Published in final edited form as:
Behav Processes. 2013 November ; 100: 116–122. doi:10.1016/j.beproc.2013.07.019.

Statistical Equivalence and Test-Retest Reliability of Delay and Probability Discounting Using Real and Hypothetical Rewards

Alexis K. Matusiewicz[a], **Anne E. Carter**[b], **Reid D. Landes**[c], and **Richard Yi**[a]

[a]Center for Addictions, Personality and Emotion Research, Department of Psychology, 2103 Cole Field House, University of Maryland, College Park, MD 20742, USA

[b]Advanced Recovery Research Center, Virginia Tech Carilion School of Medicine and Research Institute, 2 Riverside Circle, Roanoke, VA 24016, USA

[c]Department of Biostatistics, University of Arkansas for Medical Sciences, 4301 W. Markham St., Slot 781, Little Rock, AR 72205, USA

Abstract

Delay discounting (DD) and probability discounting (PD) refer to the reduction in the subjective value of outcomes as a function of delay and uncertainty, respectively. Elevated measures of discounting are associated with a variety of maladaptive behaviors, and confidence in the validity of these measures is imperative. The present research examined (1) the statistical equivalence of discounting measures when rewards were hypothetical or real, and (2) their 1-week reliability. While previous research has partially explored these issues using the low threshold of nonsignificant difference, the present study fully addressed this issue using the more-compelling threshold of statistical equivalence. DD and PD measures were collected from 28 healthy adults using real and hypothetical $50 rewards during each of two experimental sessions, one week apart. Analyses using area-under-the-curve measures revealed a general pattern of statistical equivalence, indicating equivalence of real/hypothetical conditions as well as 1-week reliability. Exceptions are identified and discussed.

Keywords

Delay Discounting; Humans; Hypothetical Outcomes; Probability Discounting; Real Outcomes; Statistical Equivalence; Test-Retest Reliability

Corresponding Author: Alexis K. Matusiewicz, M.S., Center for Addictions, Personality, and Emotion Research, Department of Psychology, University Maryland, College Park, 2103R Cole Field House, College Park, MD 20742, amatus@umd.edu, phone: (301) 405-8441, fax: (301) 405-3223.

The authors report no potential conflicts of interest.

1. Introduction

Reductions in the subjective value of a reward as a function of delay to (Bickel et al., 1999; Rachlin et al., 2000) or uncertainty of (Myerson and Green, 2004; Rachlin et al., 1991) the reward are normative. However, *excessive* discounting of delayed rewards is associated with engagement in a range of health compromising behaviors (Bickel et al., 2012; Melanko & Larkin, 2012); elevated rates of delay discounting (DD) are observed in licit and illicit drug users (Reynolds, 2006; Yi et al., 2010), pathological gamblers (Petry, 2001), binge eaters (Davis et al., 2009), obese individuals (Weller et al., 2008), and other populations that exhibit deficits of impulse control. In contrast, the literature on discounting of probabilistic outcomes is mixed. Elevated rates of probability discounting (PD) are observed in cigarette smokers (Reynolds et al., 2004; Yi et al., 2007), obese women and women with binge eating disorder (Manwaring et al., 2011; but see Madden et al., 2009 regarding pathological gamblers). Both DD and PD have been suggested as possible predictors of treatment response and/or markers of progress in treatment of addictive behaviors (Bickel & Marsch, 2001; Krishnan-Sarin et al., 2007; Landes et al., 2012; Petry, 2012; Sheffer et al., 2012). Thus, confidence in the reliability and validity of discounting measures is imperative.

1.1 Discounting of Real and Hypothetical Rewards

Discounting studies with human subjects typically employ a fungible commodity such as money, and it has been suggested that using real money rewards increases the likelihood that the participant will respond based on his or her actual preferences (Madden et al., 2004). Paradigms in which subjects experience the consequences of every trial are used almost exclusively in animal research, and are uncommon in human subjects research (for exceptions, see Lagorio & Madden, 2005; Lane, Cherek, Pietras, & Tcheremissine, 2003; Johnson, 2012; Scheres, Dijkstra, Ainslie, Balkan, Reynolds, Sonuga-Barke, Castellanos, 2006). Unfortunately, the use of real rewards significantly increases costs associated with conducting discounting research, and limits the magnitude of rewards and the duration of delays that can be queried (specifically for DD; Johnson and Bickel, 2002; Lawyer et al., 2011), even in paradigms in which participants experience the consequences of one choice trial that is randomly selected from all trials (i.e., potentially real or lottery-style rewards).

Hypothetical rewards are often used to avert these logistical challenges, and participants are instructed to respond as if the choices were real (e.g., Johnson & Bickel, 2002). To address concerns that participants may respond differently to real and hypothetical rewards (Kirby, 1997), a number of studies have compared DD of real rewards and hypothetical rewards, consistently finding no statistical difference (Baker et al., 2003; Johnson & Bickel, 2002; Lagorio & Madden, 2005; Lawyer et al., 2011; Madden et al., 2003; 2004). In addition, fMRI conducted during DD of real and hypothetical rewards indicates congruence of regions of brain activation when real and hypothetical rewards are considered (Bickel et al., 2009). Though fewer studies have compared real and hypothetical rewards on PD outcomes, at least two studies have found no significant difference in probability discounting of real and hypothetical rewards (Hinvest & Anderson, 2010; Lawyer et al., 2011; see Jikko & Okouchi, 2007 for contradictory results). Moreover, delay and probability discounting of

potentially-real rewards is highly correlated with discounting of hypothetical rewards (Lawyer et al., 2011; Yi & Landes, 2012).

Because the preponderance of available data indicates no statistical differences in discounting of real and hypothetical rewards, and evidence indicates a statistical relationship between them, some researchers have perhaps inaccurately interpreted real and hypothetical rewards as equivalent and interchangeable. Extant research, however, has used the relatively low statistical bar of non-significant difference, rather than the more compelling threshold of significant equivalence. Statistical equivalence testing offers a more rigorous test of the equivalence of real and hypothetical rewards, supporting the prudent use of hypothetical rewards in discounting research.

1.2 Test-Retest Reliability of Delay and Probability Discounting

In addition to the statistical equivalence of discounting of real and hypothetical rewards, the stability of these assessments over the course of repeated administrations has conceptual and methodological significance. Discounting is commonly conceptualized as a trait variable (Odum, 2011). Therefore, an individual's discount rate should be relatively unchanged over repeated administrations in the absence of any systematic source of variance. Adequate test-retest reliability is critical to determine whether the procedure accurately measures the construct of interest without excessive measurement error.

Two indices of test-retest reliability are common: relative reliability and absolute reliability (Baumgartner, 1989; Weir, 2005). Relative reliability refers to the consistency of an individual's rank position (relative to others) over repeated measurements. Absolute reliability refers to absolute differences in the group's mean score over successive procedurally identical assessments (e.g., stability). Relative and absolute test-retest reliability provide complementary information about the reliability and repeatability of a measure, both of which are critical to make inferences about the stability of discounting, or mark changes in discounting as a function of other factors.

Studies investigating the relative reliability of DD of real rewards have shown modest-to-strong relative reliability at test-retest intervals of 1 week (Baker et al., 2003; Johnson et al., 2007), 6 weeks (Kirby, 2009; Beck & Triplett, 2009), 17 weeks (Peter and Büchel, 2009), and 1 year (Kirby, 2009). A similar pattern has been observed with DD of hypothetical rewards at test-retest intervals of 1 week (Simpson & Vuchinich, 2000; Baker et al., 2003; Johnson et al., 2007), 5 weeks (Jimura et al., 2011), and 12 weeks (Ohmura et al., 2006). Of the two studies that assessed absolute reliability of DD for real rewards, one reported a trend for increasing discount rates over successive administrations (Kirby, 2009), while the other found no significant effect of time on DD of real rewards (Beck & Triplett, 2009). Of the three studies that assessed absolute stability of DD of hypothetical rewards, none found significant effects of time on discounting (Audrain-McGovern et al., 2009; Simpson & Vuchinich, 2000; Ohmura et al., 2006). The reliability of PD is less clear, as we are aware of only two discounting studies incorporating this analysis of PD. These studies on PD of real (Peters & Büchel, 2009) and hypothetical (Ohmura et al., 2006) rewards indicate high relative reliability, and no statistical difference between testretest assessments (Ohmura et al., 2006). Thus, the preponderance of evidence appears to support good reliability of

discounting assessments. However, particularly in the domain of absolute reliability, this conclusion may be based on the conflation of non-significant difference with statistical equivalence.

1.3 Statistical Equivalence

Despite consistent evidence that discounting of real and equivalent rewards *do not differ significantly*, and that there *are not significant differences* in discounting as a function of time, statistical or methodological equivalence remains in question; null hypothesis significance testing alone cannot support these conclusions. Because null hypothesis significance testing cannot support an *absence of effect* (Wagenmakers, 2007), statistical equivalence testing is appropriate for cases in which the focal hypothesis concerns the absence of an effect (Gallistel, 2009).

In equivalence testing, the null hypothesis is that the two sets of measurements differ, and the alternative hypothesis is that the measurements are statistically equivalent (Welleck, 2010). To establish statistical equivalence between discounting of real and hypothetical rewards requires evidence that measurements do not differ appreciably; acceptance of the alternate hypothesis supports the interpretation that there is no effect of an independent variable (e.g., reward type, time) on the dependent variable (e.g., discounting). Although the results of null hypothesis significance testing and statistical equivalence testing may not lead to dramatically different conclusions, there is a clear need for theoretical precision both in the formulation of hypotheses and the analytic approach chosen to test hypotheses. If the research is motivated by the question of equivalence of two experimental conditions, statistical equivalence testing offers a more compelling test. However, statistical equivalence testing is not widely adopted in the field of psychology (Gallistel, 2009; Rouder et al., 2011), and has not yet been applied in the study of discounting.

1.4 Current Study

The aim of the current study was to (1) examine the relationship between discounting of real and hypothetical rewards for DD and PD; and (2) to examine the one-week test-retest reliability of DD and PD of real and hypothetical rewards, using the appropriate threshold of statistical equivalence. Consistent with previous research on DD and the limited literature on PD, which suggest the absence of differences in discounting as a function of reward type, we hypothesized that discounting of real and hypothetical rewards would be statistically equivalent for both delay and probability. In addition, given previous work indicating good test-retest reliability of discount rates over brief test-retest intervals, we hypothesized statistical equivalence of repeated assessments, separated by one week, of DD and PD.

2. Material and Methods

2.1 Participants

Twenty-eight participants (11 females, 17 males) between 19 and 60 years of age were recruited locally via flyers, advertisements, and word of mouth referrals. Participants did not meet dependence criteria for any substance, were free from major psychiatric and medical disorders, and females were not pregnant at the time of participation.

2.2 Measures

2.2.1 Delay discounting procedure—DD was assessed using a binary-choice computerized program based on Holt, Green, and Myerson (2003). Participants were presented with a series of trials in which they were asked to choose between $50 that was delayed and a smaller sum of money that was available immediately. The value of the immediate outcome was titrated across 6 trials to determine the present, subjective value (indifference point) of $50 at each delay. In the hypothetical reward condition, indifference points were determined for $50 delayed by 1 day, 1 week, 1 month, 6 months, 1 year, 5 years, and 25 years. In the real reward condition, indifference points were determined for $50 delayed by 1 day, 1 week, 1 month, and 6 months. One of the trials was randomly selected from the DD of real rewards procedure at the conclusion of each session. If the randomly selected trial was for the immediate reward, participants were compensated at the end of the experimental session; if the randomly selected trial was for the delayed reward, the reward was sent to the participant by mail following the delay specified for the selected trial. Consistent with the conventions of human discounting research, and to avoid logistical challenges associated with administering real rewards at protracted delays, a smaller range of delays was assessed in the real reward condition than in the hypothetical reward condition.

2.2.2 Probability discounting procedure—PD was assessed using a similar computerized binary choice paradigm. Participants were presented with a series of trials in which they were asked to choose between $50 that was probabilistic and a smaller sum of money that was certain. In the hypothetical reward condition, indifference points were determined for $50 with probabilities of 95%, 75%, 50%, 25%, 5% and 1%. In the real reward condition, indifference points were determined for $50 with probabilities of 50%, 25%, 5%, and 1%. The smaller range of probabilities in the real reward condition (vs. hypothetical reward condition) was intended to mirror the DD procedure. In the real reward condition, one of the trials was randomly selected at the conclusion of each session. If the selected trial was one in which the participant had selected the certain reward, the participant was compensated at the end of the experimental session. If the selected trial was one in which the participant had selected the probabilistic reward, the participant drew a marble from an opaque bag with a distribution of win/no-win marbles that mirrored the probability specified in the randomly selected trial. Participants who won were compensated at the end of the experimental session.

2.3 Procedure

Participation was completed over the course of 3 visits, consisting of an informed consent session and 2 experimental sessions. During the informed consent session, participants completed a screening for significant medical, psychiatric, and drug use history, in addition to other self-report measures not included in the current analyses. The following 2 experimental sessions were procedurally identical, and completed 1 week apart. In each experimental session participants completed the DD and PD measures for both real and hypothetical rewards. At the end of each experimental session, one trial from the real rewards condition for both DD and PD was selected at random and participants received the outcome he or she chose for that trial. The order of discounting procedures was

counterbalanced across participants (i.e., half completed DD first; half completed the real reward condition first). For each participant, order of conditions was consistent in both sessions.

2.4 Data Scoring and Analysis

Discounting rates were determined using the model-free Area-Under-the-Curve (AUC; Myerson et al., 2001), where high and low AUC values indicate less and more discounting, respectively. The relation between time and subjective value was a negatively decelerating function, as was the relation between probability of reward and subjective value. To facilitate direct comparisons of AUC values for real and hypothetical rewards, the AUC values in the hypothetical reward conditions were calculated using the same four delays and probabilities as in the real reward conditions. Spearman rank correlations were calculated to evaluate the relation between discounting of real and hypothetical rewards within each session, and to establish relative reliability across sessions. Analyses were conducted separately for DD and PD.

Tests of statistical equivalence require specification of an equivalence region, within which two values can be said to be essentially equal, while allowing for minor deviations (Wellek, 2010). Because it is unclear which of the two measures will be greater, the bounds of the equivalence region are defined as ratios of the two measurements of central tendency: •edian$_1$/median$_2$•• A typical statistical equivalence region is 4/5 and 5/4, which has been adopted in biomedical research to establish the bioequivalence of new pharmacological agents to established treatments (FDA, 2001; Luzar-Stiffler and Stiffler, 2002); in other words, according to these guidelines, two measurements may be considered equivalent if one value is between 80 and 125% of the other value. To establish statistical equivalence, two conditions must be met. First, the 100(1–2<alpha>)% confidence interval must fall within the bounds of the established equivalence region (.8 to 1.25). Second, the 100(1-<alpha>)% confidence interval must cover 1.0; if this confidence interval does not include 1.0, the two measures are statistically different at the specified alpha level. If two measures are statistically different, then convention says they cannot also be statistically equivalent, even if the defined confidence interval lies within the defined equivalence region.

Statistical equivalence of discounting of real and hypothetical rewards was determined by comparing discounting of real and hypothetical rewards in session 1, and again for measurements taken in session 2. Absolute reliability of discounting was determined by comparing across session 1 and 2. Analyses were conducted separately for DD and PD.

3. Results

3.1 Discounting of Real and Hypothetical Rewards

Figure 1A presents raw discounting estimates for the ratio of DD of real/hypothetical rewards, and the ratio of PD of real/hypothetical rewards. Filled points fall within the designated equivalence region, while open points do not. Figure 1B summarizes the evidence for statistical equivalence of DD of real and hypothetical rewards, and PD of real and hypothetical rewards within each experimental session. Each horizontal line is marked

at the median ratio of discounting of real/hypothetical rewards for each condition. The thick bars represent the 90% (1-2[alpha]%) confidence interval for the median ratio. The thin bars represent the 95% (1-[alpha]%) confidence interval for the median ratio. Recall that there are two criteria for statistical equivalence: first, the 90% confidence interval must fall within the bounds of the established equivalence region (.8 to 1.25), marked here with broken vertical lines. Second, the 95% confidence interval must include 1.0, to establish that the two measurements are not, in fact, statistically different.

Figure 1B reveals that the 90% confidence interval for the ratio of DD of real/hypothetical rewards fell within the established equivalence region for both sessions. In addition, for both sessions, the 95% confidence interval for the ratio of DD of real/hypothetical rewards included 1.0, providing support for the statistical equivalence of DD of real and hypothetical rewards. Likewise, in both sessions, the 90% confidence interval for the ratio of PD of real/hypothetical rewards fell within the specified equivalence region and, in both sessions, the 95% confidence interval included 1.0. Thus, PD of real rewards was statistically equivalent to PD of hypothetical rewards.

Correlational analyses provide further evidence of the statistical relations between discounting of real and hypothetical rewards (Table 1): DD of real rewards was highly correlated with DD of hypothetical rewards (ρ = .80 to .85), and PD of real rewards was moderately correlated with PD of hypothetical rewards (ρ = .52 to .66).

3.2 Test-Retest Reliability of Delay and Probability Discounting

Figure 2A provides the ratio of DD in session 1/session 2, and the ratio of PD in session 1/session 2, for each subject as a function of reward type. Filled points fall within the designated equivalence region, while open points do not. Figure 2B summarizes the evidence for statistical equivalence (i.e., absolute stability) of DD and PD over one week. As before, each horizontal line is marked at the median ratio of discounting in session 1/session 2. The thick bars represent the 90% confidence interval for the median ratio, and must fall entirely within the bounds of the equivalence region (.8 to 1.25) to support statistical equivalence. The thin bars represent the 95% confidence interval for the median ratio of session 1/session 2, and must include 1.0 to support statistical equivalence.

Figure 2B shows that DD of real rewards in session 1 was statistically equivalent to DD of real rewards in session 2. However, for DD of hypothetical rewards, the 90% confidence interval for the median ratio of session 1/session 2 fell outside the lower boundary of the equivalence region, indicating that these measurements were not statistically equivalent. In fact, the 95% confidence interval for the median ratio did not cross 1.0, suggesting that DD of hypothetical rewards differed significantly in session 1 and session 2. PD of real rewards in session 1 was statistically equivalent to PD of real rewards in session 2. Likewise, PD of hypothetical rewards was statistically equivalent across sessions.

Table 1 reveals relations of discounting across sessions, an index of relative reliability. For DD of real and hypothetical rewards, correlations revealed a significant, positive relationship between the discount rates obtained in each session (ρ = .70 to .73). For PD,

correlations revealed a significant, positive relation for hypothetical rewards (ρ = .51), and a non-significant, positive relation for real rewards (ρ = .34, *p* = .08).

3.3. Post-Hoc Analyses

The unexpectedly low correlation of PD of real rewards across sessions led us to suspect that the outcome of the real money procedures in session 1 may have affected choice in the session 2 real money procedures. Recall that, as part of the real reward conditions in each experimental session, participants' compensation was determined by randomly selecting one trial from the DD procedure and another from the PD procedure. Thus, in the DD of real rewards condition, participants experienced one of two outcomes during the first session: (1) receive an immediate amount of money of less than $50, or (2) receive a delayed $50 reward. Participants experienced one of three different outcomes during the first session in the PD of real money condition: (1) receive an amount of money smaller than $50 (certain), (2) win $50, or (3) not win any money. Post-hoc analyses were conducted to explore the possibility that the outcome experienced in the real rewards PD trial affected PD of real rewards on the subsequent trial.

Participants were categorized on the basis of their real reward outcomes for delay (smaller immediate vs. larger, delayed) and probability discounting (certain vs. probabilistic). There were 3 who received smaller, immediate and certain outcomes, 4 who received the smaller, immediate and probabilistic outcomes, 10 who received the larger, delayed and certain outcomes, and 11 who received the larger, delayed and probabilistic outcomes. (With one exception, all participants who received the delayed outcome in the first session received their compensation after the second session). Within each type of discounting (delay or probability), a difference score was calculated by subtracting each participant's session 1 AUC from the session 2 AUC. These differences were analyzed in a repeated measures ANOVA accounting for outcomes received, discounting type (delay or probability), and their interaction. The latter two were within-subject factors, for which we used an exchangeable (i.e., compound symmetric) structure to model the within-subject covariance. Error degrees of freedom were estimated with the Kenward-Roger method.

For delay discounting, participants who received the smaller, immediate reward in session 1, showed a decrease of .067 in Session 2 (i.e., showed an increase in discounting). Those who received the larger delayed reward in session 1 showed an increase of .086 (i.e., showed a decrease in discounting). The 0.153 difference approached statistical significance (t[43.5]=1.94, p=.059).

For probability discounting, those who received a probabilistic outcome in session 1 showed a 0.175 increase in AUC in session 2; this was a statistically significant increase (t[43.5]=3.29, p=.002) from session 1 values. Those who received a certain outcome showed a .083 decrease in their session 2 AUC. The 0.258 difference from the change experienced by those receiving probabilistic outcomes was significant (t[43.5]=3.25, p = .002). Follow-up analyses revealed that those who did not win the probabilistic reward showed a decrease in discounting of .114 in session 2, and those who won the probabilistic reward showed a decrease of .184. The .070 difference was not statistically significant (*t*[25]=0.71, *p*=.48).

4. Discussion

The current study sought to demonstrate the statistical equivalence of discounting (delay and probability) of real and hypothetical monetary rewards. In addition, this study examined the one-week test-retest reliability of discounting using indices of absolute and relative reliability. First, findings support statistical equivalence of DD of real and hypothetical rewards. Likewise, findings indicate that PD of real rewards is statistically equivalent to PD of hypothetical rewards. By using the more compelling threshold of statistical equivalence, the current analyses allow us to conclude that, not only is discounting of real and hypothetical rewards not significantly different, but that it is equivalent. This is an important distinction, since the current results provide perhaps the most convincing evidence of the acceptability of using hypothetical rewards in DD and PD research. While secondary, the high correlation of discounting of real and hypothetical rewards also serves to increase confidence in the acceptability of using hypothetical rewards. Insofar as hypothetical reward procedures are less costly, easier to administer, and allow researchers to consider extended delays and large sums of money, the established equivalence of real and hypothetical rewards has noteworthy practical implications.

With regard to the test-retest reliability of discounting, findings were less consistent, varying as a function of type of discounting, reward, and reliability analysis. In keeping with previous research, (Johnson and Bickel, 2002; Baker et al., 2003; Johnson et al., 2007), the present findings support both absolute and relative reliability of DD of real rewards. In other words, discounting of real, delayed rewards was equivalent when assessed 1 week apart (i.e., absolute reliability), and the rank order of each participant in the group was highly consistent over that time span (i.e., relative reliability). In contrast, DD of hypothetical rewards was not statistically equivalent when measurements were taken 1 week apart, despite exhibiting relative stability; in fact, the two measurements differed significantly. Mixed evidence for absolute stability of DD of hypothetical rewards suggests that the overall levels of discounting may show short-term intra-individual fluctuations, suggesting the need for additional research to identify factors that promote (in)stability of DD. However, given the relative consistency of all other results of the current study, as well as that of the existing literature, we believe this aberrant finding could be a result of type-I error.

Next, the present findings extend knowledge of the test-retest reliability of PD. Results suggest absolute reliability of PD for both real and hypothetical rewards. However, in contrast to earlier studies, which reported good relative test-retest reliability of PD for real (Peters and Büchel, 2009) and hypothetical rewards (Ohmura et al., 2006), the present study found a nonsignificant correlation across sessions for PD of real rewards. Although low rank order correlations may be interpreted as evidence of poor test-retest reliability, and therefore call into question the psychometric soundness of this measure of PD, our post-hoc analyses suggest an alternative explanation.

In the first session, participants experienced different outcomes with probabilistic real rewards. Half of the participants were compensated based on a choice trial for which they selected a certain, smaller (<$50) reward, six participants were compensated based on a trial

in which they chose the probabilistic reward but did not win anything, and the remaining eight participants were compensated based on a trial for which they chose a probabilistic reward and won $50. Our findings indicate that experiencing a certain outcome on the first probabilistic trial was associated with slightly greater risk aversion in the next session. In contrast, whether or not they won, participants who had direct experience with probabilistic rewards in session 1 showed less PD in the second session. This effect was modest in the group that did not win the probabilistic reward, and dramatic in the group that won the probabilistic reward, suggesting that favorable experience with probabilistic outcomes may be associated with greater risk seeking. Although this analysis is qualified by the small sample, findings suggest that PD procedures that use real reward outcomes have the potential for non-independence of observations over multiple assessments.

A potential caveat of these findings is that the study employed a potentially-real, or lottery-style, real rewards condition, in which participants' compensation was determined by a randomly selected trial from the real rewards condition. Other types of real reward paradigms, for instance, those in which the participant experiences the contingencies of every single trial of the choice procedure (Lagorio & Madden, 2005) or experiences choice contingencies in real time (Reynolds, 2006) may not evidence the same patterns of statistical equivalence with respect to reward type or consistency over time. Despite this limitation, the current study provides a rigorous test of the empirical equivalence of procedural variations in discounting research, and extends knowledge of the test-retest reliability of discounting measures for real and hypothetical monetary rewards. Findings serve to enhance confidence in the reliability and validity of delay and probability discounting assessments.

Acknowledgments

Data were collected at the University of Arkansas for Medical Sciences. The National Institute on Drug Abuse Research Grants R01DA11692, R03DA021707 and T32DA28855 supported this work. RDL received support from the National Center for Advancing Translational Science under grant UL1TR000039.

References

Audrain-McGovern J, Rodriguez D, Epstein LH, Cuevas J, Rodgers K, Wileyto EP. Does delay discounting play an etiological role in smoking or is it a consequence of smoking? Drug Alcohol Depen. 2009; 103:99–106.

Baker F, Johnson MW, Bickel WK. Delay discounting in current and never-before cigarette smokers Similarities and differences across commodity, sign and magnitude. J. Abnorm. Psychol. 2003; 112:382–392. [PubMed: 12943017]

Baumgartner, TA. Norm-referenced measurement: reliability. In: Safrit, MJ.; Woods, TM., editors. Measurement concepts in physical education and exercise science. Champaig, IL: Human Kinetics; 1989. p. 45-72.

Beck RC, Triplett MF. Test-retest reliability of a group-administered paper-pencil measure of delay discounting. Exp. Clin. Psychopharmacol. 2009; 17:345–355. [PubMed: 19803634]

Bickel WK, Marsch LA. Toward a behavioral economic understanding of drug dependence: Delay discounting processes. Addiction. 2001; 96:73–86. [PubMed: 11177521]

Bickel WK, Pitcock JA, Yi R, Angtuaco EJ. Congruence of BOLD response across intertemporal choice conditions: Fictive and real money gains and losses. J. Neurosci. 2009; 29:8839–8846. [PubMed: 19587291]

Bickel WK, Jarmolowicz DP, Mueller ET, Koffarnus MN, Gatchalin KM. Excessive discounting of delayed reinforcers as a trans-disease process contributing to addiction and other disease-related vulnerabilities: Emerging evidence. Pharmacol. Ther. 2012; 134:287–297. [PubMed: 22387232]

Bickel WK, Odum AL, Madden GJ. Impulsivity and cigarette smoking: delay discounting in current, never, and ex-smokers. Psychopharmacology. 1999; 146:447–454. [PubMed: 10550495]

Blackwelder WC. Proving the null hypothesis in clinical trials. Control. Clin. Trials. 1982; 3:345–353. [PubMed: 7160191]

Chaplin WT, John OP, Goldberg LR. Conceptions of states and traits. J. Pers. Soc. Psychol. 1988; 54:541–557. [PubMed: 3367279]

Davis C, Patte K, Curtis C, Reid C. Immediate pleasures and future consequences. A neuropsychological study of binge eating and obesity. Appetite. 1997; 54:208–213. [PubMed: 19896515]

Food and Drug Administration, Center for Drug Evaluation and Research (CDER). Guidance for Industry: Statistical Approaches to Establishing Bioequivalence, BP. 2001.

Gallistel CR. The importance of proving the null. Psychol Rev. 2009; 116:439–453. [PubMed: 19348549]

Hinvest NS, Anderson IM. The effects of real versus hypothetical reward on delay and probability discounting. Q. J. Exp. Psychol. A. 2010; 63:1072–1084.

Jikko Y, Okouchi H. Real and hypothetical rewards in probability discounting. Jpn. J. Psychol. 2007; 78:269–276.

Jimura K, Myerson J, Hilgard J, Keighley J, Braver TS. Domain independence and stability in young and older adults' discounting of delayed rewards. Behav. Processes. 2011; 87:253–259. [PubMed: 21550384]

Johnson MW. An efficient operant choice procedure for assessing delay discounting in humans: Initial validation in cocaine-dependent and control individuals. Exp. Clin. Psychopharmacol. 2012; 20:191–204. [PubMed: 22329554]

Johnson MW, Bickel WK. Within-subject comparison of real and hypothetical money rewards in delay discounting. J. Exp. Anal. Behav. 2002; 77:129–146. [PubMed: 11936247]

Johnson MW, Bickel WK, Baker F. Moderate drug use and delay discounting: A comparison of heavy, light and never smokers. Exp. Clin. Psychopharmacol. 2007; 15:187–194. [PubMed: 17469942]

Kirby KN. One-year temporal stability of delay discount rates. Psychon. Bull. R. 2009; 16:457–462.

Kraemer HC, Glick ID, Klein DF. Clinical trials design lessons from the CATIE study. Am. J. Psychiatry. 2009; 166:1222–1228. [PubMed: 19797435]

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

Lagorio CH, Madden GJ. Delay discounting of real and hypothetical rewards III: Steady-state assessments, forced-choice trials, and all real rewards. Behav. Processes. 2005; 31:173–187. [PubMed: 15845306]

Landes RD, Christensen DR, Bickel WK. Delay discounting decreases in those completing treatment for opioid dependence. Exp. Clin. Psychopharmacol. 2012; 20:302–309. [PubMed: 22369670]

Lane SD, Cherek DR, Rhoades HM, Pietras CJ, Tcheremissine OV. Relationships among laboratory and psychometric measures of impulsivity: Implications in substance abuse and dependence. Addict. Disord. Their Treat. 2003; 2:33–40.

Lawyer SR, Schoepflin F, Green R, Jenks C. Discounting of hypothetical and potentially real outcomes in nicotine-dependent and nondependent samples. Exp. Clin. Psychopharmacol. 2011; 11:263–274. [PubMed: 21707190]

Luzvar-Stiffler V, Stiffler C. Equivalence testing the easy way. J. Comput. Inform. Tech. 2002; 10:233–239.

Madden GJ, Begotka AM, Raiff BR, Kastern LL. Delay discounting of real and hypothetical rewards. Exp. Clin. Psychopharmacol. 2003; 11:139–145. [PubMed: 12755458]

Madden GJ, Raiff BR, Lagorio CH, Begotka AM, Mueller AM, Helhi DJ, Wegener AA. Delay discounting of potentially real and hypothetical rewards: II. Between-and within-subject comparisons. Exp. Clin. Psychopharmacol. 2004; 12:251–261. [PubMed: 15571442]

Madden GJ, Petry NM, Johnson PS. Pathological gamblers discount probabilistic rewards less steeply than matched controls. Exp. Clin. Psychopharmacol. 2009; 17:283–290. [PubMed: 19803627]

Manwaring JL, Green L, Myerson J, Strube MJ, Wilfley DE. Discounting of various types of rewards by women with and without binge eating disorder: Evidence for general rather than specific differences. Psychol. Rec. 2011; 61:561–582. [PubMed: 24039301]

Melanko S, Larkin KT. Preference for immediate reinforcement over delayed reinforcement: relation between delay discounting and health behavior. J. Behav. Med. 2012; 160:1–10.

Green L, Myerson J. A discounting framework for choice with delayed and probabilistic rewards. Psychol. Bull. 2004; 130:769–792. [PubMed: 15367080]

Odum AL. Delay discounting: Trait variable? Behav. Processes. 2011; 87:1–9. [PubMed: 21385637]

Ohmura Y, Takahashi T, Kitamura N, Wehr P. Three-month stability of delay and probability discounting measures. Exp. Clin. Psychopharmacol. 2006; 14:318–328. [PubMed: 16893275]

Peters J, Büchel C. Overlapping and distinct neural systems code for subjective value during intertemporal and risky decision making. J. Neurosci. 2009; 29:15727–15734. [PubMed: 20016088]

Petry NM. Pathological gamblers, with and without substance use disorders, discount delayed rewards at high rates. J. Abnorm. Psychol. 2001; 110:482–487. [PubMed: 11502091]

Petry NM. Discounting of probabilistic rewards is associated with gambling abstinence in treatment-seeking pathological gamblers. J. Abnorm. Psychol. 2012; 121:151–159. [PubMed: 21842965]

Rachlin H, Brown J, Cross D. Discounting in judgments of delay and probability. J. Behav. Decis. Making. 2000; 13:145–160.

Rachlin H, Raineri A, Cross D. Subjective delay and probability. J. Exp. Anal. Behav. 1991; 55:233–244. [PubMed: 2037827]

Reynolds B, Richards J, Horn K, Karraker K. Delay and probability discounting as related to cigarette smoking status in adults. Behav. Processes. 2004; 65:35–42. [PubMed: 14744545]

Reynolds B. A review of delay-discounting research humans: Relations to drug use and gambling. Behav. Pharmacol. 2006; 17:651–667. [PubMed: 17110792]

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon B Rev. 2009; 16:225–237.

Scheres A, Dijkstra M, Ainslie E, Balkan J, Reynolds B, Sonuga-Barke E, Castellanos FX. Temporal and probabilistic discounting of rewards in children and adolescents: effects of age and ADHD symptoms. Neuropsychologia. 2006; 44:2092–2103. [PubMed: 16303152]

Sheffer C, MacKillop J, McGeary J, Landes R, Carter L, Yi R, Jones B, Christensen D, Stitzer M, Jackson L, Bickel WK. Delay discounting, locus of control and cognitive impulsiveness independently predict tobacco dependence treatment outcomes in a highly dependent, lower socioeconomic group of smokers. Am. J. Addict. 2012; 21:221–232. [PubMed: 22494224]

Simpson CA, Vuchinich RE. Reliability of a measure of temporal discounting. Psychol. Rec. 2000; 50:3–16.

Tryon WW. Evaluating statistical difference, equivalence and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. Psychol. Meth. 2001; 6:371–386.

Wagenmakers EJ. A practical solution to the pervasive problems of p values. Psychon B Rev. 2007; 14:779–804.

Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Cond. Res. 2005; 19:231–240. [PubMed: 15705040]

Welleck, S. Testing statistical hypotheses of equivalence and noninferiority. second ed. Boca Raton: Chapman and Hall; 2010.

Weller R, Cook E III, Avsar K, Cox J. Obese women show greater delay discounting than healthy-weight women. Appetite. 2008; 51:563–569. [PubMed: 18513828]

Yi R, Landes RD. Temporal and probability discounting by cigarette smokers following acute smoking abstinence. Nicotine Tob. Res. 2012; 14:547–558. [PubMed: 22311959]

Yi R, Chase WD, Bickel WK. Probability discounting among cigarette smokers and nonsmokers: molecular analysis discerns group differences. Behav. Pharmacol. 2007; 18:633–640. [PubMed: 17912047]

Yi, R.; Mitchell, SH.; Bickel, WK. Delay discounting and substance abuse-dependence. In: Madden, GM.; Bickel, WK., editors. Impulsivity: The behavioral and neurological science of discounting. Washington, DC: American Psychological Association; 2010. p. 191-211.

Highlights

- There is evidence for the statistical equivalence of discounting of real and hypothetical rewards.
- Results support absolute and relative test-retest reliability of delay discounting over one week.
- Probability discounting showed absolute but not relative test-retest reliability.
- Findings serve to enhance confidence in discounting assessment procedures.

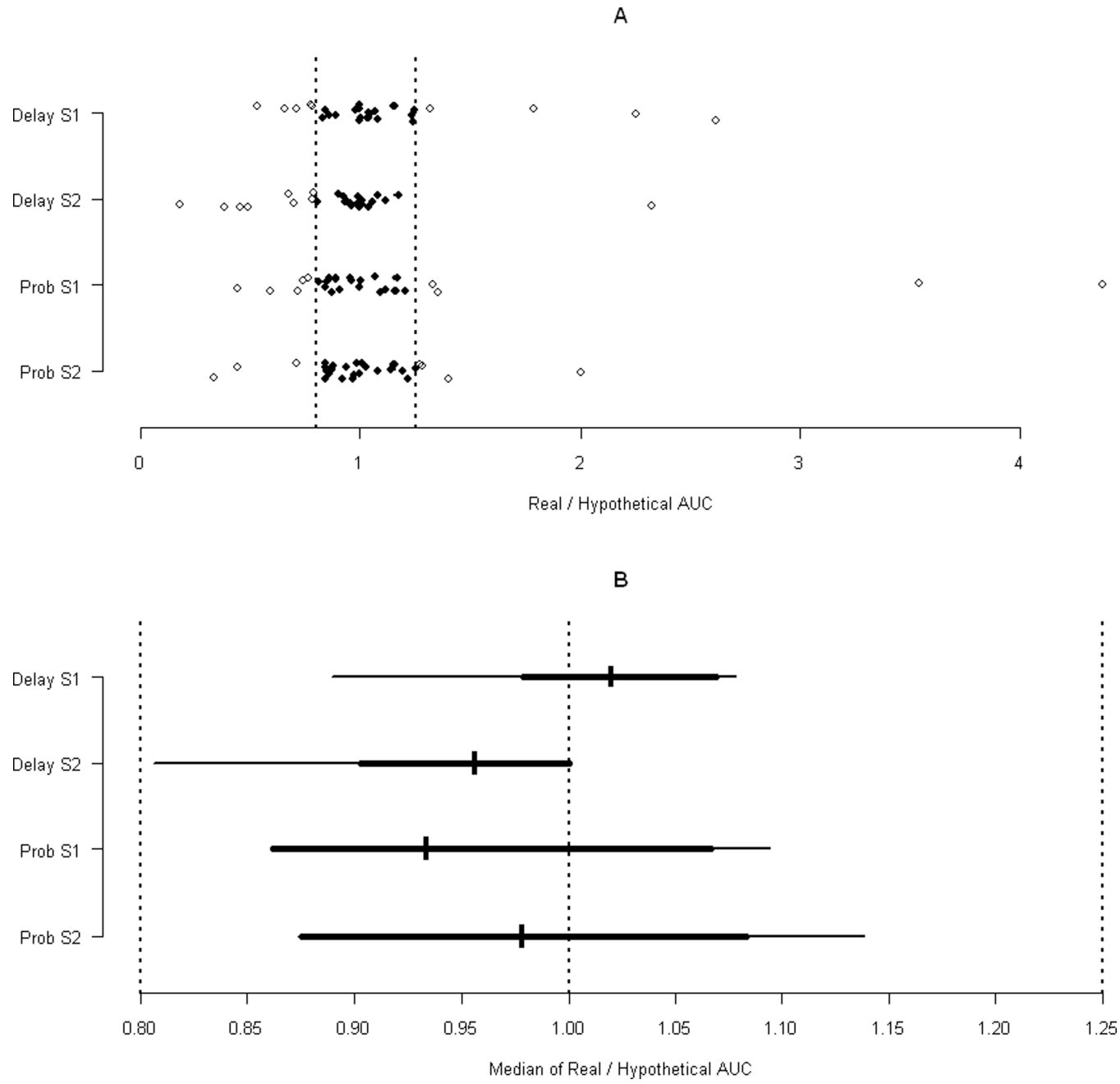

Figure 1. Equivalence of discounting of real and hypothetical rewards as a function of type of discounting and session

(A) Observed discounting expressed as a ratio of real/hypothetical rewards, jittered vertically to better distinguish individual points. S1 and S2 represent session 1 and session 2, respectively. The vertical dashed lines represent the equivalence region (0.8, 1.25). Filled points fall within the equivalence region, and open points do not.

(B) Observed discounting expressed as a ratio of real/hypothetical rewards. S1 and S2 represent session 1 and session 2, respectively. Each line is marked at the the median ratio of real/hypothetical rewards. Thick bars represent the 90% confidence intervals for the median ratios, and thin bars represent the 95% confidence intervals for the median ratios. The

nonparametric nature of the 90% and 95% confidence intervals is such that an upper *or* lower bound can be common to both. Two measurements are considered statistically equivalent if (*i*) the endpoints of the thick bar fall entirely within the equivalence region, (0.8, 1.25), and (*ii*) the thin bar covers 1.0. If the thin bar does not cross 1.0, the two measures are statistically different, and therefore cannot be statistically equivalent.

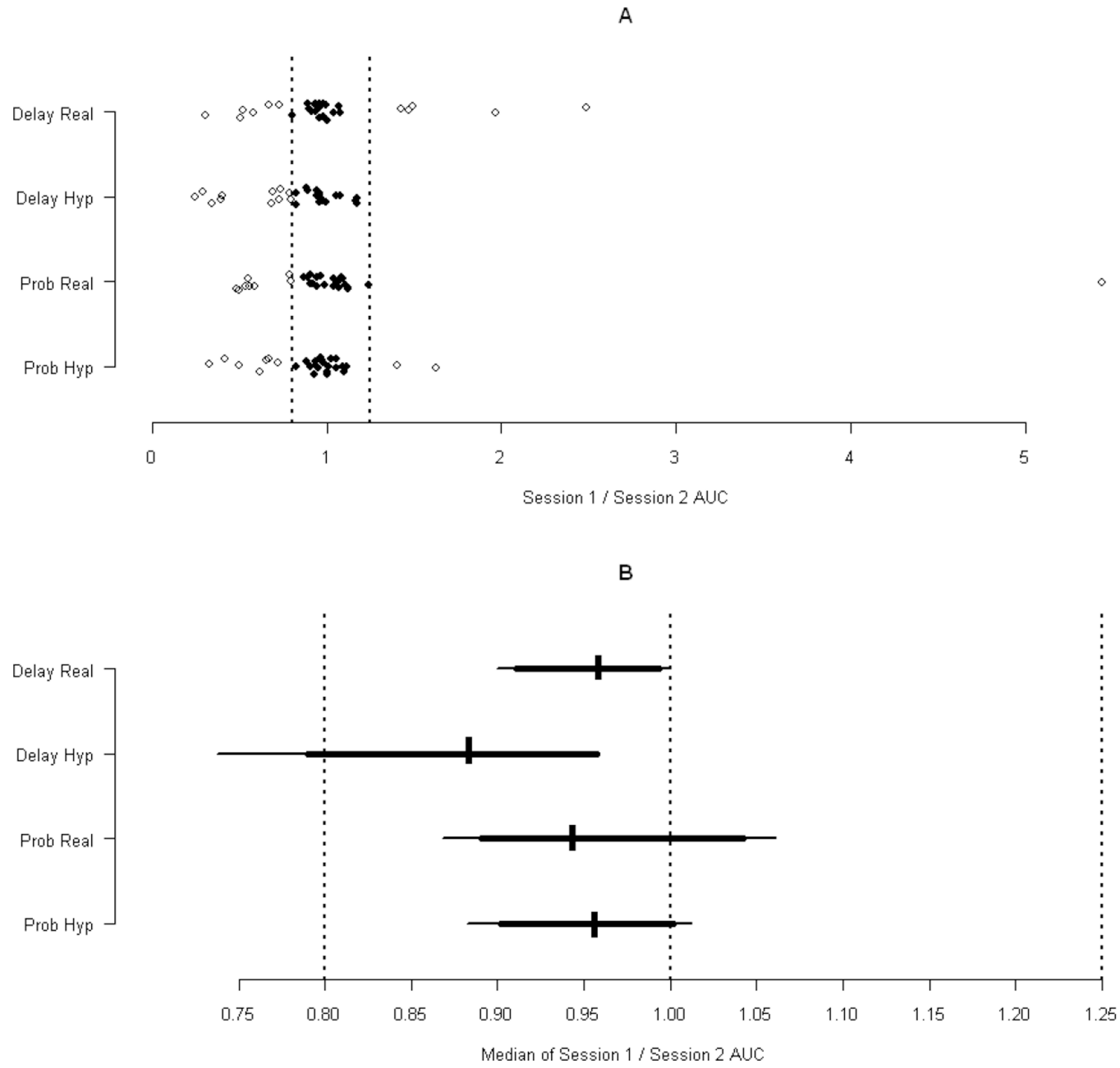

Figure 2. Equivalence of two measures of discounting taken one week apart, as a function of type of discounting and reward type (real, hypothetical [hyp])

(A) Observed discounting expressed as a ratio of session 1/session 2, jittered vertically to better distinguish individual points. The vertical dashed lines represent the equivalence region (0.8, 1.25). Filled points fall within the predeterimined equivalence region (0.8, 1.25); open do not.

(B) Observed discounting expressed as a ratio of session 1/session 2. Each line is marked at the median ratio of session 1/session 2 discounting. Thick bars represent the 90% confidence intervals for the median ratios, and thin bars represent the 95% confidence intervals for the median ratios. The nonparametric nature of the 90% and 95% confidence intervals is such

that an upper *or* lower bound can be common to both. Two measurements are considered statistically equivalent if (*i*) the endpoints of the thick bar fall entirely within the equivalence region, (0.8, 1.25), and (*ii*) the thin bar covers 1.0. If the thin bar does not cross 1.0, the two measures are statistically different, and therefore cannot be statistically equivalent.

Table 1

Spearman correlation coefficients for delay and probability discounting (AUC) of real and hypothetical rewards.

		Session 1		Session 2	
		Real	**Hypothetical**	**Real**	**Hypothetical**
Session 1	Real		.52*	.34	
	Hypothetical	.86*			.51*
Session 2	Real	.73*			.66*
	Hypothetical		.70*	.80*	

Note. Values below the diagonal are correlations for delay discounting, while values above the diagonal are correlations for probability discounting.

* $p < .05$