al., 2013; Caroselli et al., 2006; Chiu et al., 2008; Chiu & Lin, 2007; Dunn et al., 2006; Lin et al., 2007; Steingroever et al., 2013; Yechiam & Busemeyer 2005). This pattern, the "prominent Deck B phenomenon," can negatively affect net scores among healthy control participants when the long-term outcome scoring approach ([C + D] − [A + B]) is utilized (e.g., Caroselli et al., 2006; Chiu et al., 2012; Lin et al., 2007; Toplak et al., 2005). This tendency can complicate the drive to create a new threshold of nonimpaired/impaired performance on the IGT, as this threshold will need to come down on the side of either long-term outcomes or frequency of losses as the critical indicator of decision making on the task.

Construct Validity and Reliability Concerns

Although we do not explicitly examine construct validity in the present study, it is important to provide an overview of continuing concerns about the construct validity of the IGT as such concerns are directly related to scoring and interpretation. First, what type of decision making does the IGT assess (e.g., Buelow & Suhr, 2009; Dunn et al., 2006)? Early IGT iterations broadly categorized it as measuring behavioral decision making and frontal lobe functioning (e.g., Bechara et al., 1994). Since then, the task became associated with behavioral assessment of risky decision making and both "hot" and "cold" strategies, due to the presence/absence of affective responses before selections (e.g., somatic marker hypothesis; Bechara, 2004; Dunn et al., 2006). In general, hot, affective decision making focuses on quick, gut-based reactions to a situation, whereas cold, deliberative decision making is slower, void of emotions, and ultimately optimizes gains (Damasio, 1994; Seguin et al., 2007). Beyond this conceptualization of the IGT, the task also assesses decision making under ambiguity and decision making under risk (e.g., Brand et al., 2007). These differences in decision-making strategies and subsequent performance raise issues with operationalization of decision making on the IGT. Interestingly, there is more consistent evidence that the IGT

detects decision-making impairment in some neurological, substance use, and eating disorders than in other psychiatric disorders (e.g., major depressive disorder; Buelow, 2020). This more consistent evidence in neurological and substance use disorder samples, for example, may explain why the PAR IGT manual provides validity information for these specific populations. Still concerns remain regarding the IGT's reliability and whether it can track changes in decision making across time.

Growing research shows that practice effects, changes in task performance due to a previous administration of that task, commonly occur on the IGT. More specifically, practice effects on the IGT may manifest through improved performance from Time 1 to Time 2 over a short period of time (e.g., 1-3 weeks; Buelow & Barnhart, 2018b; Ernst, et al., 2003; Xu et al., 2013), as well as over months (e.g., Cardoso et al., 2010; Verdejo-García et al., 2007) to years (e.g., Tuvblad et al., 2013) between administrations. Although these patterns can indicate test–retest reliability across short periods of time, they also call into question the IGT's ability to track decision-making changes over time as a clinical marker of disease severity or treatment responsiveness. Two other forms of reliability, parallel forms and split-half, were also previously examined. The E-F-G-H version maintains similar deck distinctions but with larger gains and losses; however, inconsistent findings are seen when both the standard and parallel forms are administered (e.g., Verdejo-García et al., 2009). Mixed findings are also seen for split-half reliability (Gansler et al., 2011; Monterosso et al., 2001), but this is consistent with the wealth of research suggesting decision-making changes as the task progresses. Tasks that require participants to learn from feedback, such as the IGT, would necessarily lower split-half reliability estimates as the second half of the task should improve compared with the first half. The PAR IGT manual does not discuss task reliability, despite having a chapter devoted to validity (Bechara, 2007). Taken together, the IGT assesses behavioral decision making, a highly complex and multifaceted construct, and indeed several interpretations (i.e., scoring approaches) of decision-making task performance on the IGT exist which, in tandem with reliability concerns, obscure detection and interpretation of impaired versus nonimpaired decision making on this task.

The Present Study

The present study uses a large college student sample to further develop our understanding of differences across the long-term outcome-based and loss frequency-based scoring approaches on the IGT. Several concerns with the current IGT scoring approach are examined. It may be the case that, depending on the specific scoring approach used, lowered/impaired decision making may or may not be detected in a clinical assessment setting. On the other hand, it may be the case that decision-making difficulties are wrongfully concluded, a finding consistent with emerging research (e.g., Steingroever et al., 2013) and likely due, at least in part, to a decision-making strategy to minimize frequency of losses to the detriment of maximizing long-term outcomes. This inconsistency in identifying decision-making difficulties introduces threats to the authenticity of decision-making profiles created with the IGT. While utilizing the IGT is just one facet of a comprehensive assessment battery, unclear interpretations of results can nonetheless paint an inaccurate profile of functioning and introduce challenges to interpretation of clinical data (see Dunn et al., 2006). This discrepancy is particularly notable in cases where the IGT is used as the sole index of decision making, as the IGT is the only current task standardized for clinical evaluations. In research settings, these inconsistencies introduce challenges to interpreting how situational factors affect performance, as it would be difficult to determine if impaired performance is due to experimental manipulation or a "normative" decision-making strategy to minimize losses rather than focus on long-term outcomes.

Several study aims are presented. First, we examined performance on the IGT following two scoring metrics: long-term outcome-based ($[C + D] - [A + B]$) and loss frequency-based scoring ($[B + D] - [A + C]$). Although we do not make a specific hypothesis regarding proportions of participants falling in each IGT T-score range, recent research suggests healthy controls are not performing to the same level as evidenced in earlier studies. Thus, we expect our rates of participants falling in the lower T-score ranges ($T < 40$, $40 \leq T \leq 44$) will exceed the rates of participants falling in these ranges in the IGT normative data (e.g., 12% to 23% for $T < 40$, 10% to 16% for $40 \leq T \leq 44$), when the long-term outcome scoring based PAR T-scores are examined.

We also predict differences in impairment based on the specific scoring approach utilized. As no normative data is provided in the IGT manual for the frequency-based scoring, we rely on previous guidelines (e.g., total net score over 10; Bechara et al., 2001). We hypothesize that more participants will score in the intact range when the frequency-based scoring, rather than the long-term outcome scoring, is applied. Some participants self-reported a previous history of ADHD, psychiatric diagnosis, or TBI, factors that can affect performance on the IGT. We predict this subgroup will have a higher rate of impaired performance on the IGT, compared with those without these diagnoses, across multiple scoring approaches: (1) raw scores following the long-term outcome and frequency-based scoring approaches, (2) T-scores based on the PAR normative data, and (3) net scores falling above or below 10 using the frequency scoring.

We next turn to the sensitivity and specificity of T-scores and other approaches to classify performance on the IGT,

utilizing receiver operating characteristic (ROC) analyses. No specific hypotheses were made for these analyses. We compared a subgroup of individuals with a self-reported history of ADHD, psychiatric diagnosis, or TBI to a subgroup without this self-reported history. We examined how sensitivity (true positive rate) and specificity (true negative rate) changed across scoring approaches. In particular, we examined: (1) T-scores for total net scores and scores by 20-card blocks of trials; (2) total net scores falling above 10 (long-term outcome scoring); (3) total net scores falling above 10 (frequency-based scoring); (4) total net scores, across all 100 trials or just the final 60 trials, falling at a different criterion than 10 that increases sensitivity and specificity; (5) Deck D selections, across all 100 trials or just the final 60 trials; and (6) Deck A selections, across all 100 trials or just the final 60 trials. These final two options are examined as the research literature consistently holds that Deck D is an advantageous deck and Deck A is a disadvantageous deck. We examine whether focusing on one of these decks improves the sensitivity and specificity of task classification.

Method

Participants and Procedure

The present study utilized a sample of convenience. Data were compiled across studies occurring in the laboratory from 2011 to 2018 and included control participants. Although much of the data were previously published, the present study represents different analyses and aims. All procedures were approved by the university's institutional review board and participants provided informed consent. They completed the IGT as part of a larger study. Demographic data was reported prior to debriefing and participants received course credit (not a monetary payment) for participation. Data were compiled for 1,953 undergraduate students at an open-access regional campus of a large Midwestern university in the United States (M_{age} = 18.96 [*SD* = 2.68; *range* = 18-51]; 41.8% male; 69.1% European American or White, 13.3% African American or Black).

Participants could self-report a previous diagnosis of ADHD, psychiatric disorder, or history of TBI. Based on this self-reported data, we identified 126 participants with a diagnosis of ADHD, 163 with a psychiatric diagnosis other than ADHD, 52 with a history of TBI, and 65 with a history of more than one diagnosis. Two groups were created. The first group included participants reporting no prior diagnostic history (*n* = 1,547, M_{age} = 18.88 [*SD* = 2.58; *range* = 18-51]; 41.7% male; 66.1% European American or White, 15.6% African American or Black). The second group comprised those participants reporting some diagnostic history (*n* = 406; M_{age} = 19.22 [*SD* = 2.97; *range* = 18-41]; 42.2% male; 79.8% European American or White).

Measures

Iowa Gambling Task. The computerized IGT, as previously described, was utilized (Bechara, 2007). The two most frequently utilized scoring approaches are the long-term outcome scoring and frequency-based scoring approaches. The long-term outcome scoring approach, per the PAR IGT manual (Bechara, 2007), focuses on long-term outcomes to indicate advantageous decision making. The number of disadvantageous selections (Decks A + B) are subtracted from the number of advantageous selections (Decks C + D), either across the entire 100 trials or by 20-card blocks of trials. Per the frequency-based scoring approach, the number of high loss frequency selections (Decks A + C) are subtracted from the number of low loss frequency selections (Decks B + D). For both, positive scores indicate more advantageous decisions and negative scores indicate more disadvantageous decisions.

Data Analysis

To control for the Type I error rate, we interpret results at the $p \leq .001$ level. Given our large sample size, we also pay particular attention to effect size estimates for each analysis. The IGT manual was used to determine T-scores for total net scores and by 20-card blocks of trials. For participants missing any demographic information, the census-matched normative data (IGT manual Appendix B; Bechara, 2007) were used instead of the age- and education-corrected normative data (IGT manual Appendix A; Bechara, 2007).

Several analyses were conducted. First, we conducted a repeated-measures analysis of variance (ANOVA) to examine learning across blocks and scoring approaches. Specifically, we compared scores on the 20-card blocks following the long-term outcome scoring to the scores following the frequency scoring approach. We next compared participants to the PAR IGT normative data set. We conducted chi-square analyses on T-scores from the normative data, focusing on participants without a self-reported diagnostic history. The proportion of participants scoring in the impaired ($T < 40$), less than expected ($40 \leq T \leq 44$), and unimpaired ($T > 44$) ranges were compared with the equivalent proportion of participants reported in Table 4.10 in the PAR IGT manual (Bechara, 2007). To assess differences in proportions of participants with a total net score below 10 across the two scoring approaches, a *z* test for two proportions was conducted.

Next, we focus on differences between our participants self-reporting a previous history of ADHD, TBI, and/or other psychiatric diagnosis and those without this self-reported history. A mixed ANOVA compared raw scores following both scoring approaches across the two groups. Diagnostic group was the between-subjects variable and Scoring approach and Block were the within-subjects

Table 1. Means and Standard Deviations for IGT Scoring Approaches.

Block	Standard Scoring ([C + D] − [A + B])	Frequency Scoring ([B + D] − [A + C])
No previous diagnosis		
Block 1	−2.82 (5.76)	1.98 (5.25)
Block 2	0.19 (6.54)	3.24 (6.08)
Block 3	0.87 (7.85)	4.23 (7.10)
Block 4	1.38 (8.79)	4.72 (7.75)
Block 5	1.54 (9.08)	4.73 (8.41)
Total net score	1.12 (26.25)	18.90 (23.50)
Previous diagnosis		
Block 1	−3.03 (5.63)	1.96 (5.46)
Block 2	0.52 (6.39)	3.21 (5.90)
Block 3	1.76 (8.05)	3.62 (7.19)
Block 4	2.41 (8.50)	4.63 (7.81)
Block 5	1.30 (9.59)	4.72 (8.72)
Total net score	2.78 (25.56)	18.12 (23.81)

Note. Standard deviations are presented in parentheses. IGT = Iowa Gambling Task.

variables. To compare T-scores, a mixed ANOVA was conducted on Block and Diagnostic group. Chi-square analyses were used to compare the number in each diagnostic group who fell above/below the +10 net score criterion. Finally, we calculated ROC curves to assess differences in the sensitivity and specificity of various indicators of impaired decision making. More specifically, we assessed the ability of PAR IGT T-scores, raw scores based on the long-term outcome scoring approach, raw scores based on the frequency scoring approach, and Deck A and Deck D selections to maximize sensitivity and specificity. The area under the ROC curve (AUC) was examined, with values above .500 indicating greater ability to distinguish between groups. We also report the Youden Index, with scores closer to 1 indicating a better ability to distinguish between groups (e.g., Shan, 2015). We examine different threshold values for each analysis to determine a value that maximized sensitivity and specificity.

Results

Table 1 contains the means and standard deviations for each scoring approach, by 20-card blocks of trials and by total net scores (Trials 1-100). First, we assessed differences in decision making by scoring approach with repeated measures ANOVA, focusing on participants without a self-reported diagnosis. The main effect of Block was significant and represented a medium to large effect, $F(3.54, 5473.39) = 201.56$, $p < .001$, $\eta_p^2 = .12$. Independent of scoring approach, participants improved decisions across Blocks 1 to 4 (ps $< .001$), with no difference between Blocks 4 and 5, $p = .536$. The main effect of Scoring Approach was also significant and represented a large effect, $F(1, 1546) = 369.04$, $p < .001$, $\eta_p^2 = .19$, with higher total scores with the frequency-based versus long-term outcome approach, $p < .001$. Finally, the interaction effect was also significant but represented a small effect, $F(3.39, 5242.08) = 9.68$, $p < .001$, $\eta_p^2 = .006$, which implies steeper learning rates with the frequency scoring approach compared with the long-term outcome scoring approach (see Figure 2). Scores were higher in the frequency-based than long-term outcome-based scoring approaches in each of the five blocks of trials, ps $< .001$. Examining just the raw score calculations, participants learned to decide more advantageously across blocks but scored higher when the frequency-based scoring was applied.

We next compared participant scores, using just the long-term outcome scoring approach, to the normative data presented in the PAR manual. Figures for the normative data classifications, individual deck selections, and ROC analyses can be found in supplemental materials (Supplemental Figures 1–4 [available online]). The chi-square analyses included just those participants without a self-reported previous diagnosis. Looking at scores falling in the impaired range ($T < 40$), a lower percentage of current participants fell in this range in Block 1 but higher percentages in Blocks 3 and 4 compared with the PAR normative sample (all ps $<$ .001). For scores falling in the less than expected range ($40 \leq T \leq 44$), a higher percentage of current participants were in this range in Block 2, 4, and 5 compared with the PAR normative sample (all ps $< .001$). Finally, the present participants "passed" ($T > 44$) the IGT at a higher rate than the normative sample in Block 1, but fared worse than the normative sample in Blocks 4, 5, and overall (all ps $< .001$). Taken together, our sample, self-reporting no diagnosis of ADHD, other psychiatric condition, or TBI, performed on average worse than would be expected based on the PAR IGT scoring approach and normative data set.

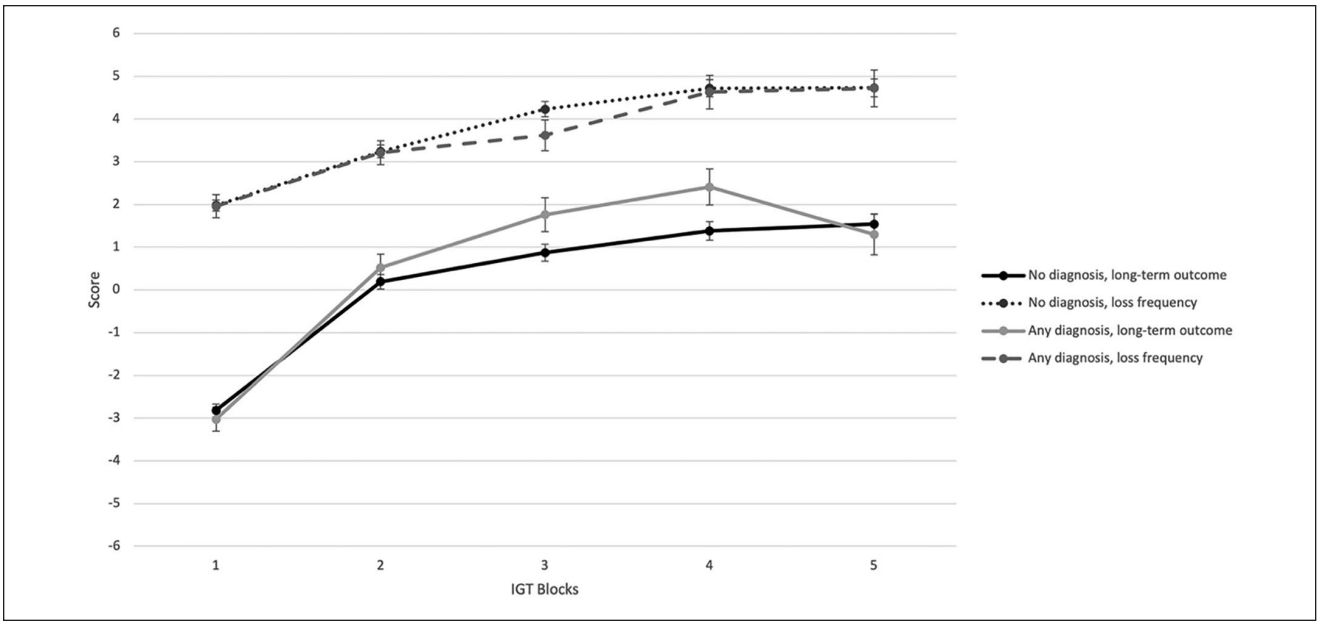

Figure 2. IGT Scores by Block and Scoring Approach.
Note. IGT = Iowa Gambling Task.

To better understand differences across the two scoring approaches, we examined differences in the proportion of participants falling below the total net score (Trials 1-100) of 10 across each scoring approach. When Decks C and D are advantageous (e.g., long-term outcome scoring), 1,085 participants (70.1%) fall below a net score of 10 and are considered impaired. When Decks B and D are advantageous (e.g., frequency scoring), 607 participants (39.2%) fall below the net score of 10, a significantly smaller proportion than are failing with the long-term outcome scoring approach, $z = 17.26$, $p < .00001$. This finding may be due to a greater number of selections from Deck B than Deck C throughout all IGT blocks (Supplemental Figure 2 [available online]).

Next, we focus on differences between our two groups of participants, those self-reporting a previous history of ADHD, TBI, and/or other psychiatric diagnosis and those without this self-reported history. First, raw scores were examined with a mixed ANOVA. There was a significant, moderate main effect of Block, $F(3.51, 6845.79) = 182.554$, $p < .001$, $\eta_p^2 = .09$, such that independent of scoring approach or diagnostic group, performance improved across blocks, ps $< .001$ (no differences between Blocks 3 and 5, $p = .008$, and Blocks 4 and 5, $p = .177$). There was a significant, moderate to large, main effect of Scoring Approach, $F(1, 1951) = 262.90$, $p < .001$, $\eta_p^2 = .12$. Total scores were higher with the frequency scoring versus the long-term outcome scoring, $p < .001$. No significant main effect of Group emerged, $F(1, 1951) = 0.31$, $p = .579$, $\eta_p^2 = .00$. In addition, the Block × Group, $F(3.51, 6845.79) = 1.18$, $p = .317$, $\eta_p^2 = .001$, and Group × Scoring Approach, $F(1, 1951) = 1.59$, $p = .208$, $\eta_p^2 = .001$, interactions were not significant, providing further evidence of a lack of differences in IGT performance between those with and without self-reported previous diagnoses. There was, however, a small, significant Block × Scoring Approach interaction, $F(3.32, 6479.17) = 13.46$, $p < .001$, $\eta_p^2 = .01$. Across blocks, scores were higher with frequency versus long-term outcome scoring, ps $< .001$. There was no significant three-way interaction, $F(3.32, 6479.17) = 2.33$, $p = .066$, $\eta_p^2 = .001$.

As limited between-group differences emerged in IGT performance based on raw scores, additional analyses examined whether the PAR IGT T-scores (long-term outcome scoring) and the net total score of 10 (both scoring approaches) identified between-group differences. For the T-scores, there was a significant, moderate, main effect of Block, $F(3.38, 6597.35) = 120.31$, $p < .001$, $\eta_p^2 = .06$, as T-scores *decreased* across blocks, ps $< .001$ (except Block 3 to Block 4, $p = .461$). The main effect of Diagnostic Group was not significant, $F(1, 1951) = 0.86$, $p = .355$, $\eta_p^2 = .00$, nor was the Block × Group interaction, $F(3.38, 6597.35) = 2.53$, $p = .048$, $\eta_p^2 = .001$. T-scores did not reliably distinguish between control and diagnostic groups. When the +10 net score criterion was examined, no group differences emerged in the proportion of participants in each group scoring above 10 for the long-term outcome scoring, $\chi^2(1, N = 1{,}953) = 0.31$, $p = .580$, nor for the frequency-based scoring, $\chi^2(1, N = 1{,}953) = 1.11$, $p = .291$. In sum, none of the major scoring approaches (frequency, long-term outcome) or calculations (raw scores, T-scores, comparisons to net +10) lead to differences in

performance across those without a self-reported diagnosis and those with a self-reported diagnosis.

Finally, ROC curves were examined to assess whether one scoring approach maximized sensitivity and specificity across the diagnostic and nondiagnostic groups. As nearly all tested ROC curves showed poor sensitivity, AUC close to .500, and low Youden Index scores, we present two representative figures in the online supplemental materials. Regardless of whether total score or one of the 20-card block scores were utilized, cut-off T-scores of 40 and 44, as suggested by the PAR manual, resulted in low sensitivity and moderate specificity, with AUC falling in the .472 to .516 range. Additional ROC curve analyses to improve sensitivity and specificity revealed no clear point of improvement: as sensitivity increased, specificity decreased, and the Youden Index was near 0 for all scores. In Table 2, we present results of ROC analyses for a representative sample of alternative threshold values.

Our final attempt to find a performance indicator maximizing sensitivity and specificity shifts to an examination of the raw scores. In particular, we created ROC curves for raw scores based on both scoring approaches, then assessed the ability of the +10 criterion to differentiate between groups. Table 2 shows the true and false positive rates for thresholds of −2, 0, and +20, as a comparison. For raw scores, neither the long-term outcome nor frequency scoring approaches led to a better means of distinguishing between nondiagnostic and diagnostic groups. Any attempt to choose a threshold value that maximized sensitivity had the effect of increasing the false positive rate as well, consistent with AUC values falling near the .500 mark and Youden Index scores near 0. Finally, examining the number of Deck D or Deck A selections individually again resulted in AUC values near .500 and Youden Index scores near 0.

Discussion

The present study sought to examine how different scoring approaches classify college student participants on the IGT. Recent research highlighted higher than expected rates of disadvantageous decision making on the IGT in healthy adults (Steingroever et al., 2013), furthering the discussion that decision-making impairment on the IGT does not match real-world behavior (Dunn et al., 2006). These studies highlight a need for research to examine how the IGT norms fare in real-world samples. To this end, the present study explored two common scoring approaches, the long-term outcome scoring (utilized in the PAR T-score approach) and the loss frequency scoring, in addition to several alternative scoring metrics focused on Decks A and D.

We first examined the overall pattern of decision making across blocks following the long-term outcome and frequency scoring approaches. Following both scoring approaches, participants learned to select more advantageously over time (large effect size), selecting more from either Decks C and D (long-term outcome scoring) or Decks B and D (frequency scoring). However, total scores were greater in the frequency than in the long-term outcome scoring (large effect size), likely driven by more selections from Deck B than Deck C across blocks. These results are consistent with results suggesting that participants are increasingly using a decision-making strategy to minimize the frequency of losses, rather than to maximize the long-term outcomes (Steingroever et al., 2013). The commonality of the prominent Deck B phenomenon across studies has called into question whether Deck C is an advantageous deck, since participants lose money on 50% of the trials (e.g., Chiu et al., 2012; Chiu & Lin, 2007; Lin et al., 2007). The PAR manual (Bechara, 2007) does provide some support for this phenomenon, indicating that selections from this deck are less indicative of impaired decision making than Deck A selections, since many controls choose from Deck B. Yet no normative data are provided for the individual deck selections beyond percentile ranges across the 100 trials. The high level of variability seen in how participants learn to decide advantageously on the IGT, either by focusing on maximizing long-term outcomes or by focusing on minimizing the frequency of immediate losses, makes it concerning that the current normative data comes down on the side of just one of these scoring approaches (i.e., maximizing long-term outcomes).

However, selecting from Decks B and D—minimizing immediate loss frequency—may in fact lower the amount of money earned, which is counter to the task's instruction to maximize profit across the 100 trials. As an exploratory analysis, we assessed for correlations between total money earned and the following metrics: (1) total score, long-term outcome scoring; (2) total score, loss frequency scoring; and (3) total selections from the individual decks. There was a strong positive correlation between long-term outcome total score and money earned, $r = .791$, $p < .001$, but no relationship between loss frequency scoring and money earned, $r = -.022$, $p = .339$. Looking at the individual deck selections, all correlations with total money earned were significant at $p < .001$. More money was earned with greater Deck C ($r = .319$) and Deck D ($r = .568$) selections but fewer Deck A ($r = -.441$) and Deck B ($r = -.634$) selections. Even if participants believe that minimizing the frequency of losses is a better decision-making strategy, this does not lead to greater money earned (i.e., greater long-term outcomes) on the task.

We next compared performance of the present participants with those included in the PAR normative data set. We predicted the rates of participants falling in the lower T-score ranges would be greater than in the normative sample. Although the present participants fared better than the normative sample during Block 1, our participants ultimately failed the IGT at higher rates than the normative sample in Blocks 3 and 4. Furthermore, less than expected decision making ($40 \leq T \leq 44$) was higher among the

Table 2. Sensitivity, Specificity, and Youden Index Results Across ROC Analyses.

Indicator	AUC	SNS	SPC	Youden	SNS	SPC	Youden	SNS	SPC	Youden	SNS	SPC	Youden
			T = 36			T = 40			T = 44			T = 50	
Block 1 T	.511	.044	.950	−.006	.121	.884	.005	.209	.811	.020	.448	.538	−.014
Block 2 T	.480	.057	.944	.001	.140	.833	−.027	.397	.565	−.038	.729	.240	−.031
Block 3 T	.472	.150	.833	−.017	.227	.732	−.041	.382	.580	−.038	.712	.231	−.057
Block 4 T	.473	.116	.863	−.021	.227	.723	−.050	.515	.477	−.008	.700	.280	−.020
Block 5 T	.516	.244	.803	.047	.330	.521	−.149	.544	.477	.021	.714	.293	.007
Total T	.486	.138	.863	.001	.244	.740	−.016	.453	.535	−.012	.727	.255	−.018
			−2 Net			0 Net			+10 Net			+20 Net	
LT Score, Raw Total	.483	.436	.543	−.021	.525	.446	−.029	.687	.299	−.014	.773	.211	−.016
Freq. Score, Raw Total	.512	.187	.822	.009	.227	.785	.012	.421	.608	.029	.591	.434	.025
LT Score, Raw Last 60	.484	.379	.615	−.006	.488	.491	−.021	.675	.305	−.02	.773	.196	−.031
Freq. Score, Raw Last 60	.509	.148	.865	.013	.239	.761	.000	.488	.529	.017	.687	.341	.028
			10 Selections			20 Selections			30 Selections			40 Selections	
Total Deck A	.479	.795	.147	−.058	.320	.690	.010	.012	.977	−.011	.000	.993	−.007
Total Deck D	.476	.046	.956	.002	.263	.700	−.037	.641	.336	−.023	.849	.165	.014

Note. ROC = receiver operating characteristic; AUC = area under the curve; LT score = long-term outcome based scoring; Freq. score = frequency-based scoring; SNS = sensitivity; SPC = specificity.

present participants for Blocks 2, 4, and 5 compared with the IGT normative data. Thus, our sample of healthy participants—those that did not self-report diagnoses of ADHD, other psychiatric conditions, or TBI—performed significantly worse than performance metrics outlined by the PAR manual. These data align with recent research (Steingroever et al., 2013) and call attention to how decisions about intact versus impaired decision-making task performance are made. It should be noted that PAR T-scores were created based on IGT performance in a large sample of healthy adults. That our sample of healthy adults is falling into a lower than expected rate of performance compared with this normative sample is concerning but may again lend credence to the theory that individuals are increasingly viewing minimizing the frequency of losses as a preferred decision-making strategy despite long-term consequences for the total money earned and normative comparisons.

As an additional exploratory analysis, we assessed whether decision making on the IGT changed as a function of time, given the concerns raised in Steingroever et al. (2013). Correlations were assessed between the academic semester of data collection (time) and performance following both scoring approaches. No significant correlations emerged between time and long-term outcome scores ($ps > .005$), nor between time and frequency-based scores ($ps > .116$). Thus, based on this data set, the IGT does appear to be measuring the same underlying decision-making construct across time, but this discrepancy between early and more recent healthy control sample performances across studies remains.

An added concern related to the high rate of failure of the IGT in the present participants is whether they reflect a low score on a single test or instead reflect decision-making impairments that should present across tests. Across the five blocks, 11.7% to 29.2% of our student participants scored in the impaired range using the PAR normative data (long-term outcome scoring). This high rate of impaired performance is consistent with the 20% to 65% classified as impaired in Steingroever et al.'s (2013) review of the literature. Impairment rate increased as the task progressed, indicating more participants "failed" the IGT during the decision making under risk than the decision making under ambiguity trials. As we show with ROC analyses, no clear threshold for intact decision making emerged across the tested metrics. AUC values fell close to .500, increasing sensitivity also increased the false positive rate, and no scoring approach was superior in maximizing sensitivity and specificity to differentiate intact from impaired performance.

Since the long-term outcome scoring approach is preferred by the PAR IGT, this approach guides clinicians' interpretation of test performance. Previous research, though not specific to decision making, suggests obtaining a single low score across a larger neuropsychological test battery is common (e.g., Binder et al., 2009), with 42.4% to 60.8% of participants across studies obtaining a score at or below the 16th percentile (Brooks et al., 2008; Brooks et al., 2011; Iverson et al., 2008). However, this singular pattern of occasional low performance is more complicated to reconcile in the behavioral decision-making literature in which no other clinically normed measures of the construct exist outside of the IGT. However, at least some of the concern regarding low performance on one measure may relate to task reliability. With the IGT, concerns about practice effects, low test–retest reliability (e.g., Buelow & Barnhart, 2018b), and mixed parallel and split-half reliability estimates (e.g., Gansler et al., 2011; Verdejo-García et al., 2009) may complicate our understanding of whether low performance indicates impairment. Comparing performance with other clinically available decision-making tasks would allow for further investigation of possible impairments. Indeed, while other behavioral measures of decision making exist and may tap into unique facets of the construct, thus providing additional information about decisions (Buelow & Blaine, 2015), it remains unclear what to base these scores on which, in turn, makes interpretation of task performance unclear.

Conclusions drawn from the present study are informed by a lack of consensus in the field regarding the IGT's construct validity. This lack of consensus calls into question the extent to which the task can provide information about the individual's true decision-making profile and lead to questions for future exploration. If there are inconsistencies in IGT performance across individuals with similar psychiatric diagnoses, should the existing normative data be applied to all psychiatric diagnoses? If the IGT assesses an individual-differences variable, leading to significant variability within a diagnostic category, can norms truly be established? At a minimum, continued concerns about construct validity lead researchers and clinicians to question how to accurately quantify impaired and intact performance on the task. It is our hope that findings from the present work serve as a springboard for future research to investigate these and other concerns, with the goal of improving accurate detection of behavioral decision-making difficulties via the IGT. We also hope these concerns will generate the development of other clinically-approved decision-making tasks to complement or supplement the IGT.

Implications

The present study has several implications for clinicians and researchers alike. Clinicians may find difficulty interpreting the IGT in light of differences in performance across scoring approaches. Indeed, our results raise concerns about the high proportion of participants who fell in the lower ranges of T-scores. In addition, ROC analyses revealed that, irrespective of scoring approach, the IGT appears to have poor sensitivity and specificity in terms of distinguishing

intact versus impaired performance across participants with and without self-reported diagnoses of ADHD, other psychiatric disorders, or TBI. Thus, clinicians who utilize the IGT as a marker of disease state or treatment responsiveness should couple scores on this task with other behavioral decision-making measures (see Buelow & Blaine, 2015). However, to date the IGT is the only standardized task with an extensive normative data set available to clinicians. Other tasks, such as the Balloon Analogue Risk Task (Lejuez et al., 2002), Columbia Card Task (Figner et al., 2009), and Game of Dice Task (Brand et al., 2005) are frequently tested in conjunction with the IGT in the research literature but are not normed for use in clinical settings. Although coupling the IGT with these tasks in research may provide additional information about the type and extent of decision-making difficulties, these tasks are not clinically normed and should not be used for diagnostic purposes in clinical practice. Until additional tasks are standardized and normed to assess behavioral decision making in evaluations of psychiatric and neurological disorders, clinicians should utilize reliable and valid assessments of other cognitive functions to further understand potential decision-making impairments. For example, early research frequently paired the IGT with the Wisconsin Card Sort Task (Strauss et al., 2006), to demonstrate that IGT impairments were not due to other executive dysfunction (see Buelow & Suhr, 2009, for review). Tests assessing basic attention and working memory, cognitive flexibility, planning, and overall cognitive ability show correlations with the IGT (see Buelow, 2020, for review), and could be used to confirm impairments demonstrated on this task.

Unfortunately, no one scoring metric emerged as both sensitive and specific to decision-making impairments on the IGT, likely due to the existence of two "good" decision-making strategies on the task (i.e., minimize losses, maximize long-term gains). Conducting additional analyses on individual deck selections could allow for a better understanding of whether participants preferred the pathological deck (Deck A) or the less-than-optimal decision-making strategy deck (Deck B), in turn leading to different interpretations of these data. Another implication from the present study is that a newer, more fine-grained scoring approach is needed to better tease apart below average and impaired decision making. If the long-term outcome scoring approach overestimates impairment by classifying frequency-based selections as disadvantageous, but the frequency scoring approach underestimates impairment by combining Deck D (almost universally accepted as the best deck) with Deck B, a "middle ground" needs to be found that combines both approaches to more fully understand decision making. Researchers may need to consider the results of cognitive modeling studies, which point toward learning from feedback, trial-by-trial evaluations of outcomes, and sensitivity to gains and losses (e.g., Ahn et al., 2008; Byrne & Worthy, 2016; Dai et al., 2015) to create a more effective means of interpretation. However, parameter estimates would need to be evaluated against comparable standards to conventional task scores when used for diagnostic purposes.

Limitations and Future Research

The present study is not without limitations. Because our sample relied on college student participants, we only had access to self-reported, yes/no history of ADHD, other psychiatric diagnosis, or TBI. Due to the question wording, we were unable to further differentiate between nor confirm these self-reported diagnoses. As previously discussed, the previous research literature shows inconsistent relationships between psychiatric diagnoses and performance on the IGT, and we have no data on current use of substances that can affect task performance. While we found no relationship between these diagnoses and the patterns of impaired/intact decision making on the IGT, future research should explore different IGT scoring approaches across diverse clinical populations to determine whether higher rates of impairment are seen among those with a diagnosis. Future research may elucidate a subset of neurological and psychiatric diagnoses with consistent impairments on the IGT and for which clinical norms more accurately differentiate between impaired and nonimpaired performance.

Furthermore, our sample of undergraduate students may have higher baseline cognitive abilities than the general population. We do not have data on baseline cognitive functioning across our sample of participants but encourage future research to examine how overall cognitive functioning relates to performance across decision-making tasks. Future research should also assess effort in relation to behavioral decision-making task performance given it may impact task performance (see Westbrook & Braver, 2015). These data may aid in identifying if the IGT is sensitive to malingering or suboptimal effort, providing additional context to scores of use to researchers and clinicians alike. Future research should also consider other measures of behavioral decision making with both overlapping and unique decision-making components (see Buelow & Blaine, 2015) to better model decision-making profiles in healthy and clinical samples. Finally, implications of the present study are constrained to the IGT and thus future research adopting a similar scoring approach across a more complete decision making and executive functioning battery would offer more complete conclusions. Unfortunately, inconsistencies in the additional tasks administered across our study protocols meant that we did not have access to a large sample of co-administered assessments of other decision making and executive functioning measures. Future research can help assess the overall validity of the IGT, in the context of other tasks, to identify decision-making impairments across different psychiatric and neurological diagnoses.

Conclusions

The results of the present study indicate concerns with the clinical normative data provided in the PAR IGT manual, such that college student participants reporting no previous history of diagnosis known to affect performance on the IGT are in fact scoring lower than would be expected on the task. It is possible this finding may be due to a reliance on the alternative decision-making strategy of minimizing frequency of immediate losses, a strategy that is not reflected in the manual-provided T-score norms. However, neither scoring approach, nor alternatives based on Deck A or Deck D selections, resulted in a ROC curve with adequate sensitivity and specificity. No threshold for intact versus impaired decision making could be found, in direct contrast with the $T < 40$ guideline provided for clinicians. Indeed, the most optimal strategy in the 1990s (e.g., focusing on long-term outcomes), when the IGT was first created, may not be the most optimal strategy now due, in part, to changes in our environment that emphasize minimizing the frequency of short-term losses as an adaptive decision-making strategy. In the IGT, we have a task that can assess both minimization of short-term loss frequency and maximization of long-term gains. Our approach to scoring and understanding these data should match this potential.

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) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Wesley R. Barnhart https://orcid.org/0000-0002-9809-5225

Supplemental Material

Supplemental material for this article is available online.

References

Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, *32*(8), 1376-1402. https://doi.org/10.1080/03640210802352992

Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and Cognition*, *55*(1), 30-40. https://doi.org/10.1016/j.bandc.2003.04.001

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

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1-3), 7-15. https://doi.org/10.1016/0010-0277(94)90018-3

Bechara, A., Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience*, *19*(13), 5473-5481. https://doi.org/10.1523/JNEUROSCI.19-13-05473.1999

Bechara, A., Damasio, H., Tranel, D., & Anderson, S. W. (1998). Dissociation of working memory from decision making within the human prefrontal cortex. *Journal of Neuroscience*, *18*(1), 428-437. https://doi.org/10.1523/JNEUROSCI.18-01-00428.1998

Bechara, A., Dolan, S., Denburg, N., Hindes, A., Anderson, S. W., & Nathan, P. E. (2001). Decision-making deficits, linked to a dysfunctional ventromedial prefrontal cortex, revealed in alcohol and stimulant abusers. *Neuropsychologia*, *39*(4), 376-389. https://doi.org/10.1016/s0028-3932

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, *24*(1), 31-46. http://doi.org/10.1093/arclin/acn001

Blair, R. J. R., Colledge, E., & Mitchell, D. G. V. (2001). Somatic markers and response reversal: Is there orbitofrontal cortex dysfunction in boys with psychopathic tendencies? *Journal of Abnormal Child Psychology*, *29*(6), 499-511. http://doi.org/10.1023/A:1012277125119

Brand, M., Kalbe, E., Labudda, K., Fujiwara, E., Kessler, J., & Markowitsch, H. J. (2005). Decision-making impairments in patients with pathological gambling. *Psychiatry Research*, *133*(1), 91-99. http://doi.org/10.1016/j.psychres.2004.10.003

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*(1), 86-99. https://doi.org/10.1080/13803390500507196

Brooks, B. L., Holdnack, J. A., & Iverson, G. L. (2011). Advanced clinical interpretation of the WAIS-IV and WMS-IV: Prevalence of low scores varies by level of intelligence and years of education. *Assessment*, *18*(2), 156-167. http://doi.org/10.1177/1073191110385316

Brooks, B. L., Iverson, G. L., Holdnack, J. A., & Feldman, H. H. (2008). Potential for misclassification of mild cognitive impairment: A study of memory scores on the Wechsler Memory Scale-III in healthy older adults. *Journal of the International Neuropsychological Society*, *14*(3), 463-478. https://doi.org/10.1017/S1355617708080521

Buelow, M. T. (2020). *Risky decision making in psychological disorders*. Elsevier.

Buelow, M. T., & Barnhart, W. R. (2018a). An initial examination of performance on two versions of the Iowa gambling task. *Archives of Clinical Neuropsychology*, *33*(4), 502-507. https://doi.org/10.1093/arclin/acx103

Buelow, M. T., & Barnhart, W. R. (2018b). Test–retest reliability of common behavioral decision making tasks. *Archives of Clinical Neuropsychology*, *33*(1), 125-129. http://doi.org/10.1093/arclin/acx038

Buelow, M. T., & Blaine, M. T. (2015). The assessment of risk decision making: A factor analysis of performance on the Iowa gambling task, balloon analogue risk task, and Columbia card task. *Psychological Assessment*, *27*(3), 777-785. https://doi.org/10.1037/a0038622

Buelow, M. T., & Brunell, A. B. (2020). Narcissism, the experience of pain, and risky decision making. *Frontiers in Psychology*, *11*, Article 1128. https://doi.org/10.3389/fpsyg.2020.01128

Buelow, M. T., Okdie, B. M., & Blaine, A. L. (2013). Seeing the forest through the trees: Improving decision making on the Iowa Gambling Task by shifting focus from short- to long-term outcomes. *Frontiers in Psychology*, *4*, Article 773. https://doi.org/10.3389/fpsyg.2013.00773

Buelow, M. T., & Suhr, J. A. (2009). Construct validity of the Iowa Gambling Task. *Neuropsychology Review*, *19*(1), 102-114. https://doi.org/10.1007/s11065-009-9083-4

Byrne, K. A., & Worthy, D. A. (2016). Toward a mechanistic account of gender differences in reward-based decision-making. *Journal of Neuroscience, Psychology, and Economics*, *9*(3-4), 157-168. https://doi.org/10.1037/npe0000059

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*(4), 279-285. https://doi.org/10.1590/S0047-20852010000400003

Caroselli, J. S., Hiscock, M., Scheibel, R. S., & Ingram, F. (2006). The simulated gambling paradigm applied to young adults: An examination of university students' performance. *Applied Neuropsychology*, *13*(4), 203-212. https://doi.org/10.1207/s15324826an1304_1

Chiu, Y.-C., & Lin, C.-H. (2007). Is deck C an advantageous deck in the Iowa gambling task? *Behavioral and Brain Functions*, *3*, Article 37. https://doi.org/10.1186/1744-9081-3-37

Chiu, Y.-C., Lin, C.-H., & Huang, J.-T. (2012). Prominent deck B phenomenon: Are decision-makers sensitive to long-term outcome in the Iowa gambling task? In A. E. Cavanna (Ed.), *Psychology research progress: Psychology of gambling: New research* (pp. 93-118). Nova Science.

Chiu, Y.-C., Lin, C.-H., Huang, J.-T., Lin, S., Lee, P.-L., & Hsieh, J.-C. (2008). Immediate gain is long-term loss: Are there foresighted decision makers in the Iowa gambling task? *Behavioral and Brain Functions*, *4*, Article 13. https://doi.org/10.1186/1744-9081-4-13

Dai, J., Kerestes, R., Upton, D. J., Busemeyer, J. R., & Stout, J. C. (2015). An improved cognitive model of the Iowa and Soochow gambling tasks with regard to model fitting performance and tests of parameter consistency. *Frontiers in Psychology*, *6*, Article 229. https://doi.org/10.3389/fpsyg.2015.00229

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. Putnam.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience & Biobehavioral Reviews*, *30*(2), 239-271. https://doi.org/10.1016/j.neubiorev.2005.07.001

Ernst, M., Kimes, A. S., London, E. D., Matochik, J. A., Eldreth, D., & Tata, S. (2003). Neural substrates of decision making in adults with attention deficit hyperactivity disorder. *American Journal of Psychiatry*, *160*(6), 1061-1070. https://doi.org/10.1176/appi.ajp.160.6.1061

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*(3), 709-730. http://dx.doi.org/10.1037/a0014983

Gansler, D. A., Jerram, M. W., Vannorsdall, T. D., & Schretlen, D. J. (2011). Does the Iowa Gambling Task measure executive function? *Archives of Clinical Neuropsychology*, *26*(8), 706-717. http://doi.org/10.1093/arclin/acr082

Iverson, G. L., Brooks, B. L., & Holdnack, J. A. (2008). Misdiagnosis of cognitive impairment in forensic neuropsychology. In R. L. Heilbronner (Ed.), *Neuropsychology in the courtroom: Expert analysis of reports and testimony* (pp. 243-266). Guilford Press.

Ko, C. H., Hsiao, S., Liu, G., Yen, J., Yang, M., & Yen, C. (2010). The characteristics of decision making, potential to take risks, and personality of college students with internet addiction. *Psychiatry Research*, *175*(1-2), 121-125. https://doi.org/10.1016/j.psychres.2008.10.004

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75-84. https://doi.org/10.1037/1076-898X.8.2.75

Lin, C.-H., Chiu, Y.-C, Lee, P.-L., & Hsieh, J.-C. (2007). Is deck B a disadvantageous deck in the Iowa gambling task? *Behavioral and Brain functions*, *3*, Article 16. https://doi.org/10.1186/1744-9081-3-16

Monterosso, J., Ehrman, R., Napier, K. L., O'Brien, C. P., & Childress, A. R. (2001). Three decision-making tasks in cocaine-dependent patients: Do they measure the same construct? *Addiction*, *96*(12), 1825-1837. https://doi.org/10.1046/j.1360-0443.2001.9612182512.x

Noël, X., Bechara, A., Dan, B., Hanak, C., & Verbanck, P. (2007). Response inhibition deficit is involved in poor decision making under risk in nonamnesic individuals with alcoholism. *Neuropsychology*, *21*(6), 778-786. https://doi.org/10.1037/0894-4105.21.6.778

North, N. T., & O'Carroll, R. E. (2001). Decision making in patients with spinal cord damage: Afferent feedback and the somatic marker hypothesis. *Neuropsychologia*, *39*(5), 521-524. https://doi.org/10.1016/s0028-3932(00)00107-x

Overman, W. H., Frassrand, K., Ansel, S., Trawalter, S., Bies, B., & Redmond, A. (2004). Performance on the Iowa card task by adolescents and adults. *Neuropsychologia*, *42*(13), 1838-1851. https://doi.org/10.1016/j.neuropsychologia.2004.03.014

Overman, W. H., & Pierce, A. (2013). Iowa Gambling Task with non-clinical participants: Effects of using real + virtual cards and additional trials. *Frontiers in Psychology*, *4*, 935. https://doi.org/10.3389/fpsyg.2013.00935

Premkumar, P., Fannon, D., Kuipers, E., Simmons, A., Frangou, S., & Kumari, V. (2008). Emotional decision-making and its dissociable components in schizophrenia and schizoaffective disorder: A behavioural and MRI investigation. *Neuropsychologia*, *46*(7), 2002-2012. https://doi.org/10.1016/j.neuropsychologia.2008.01.022

Seguin, J. R., Arseneault, L., & Tremblay, R. E. (2007). The contribution of "cool" and "hot" components of decision-making in adolescence: Implications for developmental psychopathology. *Cognitive Development*, *22*(4), 530-543. https://doi.org/10.1016/j.cogdev.2007.08.006

Shan, G. (2015). Improved confidence intervals for the Youden Index. *PLOS ONE*, *10*(7), e0127272. https://doi.org/10.1371/journal.pone.0127272

Shurman, B., Horan, W. P., & Nuechterlein, K. H. (2005). Schizophrenia patients demonstrate a distinctive pattern of decision-making impairment on the Iowa gambling task. *Schizophrenia Research*, *72*(2-3), 215-224. https://doi.org/10.1016/j.schres.2004.03.020

Steingroever, H., Wetzels, R., Horstmann, A., Neumann, J., & Wagenmakers, E.-J. (2013). Performance of healthy participants on the Iowa gambling task. *Psychological Assessment*, *25*(1), 180-193. https://doi.org/10.1037/a0029929

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford University Press.

Tomb, I., Hauser, M., Deldin, P., & Caramazza, A. (2002). Do somatic markers mediate decisions on the gambling task? *Nature Neuroscience*, *5*(11), 1103-1104. https://doi.org/10.1038/nn1102-1103

Toplak, M. E., Jain, U., & Tannock, R. (2005). Executive and motivational processes in adolescents with Attention-Deficit-Hyperactivity Disorder (ADHD). *Behavioral and Brain Functions*, *1*, Article 8. https://doi.org/10.1186/1744-9081-1-8

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*(2), 245-255. http://dx.doi.org/10.1016/j.adolescence.2012.10.006

Verdejo-García, A., Benbrook, A., Funderburk, F., David, P., Cadet, J.-L., & Bolla, K. I. (2007). The differential relationship between cocaine use and marijuana use on decision-making performance over repeat testing with the Iowa gambling task. *Drug and Alcohol Dependence*, *90*(1), 2-11. http://doi.org/10.1016/j.drugalcdep.2007.02.004

Verdejo-García, A., López-Torrecillas, F., Calandre, E. P., Delgado-Rodríguez, A., & Bechara, A. (2009). Executive function and decision-making in women with fibromyalgia. *Archives of Clinical Neuropsychology*, *24*(1), 113-122. https://doi.org/10.1093/arclin/acp014

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 395-415. http://doi.org/10.3758/s13415-015-0334-y

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*(3), 477-486. http://doi.org/10.2224/sbp.2013.41.3.477

Yechiam, E., & Busemeyer, J. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review*, *12*(3), 387-402. https://doi.org/10.3758/BF03193783