FULL-LENGTH REPORT

Journal of Behavioral Addictions 7(4), pp. 1122–1131 (2018)
DOI: 10.1556/2006.7.2018.117
First published online December 6, 2018

Obtaining quality data using behavioral measures of impulsivity in gambling research with Amazon's Mechanical Turk

MAGDALEN G. SCHLUTER*, HYOUN S. KIM and DAVID C. HODGINS

Department of Psychology, University of Calgary, Calgary, AB, Canada

(Received: February 20, 2018; revised manuscript received: July 23, 2018; second revised manuscript received: October 12, 2018; accepted: November 10, 2018)

Background and aims: To date, no research has examined the viability of using behavioral tasks typical of cognitive and neuropsychology within addiction populations through online recruitment methods. Therefore, we examined the reliability and validity of three behavioral tasks of impulsivity common in addiction research in a sample of individuals with a current or past history of problem gambling recruited online. *Methods:* Using a two-stage recruitment process, a final sample of 110 participants with a history of problem or disordered gambling were recruited through MTurk and completed self-report questionnaires of gambling involvement symptomology, a Delay Discounting Task (DDT), Balloon Analogue Risk Task (BART), Cued Go/No-Go Task, and the UPPS-P. *Results:* Participants demonstrated logically consistent responding on the DDT. The area under the empirical discounting curve (AUC) ranged from 0.02 to 0.88 ($M = 0.23$). The BART demonstrated good split-third reliability (ρs = 0.67 to 0.78). The tasks generally showed small correlations with each other (ρs = ±0.06 to 0.19) and with UPPS-P subscales (ρs = ±0.01 to 0.20). *Discussion and conclusions:* The behavioral tasks demonstrated good divergent validity. Correlation magnitudes between behavioral tasks and UPPS-P scales and mean scores on these measures were generally consistent with the existing literature. Behavioral tasks of impulsivity appear to have utility for use with problem and disordered gambling samples collected online, allowing researchers a cost efficient and rapid avenue for conducting behavioral research with gamblers. We conclude with best-practice recommendations for using behavioral tasks using crowdsourcing samples.

Keywords: behavioral tasks, gambling, impulsivity, crowdsourcing, MTurk

INTRODUCTION

Over the past several years, the use of Amazon's Mechanical Turk (MTurk; www.mturk.com) for experimental- and survey-based psychological research has surged, including recruitment of clinical samples (Chandler & Shapiro, 2016). MTurk is an online crowdsourcing platform that allows large groups of individuals (MTurkers) to complete tasks for small monetary payment. As the use of MTurk for psychological research has grown in popularity, questions have been raised regarding the reliability of data collected and the validity of inferences made from MTurk samples (Huff & Tingley, 2015; Paolacci & Chandler, 2014). Although the extant literature provides some confidence for the reliability and validity of self-report data on MTurk (Chandler & Shapiro, 2016; Kim & Hodgins, 2017; Mishra & Carleton, 2017), whether this extends to data that require sustained attention (e.g., neuropsychological tasks) remains unknown. In this study, we aimed to address questions regarding the quality of MTurk data using a clinical sample of individuals with current or past problem gambling behavior.

One reason for MTurk's rising popularity in psychological research is due to the advantages it provides as a recruitment tool. Better-quality data can be collected in less time than alternative methods of recruiting convenience samples (e.g., through Facebook; Shao et al., 2015), and workers can be kept completely anonymous, reducing the risk of experimenter influence on the results (Crump, McDonnell, & Gureckis, 2013). Samples recruited from MTurk also tend to be more demographically diverse than other convenience samples, such as undergraduate students (Chandler & Shapiro, 2016). Specifically, the mean age of MTurkers tends to be higher than undergraduate samples and lower than other adult convenience samples and the racial/ethnicity composition tends to be less non-Hispanic white than undergraduate samples (Berinsky, Huber, & Lenz, 2012). Workers also tend to be younger, less religious, more liberal, and have a higher mean education than the overall population and other non-probability samples (Chandler & Shapiro, 2016).

Furthermore, data collected through MTurk have demonstrated high convergent and concurrent validity (Chandler & Shapiro, 2016), scale reliability (Buhrmester, Kwang, & Gosling, 2011), test–retest reliability (Chandler & Shapiro,

* Corresponding author: Magdalen G. Schluter, MSc; Department of Psychology, University of Calgary, 2500 University Drive NW, Calgary T2N 1N4, AB, Canada; Phone: +1 403 210 9500; E-mail: Magdalen.schluter@ucalgary.ca

ISSN 2062-5871 © 2018 The Author(s)

2016; Kim & Hodgins, 2017), and comparable effect sizes to those seen in the existing literature on a variety of psychological measures (Shapiro, Chandler, & Mueller, 2013). In addition, populations that are traditionally difficult to recruit in large numbers or who are underrepresented in traditional recruitment techniques can be reached through MTurk (Blumberg & Luke, 2007; Chandler & Shapiro, 2016). Importantly for this paper, MTurk can be used to recruit populations with substance abuse and gambling addiction (Kim & Hodgins, 2017). However, issues of inattention and inaccurate reporting of symptoms may be particularly concerning among addiction populations, where issues of truthful responding are especially prominent (Godinho, Kushnir, & Cunningham, 2016).

Several studies have directly addressed issues of data quality among addiction populations. Kim and Hodgins (2017) examined the reliability and validity of self-report MTurk data collected from gambling, alcohol, and cannabis users. Overall, data from the gambling and alcohol samples showed good internal consistency, 1-week test–retest reliability, and external validity. Moreover, the reliability indices were like those found in the existing literature, with the exception of the cannabis sample, suggesting caution when collecting data from this population (Kim & Hodgins, 2017). In a related vein, Mishra and Carleton (2017) investigated the usage of crowdsourcing for gambling research. In a series of studies, they demonstrated good convergent validity and adequate test–retest reliability across several measures of problem gambling and general gambling involvement, associations with reports of personality, impulsivity, and behavioral risk-taking with magnitudes consistent with existing literature. Similar to the previous research suggesting higher rates of psychopathology in crowdsourcing samples (Arditte, Çek, Shaw, & Timpano, 2016), Mishra and Carleton (2017) found rates of problematic gambling higher than the general population. On the whole, these studies suggest that although crowdsourcing samples are not representative of the general population, MTurk may be used to efficiently recruit addiction samples and collect high-quality data.

This study extends this examination of the utility of MTurk by investigating its use in behavioral tasks typical of cognitive or neuropsychology within addiction samples. Research using behavioral tasks has consistently shown aberrant impulsive choice (MacKillop et al., 2011), response inhibition (De Wit, 2009), and risk-taking behaviors in addiction populations (De Wit, 2009; MacKillop et al., 2011). The ability to collect large and diverse samples quickly using these tasks would be of benefit to the field. Such tasks typically involve recording of response times, sustained attention, and often complex instructions, which may be problematic for web-based administration, especially among participants recruited through crowdsourcing. That said, several behavioral tasks have been validated with MTurk among non-clinical samples (Crump et al., 2013). However, whether these results translate to clinical samples, including problem gamblers, have yet to been tested.

Overview of the present research

In sum, MTurk is increasingly being used in psychological research with clinical samples (Chandler & Shapiro, 2016; Shapiro et al., 2013) and while issues of data quality have been addressed with regard to self-report measures (Kim & Hodgins, 2017; Mishra & Carleton, 2017), this has not been done with behavioral tasks common to addiction research. Therefore, the aim of the present research was to investigate the reliability and validity of three common behavioral tasks [Delay Discounting Task (DDT), Balloon Analogue Risk Task (BART), and Cued Go/No-Go Task] in individuals with a lifetime history of problem gambling recruited through MTurk. Logical consistency of responding on the DDT and split-third reliability of the BART were examined as indicators of task reliability. Given the multifaceted nature of impulsivity, the UPPS-P Impulsive Behavior Scale was used to provide a measure of convergent and divergent validity. A comparison of the behavioral tasks disordered gambling symptoms provided a measure of external validity. The results of these tasks (including means and standard deviations) and correlations with self-reported impulsivity were compared to previous research to further assess the utility of using behavioral tasks of impulsivity on MTurk with clinical samples.

METHODS

Participants

A total of 304 individuals residing in the United States (55% male; $M_{\text{age}} = 37.57$, $SD_{\text{age}} = 11.42$) participated in this study. Participation was limited to individuals aged 18 years and older who felt that they had experienced a problem with gambling behavior at some point in their life. Of those who gave permission to be recontacted for Part 2 (99%), 268 (89%) were eligible and 116 completed Part 2.

Procedure

The study was hosted through TurkPrime, a platform designed for social science research that integrates with the MTurk platform and recruits participants through MTurk. A recruitment notice was posted on Amazon's MTurk to recruit our sample of interest, "*For this study, we are recruiting people who may feel they currently have a problem with gambling activities or have had a problem with gambling in the past. Examples of gambling would include: card games (e.g., poker, blackjack), slot machines, sports betting (including horse racing), lotteries and other type of Casino games (e.g., roulette).*" Participants were further informed about the follow-up study. MTurkers were compensated approximately $0.10 per minute for their participation, a rate that has been suggested as an ethical remuneration for MTurk studies (Chandler & Shapiro, 2016). Consistent with previously employed methods, participation was limited to MTurkers with an approval rating of 95% (i.e., MTurkers who have been approved 95% of the time or greater for the tasks they have previously completed) and duplicate IP addresses were blocked to prevent multiple responding (Kim & Hodgins, 2017).

Participants were redirected to a survey hosted by Qualtrics where they provided informed consent. Given concerns that MTurk workers may misreport symptoms in order to be eligible for studies with explicit inclusion criteria, best research practices recommend a two-stage screening process, and use of subtle measures of reliability

(Chandler & Paolacci, 2017; Chandler & Shapiro, 2016). Therefore, to ensure that our sample consisted of our target population, Part 1 included three screening items to assess for lifetime problem gambling behavior [NORC DSM-IV Screen for Gambling Problems questions pertaining to loss of Control, Lying, and Preoccupation (NODS-CLiP); Toce-Gerstein, Gerstein, & Volberg, 2009]. Participants who met the eligibility criteria completed a battery of self-report measures and were then directed to a consent to be recontacted to complete Part 2. Participants who were interested and eligible were invited through TurkPrime to complete Part 2. Participants who demonstrated inconsistent responding, characterized by responses on the NODS-CLiP that differed from those provided on the identical questions presented later were not invited to complete Part 2.

In Part 2, participants were redirected to Inquisit, a study platform hosted by Millisecond (www.millisecond.com) that administers behavioral tasks online. As part of informed consent, they were notified that they would be asked to install the Inquisit launcher, and that they would be directed to a page that provided instructions on how to uninstall the program upon completion. Participants were also given information on how to exit the behavioral tasks if they would like to stop. Then, they completed three behavioral tasks and a questionnaire.

Measures

NODS-CLiP (Toce-Gerstein & Volberg, 2004). Participants first completed the NODS-CLiP, rapid screen for disordered and problem gambling in adults. The three questions, derived from the National Opinion Research Center Diagnostic Screen (NORC) DSM-IV Screen for Gambling Problems (NODS) pertain to loss of control, lying, and preoccupation diagnostic criteria in the DSM. A threshold of one item endorsed captures 96.2% of NODS problem and disordered gamblers (Toce-Gerstein et al., 2009). Consistent with previous research (e.g., Kim, Wohl, Salmon, & Santesso, 2017; Suurvali, Hodgins, Toneatto, & Cunningham, 2008), participants who endorsed at least one item went on to complete the remaining questionnaires.

Demographics and gambling involvement. Participants completed a questionnaire assessing demographic and gambling involvement in the previous 12 months. Those who had not gambled in the preceding 12 months did not complete the gambling involvement questions.

NORC DSM Screen for Gambling Problems (NODS; Gerstein et al., 1999). The NODS is a self-report screening measure of both current and lifetime problem gambling based on DSM-IV criteria. Seventeen questions for lifetime and 17 corresponding past-year items scored yes or no and measure 10 disordered gambling criteria. Consistent with DSM-5 criteria (American Psychiatric Association, 2013), the criterion of engaging in illegal activity was excluded from analysis.

Delay Discounting Task (DDT). According to a modified version of the Richards, Zhang, Mitchell, and de Wit (1999), DDT was used. In this task, participants are asked to make choices between smaller-immediate rewards and larger-later rewards. The task was adjusted to provide suitable variability in possible scores using hypothetical rewards. Consistent with other tasks using hypothetical rewards, the standard delayed reward was increased to $1,000, and seven temporal delays were presented: 1 day, 1 week, 1 month, 3 months, 1 year, 5 years, and 25 years (Koffarnus & Bickel, 2014). The magnitude of the rewards was systematically varied using a random-adjustment algorithm to estimate indifference points for each temporal delay, where the subjective value of the immediate and delayed reward is equal (Myerson, Green, & Warusawitharana, 2001). A random-adjustment task is preferred over procedures with a set order of presentation, because participants are not able to anticipate the next question. The area under the empirical discounting curve (AUC) was used as an indicator of impulsive choice (Myerson et al., 2001). Given significant debate regarding different discounting equations, it has been recommended that the AUC should be used, as it is a theoretically neutral parameter (Madden & Bickel, 2010). Smaller AUCs reflect steeper discounting and greater impulsive choice.

Balloon Analogue Risk Task (BART; Lejuez et al., 2002). As a measure of risk-taking, participants completed the BART. In this task, participants are presented an image of a deflated balloon, which they are asked to pump up virtually. For each successful pump, they can earn hypothetical money. They are informed that the balloon will explode at some point and that no money will be earned from exploded balloons. Participants are informed that the point at which the balloon pops is random, and pumps can range from 1 to 128 for each individual trial. The task involves 30 trials (i.e., 30 balloons). The average pump count for unexploded balloons is used as an indicator of risk-taking propensity (Lejuez et al., 2002).

Cued Go/No-Go Task (Fillmore, 2003). As a measure of motor response inhibition, participants completed the Cued Go/No-Go Task. Participants are provided green and blue rectangles presented vertically (125 trials) or horizontally (125 trials). They are told to press the space bar when they see a green rectangle but to refrain from pressing when they see a blue rectangle. The vertical rectangles have a higher probability of being green (80%), and the horizontal rectangles have a high probability of being blue (80%). Participants are given a cue about the orientation of the rectangle (a silhouette) shortly before the color of the rectangle is revealed. As such, response-appropriate expectancies influence behavioral control mechanisms (Vogel-Sprott & Fillmore, 2011). Inhibition errors were used as a measure of motor response inhibition, with higher errors indicating greater trouble with response inhibition. The number of inhibition errors can range from 0 to 200. Due to different response-appropriate expectancies on horizontal and vertical cues, inhibition errors were reported individually for each orientation.

UPPS-P Impulsive Behavior Scale (Lynam, Smith, Whiteside, & Cyders, 2006). The UPPS-P is a 59-item self-report measure of personality traits that lead to impulsive behaviors. Each item is measured on a 4-point Likert scale from 1 to 4. Total scores can range from 59 to 236, with higher scores indicating greater impulsivity. It assesses five impulsivity subscales; negative urgency measures the tendency to act rashly in response to strong negative emotions (12 items), (lack of) premeditation assesses the ability to think through possible consequences before acting

(11 items), (lack of) perseverance assesses the ability to persist in completing tasks (10 items), sensation seeking assesses the preference for excitement and stimulation (12 items), and finally positive urgency measures the tendency to act rashly in response to strong positive emotions (14 items).

Data analytic plan

All analyses were conducted in R (R Core Team, 2008) software. A variety of measures of reliability and validity were calculated. Part 1 completion time was examined as a subtle measure of validity. Outliers were examined using visual plots and Z scores of +3.29 (Cohen, Cohen, West, & Aiken, 2003). To address concerns with symptom misreporting in MTurk samples (Chandler & Paolacci, 2017), we compared NODS-CLiP responses to the identical question presented later in the lifetime NODS. As a subtle measure of reliability, the logical consistency of responding in the DDT was also examined. Split-third reliability (comparing performance on the first, second, and third blocks of balloons) was examined using Spearman's correlation coefficients. In the absence of an established measure of response consistency in the Go/No-Go Task, only the relationship of this task to other measures and the magnitude of the correlations with UPPS-P scored compared to the preexisting literature were examined.

Spearman's correlation coefficients were used to examine concurrent and divergent validity of the behavioral tasks of impulsivity. External validity was examined by calculating Spearman's correlations between the behavioral task measures and NODS scores for both current and lifetime modules. The magnitude of the associations to previous results in the literature was compared.

Ethics

The study procedures were carried out in accordance with Conjoint Faculties Research Ethics Board at the University of Calgary. The institutional review board of the University of Calgary approved the study. All subjects were informed about the study and all provided consent.

Table 1. Demographic information of participants who completed Part 1 and Part 2

	Part 1		Part 2	
	N	%	*N*	%
Gender				
Male	166	54.60	58	52.73
Female	133	43.80	50	45.45
Other	2	0.60	2	1.82
Missing	3	0.90	0	0.00
Income				
Under $10,000	12	3.90	5	4.55
$10,000–$39,000	99	32.50	41	37.27
$40,000–$69,000	96	31.60	31	28.18
$70,000–$99,000	97	31.90	24	21.82
Education				
High-school diploma or less	62	20.39	23	20.91
Trades or apprenticeship	12	3.95	6	5.45
College	42	13.82	15	13.64
University below bachelors	36	11.84	16	14.55
Bachelors	116	38.16	37	33.64
Graduate degree	36	11.84	13	11.82
Ethnicity				
Caucasian	224	80.30	87	79.09
South Asian	5	1.70	0	0.00
Black	34	11.20	11	10.00
Filipino/Pacific Islander	2	0.70	0	0.00
Latin American	23	7.60	14	12.73
Chinese	6	2.00	2	1.82
Southeast Asian	1	0.30	1	0.91
Arab	1	0.30	1	0.91
Japanese	1	0.30	0	0.00
Korean	2	0.70	1	0.91
Aboriginal	1	0.30	0	0.00
Employment				
Full-time	248	81.58	89	80.91
Unemployed	18	5.92	7	6.36
Student	8	2.63	1	0.91
Part-time	25	8.22	8	7.27
Retired	5	1.64	4	3.64
Other	7	2.30	3	2.73

RESULTS

Preliminary analysis

Demographic characteristics of our sample for Parts 1 and 2 are included in Table 1. In Part 1, 204 participants reported gambling problems in the past year, among whom 122 met NODS screening criteria for disordered gambling; 244 met lifetime screening criteria for disordered gambling. In Part 2, 45 participants reported gambling problems in the past year, and 73 participants met lifetime screening criteria for disordered gambling. *T*-tests were conducted to assess demographic differences between eligible participants who completed Part 2 versus those who did not. No significant differences were found, $ps \geq .54$. In addition, a χ^2 test examined whether individuals with a history of disordered gambling (current vs. past) were more or less likely to complete Part 2. The χ^2 test was not significant, $\chi^2(1) < 0.001$, $p = 1$.

In addition, we compared this sample on gambling-specific variables to previous research. In the present sample, we observed some similarities to community-based samples regarding associations between problematic gambling and frequency and types of gambling involvement. Males reported higher hours engaged in their most frequent activity ($M = 39.06$) compared to females ($M = 16.15$), consistent with prior research demonstrating that males gamble more frequently than females (Potenza, Maciejewski, & Mazure, 2006; Wong, Zane, Saw, & Chan, 2013). In addition, lottery tickets were the form of gambling most frequently engaged in by this sample, which is consistently reported in population surveys (2–3 times/month; Gerstein et al., 1999; Tu, 2013; Welte, Barnes, Tidwell, Hoffman, & Wieczorek, 2015). However, while video lottery terminal (VLT) engagement is consistently reported as the most problematic form of gambling (Currie et al., 2006; LaPlante, Nelson, LaBrie, & Shaffer, 2011; MacLaren, 2016), casino slots games showed a similarly high association with problem gambling symptomology ($r = .40$) as that between problem gambling and VLT engagement ($r = .37$).

Reliable responding

No participants completed the survey unusually quickly compared to other participants, and four took unusually long to complete (0.01%). Eighty participants (26%) showed inconsistency in their endorsement of disordered gambling symptoms, as demonstrated by different scores on at least one NODS-CLiP item and its corresponding NODS item. Most inconsistent responding was characterized by endorsement of items on the NODS-CLiP and not on the NODS. Individuals who consistently responded demonstrated significantly higher lifetime NODS ($M = 5.91$, $t = 10.58$, $p < .001$) and past-year scores ($M = 4.48$, $t = 7.83$, $p < .001$) compared to those who inconsistently responded ($M_{\text{lifetime}} = 2.88$, $M_{\text{past-year}} = 1.75$). These latter participants were not invited to complete Part 2, resulting in a total of 93 participants who completed the behavioral tasks.

Logical consistency and task reliability

Delay Discounting Task. The two criteria used to examine logically consistent responding in the DDT have been used in discounting tasks with a standard delay value of $1,000 and 7 delay intervals, as was the case in this study (Koffarnus & Bickel, 2014). First, the indifferent point at the first delay must be at least $100 greater than the indifference point at the longest delay. Second, no more than 1 indifference point can be $200 greater than the point preceding it. These criteria were employed to detect and exclude cases where the value of the delayed reward haphazardly fluctuated across delays, which might indicate that the participant did not understand the task, or that they did not attend to the questions (see Johnson & Bickel, 2008 for a discussion on identifying non-systematic discounting data). Eight participants were removed from analysis due to failure to meet the first criteria. No additional participants failed to meet the second criteria. AUCs ranged from 0.02 to 0.88 ($M = 0.232$, $SD = 0.21$; Figure 1).

Balloon Analogue Risk Task (BART). Adjusted average pump count (APC) ranged from 4.17 to 84.11 ($M = 53.56$, $SD = 14.77$). Interitem reliability was examined by comparing APC across trials. APCs across blocks of trials demonstrated moderate to high correlations (ρs = 0.67–0.78; Table 2), indicating good interitem reliability.

DISCUSSION

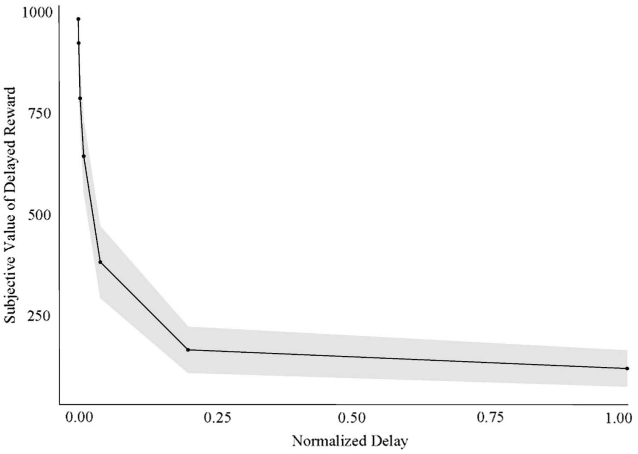

Figure 1. Overall delay discounting curve with 95% confidence intervals (CI) across delays. Area under the curve (AUC) = 0.23, 95% CI of AUC [0.18, 0.28]. Discounting curve created by plotting the average indifference point across the sample. Temporal delays are represented as a proportion of the maximum temporal delay. The shaded area illustrates the 95% CIs across all delay points: 1 day, 95% CI [977.64, 981.42]; 1 week, 95% CI [875.13, 965.34]; 1 month, 95% CI [710.23, 857.77]; 3 months, 95% CI [552.77, 730.53]; 1 year, 95% CI [292.14, 468.80]; 5 years, 95% CI [107.32, 221.15]; and 25 years, 95% CI [73.45, 162.78]

Research consistently demonstrates differences in various facets of behavioral impulsivity in disordered gambling samples (e.g., Kovacs, Richman, Janka, Maraz, & Ando, 2017). As such, behavioral tasks of impulsivity are frequently used, although their utility with online clinical samples has not been investigated. The purpose of this study was to examine the utility of using three common behavioral tasks with an online sample of individuals with current and lifetime problem and disordered gambling. Overall, this study supports the use of behavioral tasks with gambling samples recruited through MTurk. However, the results of the present research also suggest some caution and the need to act appropriately to minimize the effect of inattention and symptom misreporting.

This study screened for logically consistent responding in the DDT, which is frequently unaddressed in research using discounting tasks. Only eight participants were excluded from analyses due to failing to meet the criteria for logically consistent responding, suggesting that participants tended to respond consistently. Similarly, the study showed strong within task reliability on the BART, indicating that participants responded consistently.

Within the delay discounting literature, we found AUCs ranging between 0.21 and 0.32 among problem and disordered gambling samples (Krmpotich et al., 2015; Ledgerwood, Alessi, Phoenix, & Petry, 2009), similar to our reported mean AUC in this study (0.23). Regarding the BART, previous research has reported an APC of 37–45 in community samples (Hunt, Hopko, Bare, Lejuez, & Robinson, 2005; Krmpotich et al., 2015; White, Lejuez, & de Wit, 2008). Although addiction groups often display greater mean APC than control groups (Lauriola, Panno, Levin, & Lejuez, 2014), mean APCs have been as low as 21.3 in disordered gambling samples (Ledgerwood et al., 2009). However, the APC on the BART was slightly higher than expected compared with previous research. Next, we compared the within-session reliability of the BART. Specifically, previous research has reported strong split-third reliability (>.7; White et al., 2008)

Table 2. Spearman's ρ correlation coefficients for the adjusted average pump count (APC) on the Balloon Analogue Risk Task

	Overall APC	Trials 1–10	Trials 11–20	Trials 21–30
Trials 1–10	0.87			
Trials 11–20	0.92	0.73		
Trials 21–30	0.90	0.67	0.78	
M	53.54	55.86	51.08	55.62
SD	4.77	17.13	15.98	16.24

Note. SD: standard deviation.

across the blocks of trials, which was found in this study (.67–.78). On the whole, the results suggest that participants on MTurk demonstrate sustained attention and respond reliably to behavioral tasks as compared to individuals with disordered gambling recruited through traditional avenues.

Correlation magnitudes observed in this study between Delay Discounting AUC, BART APC, and Go/No-Go inhibition errors to each other and to UPPS-P subscale scores (Table 3). In general, the mean scores on the variables of interest in the behavioral tasks and their correlations with the UPPS-P scale scores were consistent to those seen in previous research. As impulsivity is multifaceted construct, the magnitude of these correlations is within the range that would be expected. The use of behavioral measures has revealed that the facets of impulsivity tapped into by these tasks reflect separate underlying processes in addictive disorders (De Wit, 2009). Therefore, they tend to demonstrate only modest correlations with each other. However, the strongest relationships are observed between measures of response inhibition (e.g., Go/No-Go tasks) and UPPS-P dimensions ($r = .10$–.13; Cyders & Coskunpinar, 2011). The magnitude of the relationship between inhibition errors on the horizontal target in this study demonstrated similar, if somewhat higher, correlations with UPPS-P scales. Delayed discounting of monetary rewards appear most strongly correlated with lack of planning ($r = .13$) and positive urgency (.13; Cyders & Coskunpinar, 2011). This study found a similar correlation between AUC and lack of premeditation ($r = .16$) but was unrelated to positive urgency. The BART correlates most strongly with the sensation-seeking subscale ($r = .08$–.12; Cyders & Coskunpinar, 2011), consistent with that observed in this study ($r = .08$).

Finally, the average scores on UPPS-P subscales (Table 3) was similar to that observed in other literature. For individuals with problem and disordered gambling, the highest subscale scores are typically found on sensation seeking, negative urgency, and positive urgency subscales (Clark et al., 2012; Haw, 2017; Michalczuk, Bowden-Jones, Verdejo-Garcia, & Clark, 2011). This same pattern was observed in the present project; individuals showed the highest scores on the sensation-seeking subscale ($M = 28.95$, $SD = 8.55$), followed by positive urgency ($M = 26.3$, $SD = 10.84$) and negative urgency ($M = 23.42$, $SD = 8.71$). Mean lack of premeditation was slightly lower ($M = 20.29$, $SD = 6.41$), and lack of perseveration had the lowest score ($M = 16.91$, $SD = 5.24$). Moreover, these scores were within the range of scores observed in previous research with problem and disordered gambling populations (Haw, 2017; Michalczuk et al., 2011). For example, Haw's (2017) reported mean UPPS-P subscale scores for men and women are similar to the scores we observed [*M*s = 15.61 ($SD = 5.50$; perseverance) to 28.78 ($SD = 8.98$; sensation seeking)]. The sample in the study of Michalczuk et al. (2011) was composed of individuals with disordered gambling and average UPPS-P subscale scores were higher than we observed, as we would expect. However, sensation seeking ($M = 33.4$, $SD = 5.9$), positive urgency ($M = 35.9$, $SD = 8.3$), and negative urgency ($M = 35.6$, $SD = 6.4$) continued to demonstrate the highest scores, with lack of premeditation ($M = 26.6$, $SD = 5.9$) and lack of perseverance ($M = 23.0$, $SD = 5.6$) showing the lowest. Altogether, the measures of behavioral impulsivity showed associations with each other and self-report impulsivity that are consistent with research conducted in lab settings.

Limitations

Although the study found promising findings for the utility of these behavioral assays with online clinical samples, several limitations should be considered. First, the statement of inclusion criteria in the recruitment notice may have encouraged symptom overreporting. As has been noted by other researchers (Kim & Hodgins, 2017), it is difficult (or impossible) to verify self-report data obtained online. The high proportion of inconsistent symptom responding highlights concerns with careless responding or symptom misreporting. Most of these inconsistencies arose from endorsement of symptoms on the NODS-CLiP screening items, followed by lack of endorsement on the identical NODS item later in the questionnaire. Participants were notified in the recruitment notice of the target population and therefore they may have been primed to endorse gambling symptoms in the screening items to quality for the study. Chandler and Paolacci (2017) recently demonstrated that many participants failed to meet eligibility criteria for MTurk studies and recommended that multiple-step screening processes be used to exclude ineligible participants. Thus, the two-step process of recruitment and screening used in this study allowed us to assess for inconsistent responding and exclude these participants, lending confidence to our results. In addition, this study also highlights the importance of multistep screening when it is not possible or ethical to completely blind participants to the purpose of the research. Second, this study did not include a group of individuals who completed the questionnaires and tasks in-person. Therefore, we cannot speak to the influence of method variance on the results we observed. However, the data obtained in this study generally showed similar psychometric properties when compared to the existing literature, lending some confidence to the validity of our findings. Third,

Table 3. Spearman's ρ correlation coefficients between behavioral task parameter, UPPS-P subscale and total scores, past-year NODS scores, and lifetime NODS scores

			Inhibition errors –	Inhibition errors –	UPPS-P	Negative	(Lack of)		Sensation	Positive	NODS	NODS
	AUC	APC	vertical target	horizontal target	total score	urgency	Premeditation	Perseveration	seeking	urgency	past year	lifetime
APC	0.16	1.00										
Inhibition errors – vertical target	−0.19	0.08	1.00									
Inhibition errors – horizontal target	−0.06	0.11	0.50	1.00								
UPPS-P total score	0.05	0.02	0.00	0.19	1.00							
Negative urgency	−0.01	−0.15	−0.02	0.03	**0.82**	1.00						
(Lack of) Premeditation	0.16	0.13	−0.04	0.14	**0.71**	**0.44**	1.00					
(Lack of) Perseveration	0.10	0.03	−0.02	0.15	**0.66**	**0.55**	**0.58**	1.00				
Sensation seeking	0.03	0.08	0.05	0.20	**0.66**	**0.35**	**0.35**	0.16	1.00			
Positive urgency	−0.03	0.02	0.02	0.17	**0.88**	**0.75**	**0.47**	**0.49**	**0.50**	1.00		
NODS past year	−0.15	0.05	0.04	0.09	**0.25**	**0.26**	0.17	0.21	0.13	**0.29**	1.00	
NODS lifetime	−0.05	0.14	0.02	−0.05	0.15	**0.24**	0.08	0.07	0.05	0.16	**0.84**	1.00
M	0.23	53.54	0.03	0.03	115.87	23.42	20.29	16.91	28.95	26.30	4.38	5.71
SD	0.22	14.77	0.12	0.11	31.18	8.71	6.41	5.24	8.55	10.84	2.16	2.11

Note. Bold values indicate correlation is significant at $p \leq .05$. Note that an exact *p* value was not able to be calculated due to rank ties. Therefore, all *p* values were estimates and should be interpreted with caution. NODS past-year scores included only participants who had gambled in the prior 12 months. *SD*: standard deviation; AUC: area under the empirical discounting curve for the delay discounting task; APC: adjusted average pump count for the Balloon Analogue Risk Task; NODS: NORC DSM-IV Screen for Gambling Problems.

these tasks lack a clear gold standard with which to evaluate responses. A consistent concern raised is their poor correlations with trait measures of impulsivity (Sharma, Markon, & Clark, 2014), highlighting potential issues with their construct validity. We recognize that lack of such a gold standard is a significant limitation in any research using behavioral tasks. However, impulsivity is increasingly recognized as multifaceted construct, with behavioral tasks tapping in to distinct facets compared to self-report measures of "trait" impulsivity (De Wit, 2009). Thus, the relationships observed between the UPPS-P and behavioral tasks were consistent with previous literature and our current understanding of impulsivity. In consideration of this limitation, only widely used behavioral tasks were selected, which generally show high reliability. In addition, the tasks examined in this study have demonstrated predictive validity on a variety of important outcomes, including future drug use (De Wit, 2009; Perry & Carroll, 2008), problem gambling, (Verdejo-García, Lawrence, & Clark, 2008; Vitaro, Arseneault, & Tremblay, 1999), and substance-use treatment outcomes (Stevens et al., 2014). Finally, MTurkers are a non-representative sample. In the present sample, we observed differences, as well as similarities, on associations between gambling-related variables compared to those seen in community-based samples. Of note, although VLT and casino slots are considered to be measures of separate types of gambling, they share significant similarities in play. This may contribute to the similarly high associations with problem gambling symptomology that we observed. Overall, while MTurk samples are not representative of the general gambling population, it appears to be a useful tool for recruiting samples characterized by problematic gambling.

Recommendations

Given our findings, we make the following recommendations for obtaining quality data from gambling samples on MTurk:

1. Researchers should consider multistep screening when recruiting clinical populations and examine the consistency of symptom reporting specifically.
2. Analyses should include examination of response consistency on reward-related decision-making tasks (e.g., probabilistic or delayed reward discounting).
3. Studies utilizing behavioral tasks that have not been directly examined in an online clinical population should pilot the tasks and recruitment strategy.
4. Continue to uphold best research practices with MTurk. Such recommendations can be found in the study of Chandler and Shapiro (2016) or Chandler and Paolacci (2017).
5. Be cognizant that MTurkers are non-representative samples. In regard to gambling characteristic, the proportion of problem gamblers tends to be greater than that of traditional avenues (Kim & Hodgins, 2017).

CONCLUSIONS

Overall, the random-adjusting DDT and Cued Go/No-Go Tasks appear to be viable for use with addiction samples collected through MTurk. The BART that may be more susceptible to inattention and caution is warranted when using hypothetical rewards. Further research should investigate the potential impact of real versus hypothetical rewards with online clinical samples before adopting it for web-based use. In conclusion, this is the first study to highlight the potential of using behavioral tasks with clinical MTurk samples. As we continue to push the boundaries of knowledge and how we study addiction, the use of behavioral tasks through MTurk might be a promising avenue to advance this goal.

Funding sources: None.

Authors' contribution: MGS: study concept and design, analysis and interpretation of data, and statistical analysis. HSK: study concept and design and analysis and interpretation of data. DCH: study concept and design and study supervision. All authors had full access to all data and take responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of interest: The authors declare no conflict of interest.

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment, 28*(6), 684. doi:10.1037/pas0000217

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351–368. doi:10.1093/pan/mpr057

Blumberg, S. J., & Luke, J. V. (2007). Coverage bias in traditional telephone surveys of low-income and young adults. *Public Opinion Quarterly, 71*(5), 734–749. doi:10.1093/poq/nfm047

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. doi:10.1177/1745691610393980

Chandler, J., & Paolacci, G. (2017). Lie for a Dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science, 8*(5), 500–508. doi:10.1177/1948550617698203

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*(1), 53–81. doi:10.1146/annurev-clinpsy-021815-093623

Clark, L., Stokes, P. R., Wu, K., Michalczuk, R., Benecke, A., Watson, B. J., Egerton, A., Piccini, P., Nutt, D. J., Bowden-Jones, H., & Lingford-Hughes, A. R. (2012). Striatal dopamine D2/D3 receptor binding in pathological gambling is correlated with mood-related impulsivity. *Neuroimage, 63*(1), 40–46. doi:10.1016/j.neuroimage.2012.06.067

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

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), e57410. doi:10.1371/journal.pone.0057410

Currie, S. R., Hodgins, D. C., Wang, J., El-Guebaly, N., Wynne, H., & Chen, S. (2006). Risk of harm among gamblers in the general population as a function of level of participation in gambling activities. *Addiction, 101*(4), 570–580. doi:10.1111/j.1360-0443.2006.01392.x

Cyders, M. A., & Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity? *Clinical Psychology Review, 31*(6), 965–982. doi:10.1016/j.cpr.2011.06.001

De Wit, H. (2009). Impulsivity as a determinant and consequence of drug use: A review of underlying processes. *Addiction Biology, 14*(1), 22–31. doi:10.1111/j.1369-1600.2008.00129.x

Fillmore, M. T. (2003). Drug abuse as a problem of impaired control: Current approaches and findings. *Behavioral and Cognitive Neuroscience Reviews, 2*(3), 179–197. doi:10.1177/1534582303257007

Gerstein, D., Volberg, R. A., Toce, M. T., Harwood, H., Johnson, R. A., Buie, T., Christiansen, E., Chuchro, L., Cummings, W., Engelman, L., & Hill, M. A. (1999). *Gambling impact and behavior study: Report to the national gambling impact study commission*. Chicago, IL: National Opinion Research Center.

Godinho, A., Kushnir, V., & Cunningham, J. A. (2016). Unfaithful findings: Identifying careless responding in addictions research. *Addiction, 111*(6), 955–956. doi:10.1111/add.13221

Haw, J. (2017). Impulsivity predictors of problem gambling and impaired control. *International Journal of Mental Health and Addiction, 15*(1), 154–165. doi:10.1007/s11469-015-9603-9

Huff, C., & Tingley, D. (2015). Who are these people? Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics, 2*(3), 1–12. doi:10.1177/2053168015604648

Hunt, M. K., Hopko, D. R., Bare, R., Lejuez, C., & Robinson, E. (2005). Construct validity of the Balloon Analog Risk Task (BART) associations with psychopathy and impulsivity. *Assessment, 12*(4), 416–428. doi:10.1177/1073191105278740

Johnson, M. W., & Bickel, W. K. (2008). An algorithm for identifying nonsystematic delay-discounting data. *Experimental and Clinical Psychopharmacology, 16*(3), 264–274. doi:10.1037/1064-1297.16.3.264

Kim, H. S., & Hodgins, D. C. (2017). Reliability and validity of data obtained from alcohol, cannabis, and gambling populations on Amazon's Mechanical Turk. *Psychology of Addictive Behaviors, 31*(1), 85–94. doi:10.1037/adb0000219

Kim, H. S., Wohl, M. J. A., Salmon, M., & Santesso, D. (2017). When do gamblers help themselves? Self-discontinuity increases self-directed change over time. *Addictive Behaviors, 64,* 148–153. doi:10.1016/j.addbeh.2016.08.037

Koffarnus, M. N., & Bickel, W. K. (2014). A 5-trial adjusting delay discounting task: Accurate discount rates in less than one minute. *Experimental and Clinical Psychopharmacology, 22*(3), 222–228. doi:10.1037/a0035973

Kovacs, I., Richman, M. J., Janka, Z., Maraz, A., & Ando, B. (2017). Decision making measured by the Iowa Gambling Task in alcohol use disorder and gambling disorder: A systematic review and meta-analysis. *Drug Alcohol Depend, 181,* 152–161. doi:10.1016/j.drugalcdep.2017.09.023

Krmpotich, T., Mikulich-Gilbertson, S., Sakai, J., Thompson, L., Banich, M. T., & Tanabe, J. (2015). Impaired decision-making, higher impulsivity, and drug severity in substance dependence and pathological gambling. *Journal of Addiction Medicine, 9*(4), 273–280. doi:10.1097/ADM.0000000000000129

LaPlante, D. A., Nelson, S. E., LaBrie, R. A., & Shaffer, H. J. (2011). Disordered gambling, type of gambling and gambling involvement in the British Gambling Prevalence Survey 2007. *European Journal of Public Health, 21*(4), 532–537. doi:10.1093/eurpub/ckp177

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2014). Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the Balloon Analogue Risk Task. *Journal of Behavioral Decision Making, 27*(1), 20–36. doi:10.1002/bdm.1784

Ledgerwood, D. M., Alessi, S. M., Phoenix, N., & Petry, N. M. (2009). Behavioral assessment of impulsivity in pathological gamblers with and without substance use disorder histories versus healthy controls. *Drug and Alcohol Dependence, 105*(1), 89–96. doi:10.1016/j.drugalcdep.2009.06.011

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. doi:10.1037//1076-898X.8.2.75

Lynam, D. R., Smith, G. T., Whiteside, S. P., & Cyders, M. A. (2006). *The UPPS-P: Assessing five personality pathways to impulsive behavior*. West Lafayette, IN: Purdue University.

MacKillop, J., Amlung, M. T., Few, L. R., Ray, L. A., Sweet, L. H., & Munafò, M. R. (2011). Delayed reward discounting and addictive behavior: A meta-analysis. *Psychopharmacology, 216*(3), 305–321. doi:10.1007/s00213-011-2229-0

MacLaren, V. V. (2016). Video lottery is the most harmful form of gambling in Canada. *Journal of Gambling Studies, 32*(2), 459–485. doi:10.1007/s10899-015-9560-z

Madden, G. J., & Bickel, W. K. (2010). Introduction. In G. J. Madden & W. K. Bickel (Eds.), *Impulsivity: The behavioral and neurological science of discounting* (pp. 3–8). Washington, DC: American Psychological Association.

Michalczuk, R., Bowden-Jones, H., Verdejo-Garcia, A., & Clark, L. (2011). Impulsivity and cognitive distortions in pathological gamblers attending the UK National Problem Gambling Clinic: A preliminary report. *Psychological Medicine, 41*(12), 2625–2635. doi:10.1017/S003329171100095X

Mishra, S., & Carleton, R. N. (2017). Use of online crowdsourcing platforms for gambling research. *International Gambling Studies, 17*(1), 125–143. doi:10.1080/14459795.2017.1284250

Myerson, J., Green, L., & Warusawitharana, M. (2001). Area under the curve as a measure of discounting. *Journal of the Experimental Analysis of Behavior, 76*(2), 235–243. doi:10.1901/jeab.2001.76-235

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188. doi:10.1177/0963721414531598

Perry, J. L., & Carroll, M. E. (2008). The role of impulsive behavior in drug abuse. *Psychopharmacology, 200*(1), 1–26. doi:10.1007/s00213-008-1173-0

Potenza, M. N., Maciejewski, P. K., & Mazure, C. M. (2006). A gender-based examination of past-year recreational gamblers. *Journal of Gambling Studies, 22*(1), 41–64. doi:10.1007/s10899-005-9002-4

R Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Richards, J. B., Zhang, L., Mitchell, S. H., & de Wit, H. (1999). Delay or probability discounting in a model of impulsive behavior: Effect of alcohol. *Journal of the Experimental Analysis of Behavior, 71*(2), 121–143. doi:10.1901/jeab.1999.71-121

Shao, W., Guan, W., Clark, M. A., Liu, T., Santelices, C., Cortes, D. E., & Merchant, R. C. (2015). Variations in recruitment yield, costs, speed and participant diversity across internet platforms in a global study examining the efficacy of an HIV/AIDS and HIV testing animated and live-action video among English-or Spanish-speaking Internet or social media users. *Digital Culture & Education, 7*(1), 40–86. Retrieved from http://europepmc.org/abstract/MED/27330570

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science, 1*(2), 213–220. doi:10.1177/2167702612469015

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of impulsive behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin, 140*(2), 374–408. doi:10.1037/a0034418

Stevens, L., Verdejo-García, A., Goudriaan, A. E., Roeyers, H., Dom, G., & Vanderplasschen, W. (2014). Impulsivity as a vulnerability factor for poor addiction treatment outcomes: A review of neurocognitive findings among individuals with substance use disorders. *Journal of Substance Abuse Treatment, 47*(1), 58–72. doi:10.1016/j.jsat.2014.01.008

Suurvali, H., Hodgins, D., Toneatto, T., & Cunningham, J. (2008). Treatment seeking among Ontario problem gamblers: Results of a population survey. *Psychiatric Services, 59*(11), 1343–1346. doi:10.1176/ps.2008.59.11.1343

Toce-Gerstein, M., Gerstein, D., & Volberg, R. (2009). The NODS-CLiP: A rapid screen for adult pathological and problem gambling. *Journal of Gambling Studies, 25*(4), 541–555. doi:10.1007/s10899-009-9135-y

Toce-Gerstein, M., & Volberg, R. (2004). *The NODS-CLiP: A new brief screen for pathological gambling*. Paper presented at the International Symposium on Problem Gambling and Co-Occurring Disorders, Mystic, CT.

Tu, D. (2013). *New Zealanders' participation in gambling: Results from the 2012 health and lifestyles survey*. Wellington, New Zealand: Health Promotion Agency.

Verdejo-García, A., Lawrence, A. J., & Clark, L. (2008). Impulsivity as a vulnerability marker for substance-use disorders: Review of findings from high-risk research, problem gamblers and genetic association studies. *Neuroscience and Biobehavioral Reviews, 32*(4), 777–810. doi:10.1016/j.neubiorev.2007.11.003

Vitaro, F., Arseneault, L., & Tremblay, R. E. (1999). Impulsivity predicts problem gambling in low SES adolescent males. *Addiction, 94*(4), 565–575. doi:10.1046/j.1360-0443.1999.94456511.x

Vogel-Sprott, M., & Fillmore, M. T. (2011). Learning, expectancy, and behavioral control implications for drug abuse. In T. R. Schachtman & S. S. Reilly (Eds.), *Associative learning and conditioning theory: Human and non-human applications* (pp. 213–234). New York, NY: Oxford University Press.

Welte, J. W., Barnes, G. M., Tidwell, M.-C. O., Hoffman, J. H., & Wieczorek, W. F. (2015). Gambling and problem gambling in the United States: Changes between 1999 and 2013. *Journal of Gambling Studies, 31*(3), 695–715. doi:10.1007/s10899-014-9471-4

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*(6), 565–570. doi:10.1037/a0014083

Wong, G., Zane, N., Saw, A., & Chan, A. K. K. (2013). Examining gender differences for gambling engagement and gambling problems among emerging adults. *Journal of gambling studies, 29*(2), 171–189. doi:10.1007/s10899-012-9305-1