An Impulsive Dissertation: Experimental and Behavioral Validity for a New Measure of Trait Impulsivity

By

Aaron Kirk Wirick

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

Washington State University
Department of Psychology

May 2009

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of Aaron Kirk Wirick find it satisfactory and recommend that it be accepted.

John Hinson, Ph. D., Chair

Paul M. Whitney, Ph. D.

Craig D. Parks, Ph. D.

Maureen Schmitter-Edgecombe, Ph. D.

Acknowledgment

This dissertation would have been impossible without the help of a number of important individuals. First of all, I'd like to thank my advisors, Drs. John Hinson and Paul Whitney, for five years of excellent guidance and mentorship. They not only kept a close eye on me when necessary but also gave me the freedom to pursue new research areas and methods. I'm extremely grateful for their efforts.

Secondly, I received a great deal of assistance from Drs. Maureen Schmitter-Edgecombe and Craig Parks who served on my dissertation committee. Maureen always helped me to see my research from an outside perspective and Craig made sure that my methods were sound. I'd also like to acknowledge the very helpful advice of Dr. Len Burns who helped develop my skills as a psychometrician and served as a sounding board for many of the ideas that eventually led to this dissertation.

Third, I am grateful to my fellow graduate student, Allison Matthews, for her willingness to provide a second opinion on many research decisions and an extra pair of eyes in the revision process. I also want to thank the undergraduate research assistants who collected the lion's share of the data that made this work possible. These hard-working individuals include: Alison Gothro, Michelle Harn, Fritz Schoepflin, Kevin Feiszli, Dianna Barker, Corissa White, Marilis Earle, Tricia Christensen, Josh Christensen, Vernon Rasiah, Danielle Lamoy, Casey Jobin and Jessie Aguilar.

Last but certainly not least, I would have never finished this work without the encouragement and help of my beautiful wife, Dina. Always willing to read a draft, brew some coffee or look over my work, her help was as invaluable as it was unceasing.

An Impulsive Dissertation: Experimental and Behavioral Validity for a New Measure of Trait Impulsivity

Abstract

by Aaron Kirk Wirick, Ph. D.
Washington State University
May 2009

Chair: John Hinson

Two studies evaluated the validity of a new measure of impulsivity titled the Risk Seeking and Response Inhibition Scales. This measure was derived from a two-factor theory of impulsivity and was designed to replace the questionnaires currently used in the impulsivity literature. The first study demonstrated good concurrent and discriminant validity with existing measures of impulsivity. Furthermore, marginal external validity was shown through correlations with measures of compulsive buying, alcohol problems and gambling problems. In the second study, structural models were constructed to measure the new scales' ability to predict two experimental decision making tasks, the Iowa Gambling Task and Balloon Analogue Risk Task. These models demonstrated some validity, but also highlighted some of the weaknesses still present in the new measure. Further refinement is needed before the subscales can adequately replace currently used measures. However, these studies demonstrate that a replacement is still very much needed.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Many argue that the validity of a psychological theory can be summarized by the extent to which it accurately predicts behavior. In the case of decision making, one approach to addressing the validity of a theory is to predict failures in an individual's decisions. These deficits can range from temporary lapses in judgment to enduring patterns of poor decision making. In the latter context, it invokes a personality trait of decision making that has been labeled impulsivity.

Impulsivity provides a conduit for relating laboratory results to real world decision making. As a personality trait, impulsivity can be studied in broader contexts than experimental measures and as a result has been reliably linked to many different types of problematic real world behaviors including substance abuse, gambling, financial loss and other maladaptive decisions. From this, one can presumably infer that impulsivity is a valid analogue for measuring real world problems in a lab setting. However, many theorists note that there is often a weak or nonexistent correlation between experimental results and the trait of impulsivity (Reynolds, Richards, & de Wit, 2006). Further complicating the problem is that experi-

mental tasks also correlate poorly with real world difficulties. While these complications do not imply that the experimental measures are invalid, they provide a sizable obstacle for anyone wishing to validate their results. I argue that the reason for this difficulty is not that the experimental tasks are unrelated to their personality analogues, but rather that the presence of excessive error variance in many personality measures precludes researchers from establishing a relationship. The goal of the present dissertation was to remedy that situation by the validation of a new measure of impulsivity and the use of more appropriate statistical methods to relate that measure with two experimental tasks.

In order to accomplish this goal, I begin by outlining our current understanding of impulsivity as a personality trait, including a discussion of how impulsivity can be looked at from an experimental approach. Next, I provide a brief review of the most widely used measures of impulsivity and highlight the problems that are leading to unreliable measurement. Finally, I discuss the development of the new measure of impulsivity and the methodologies employed to validate the measure in this dissertation.

1.2 Theoretical Background

Impulsivity is a multidimensional trait that varies normally in the adult population. Although many different factor structures have been suggested through the years, current theories concur that a two factor structure is preferred (Swann, Bjork, Moeller, & Dougherty, 2002; Evenden, 1999; Dawe, Gullo, & Loxton, 2004). Interestingly, these factors are the combination of two distinct approaches to understanding impulsivity.

First, from social psychology, there is impulsivity related to reward or novelty seeking (Zuckerman, 1994; Roberti, 2004). This facet of impulsivity is responsible for the attraction of individuals towards risky or unsafe decisions. For consistency, I will refer to this subfactor of impulsivity as *reward seeking*. It can be thought of as a motivational factor that draws

individuals towards choices with the lure of reward in the presence of negative outcomes.

Reward seeking is described by many theorists more precisely as a hypersensitivity to reward (Bechara, Dolan, & Hindes, 2002; Smillie & Jackson, 2006). For instance, in individuals with alcohol problems, reward seeking seems to account for the initial attraction to alcohol use (Dawe & Loxton, 2004). Alcohol appeals to these individuals as a means of experiencing euphoria and other positive emotions. Reward seeking maintains alcohol misuse by keeping the possible benefits of alcohol in the forefront of an individual's mind and masking the potentially serious drawbacks of excessive use.

Reward seeking can manifest itself in many different ways but can be loosely described as an attraction towards risky behaviors and thrilling experiences (e.g., skydiving, playing the stock market). Categorizing reward seeking in this way focuses on the negative aspect of this behavior (i.e., risk implies danger or loss). Furthermore, by leaving risk open-ended (i.e., referring generally to risk rather than specifically to skydiving, etc.), the individual can apply it to a variety of personally relevant activities. It should be noted that there are likely other manifestations of reward seeking. But if we are interested in studying impulsivity as a negative decision we can generalize reward seeking as a tendency towards risky behaviors.

The second factor of impulsivity comes from the neuropsychological and cognitive neuroscience literature (Franken, van Strien, Nijs, & Muris, 2008; Fellows & Farah, 2005; Williams & Taylor, 2006). This aspect of impulsivity relates to a lack of inhibitory control or response inhibition. In this case, impulsivity is not defined by a motivational drive but rather by a lack of cognitive control. For instance, individuals may intuitively know that a decision has led to problems in the past, but they are unable to stop themselves from making the same faulty choice again. Many theorists suggest that this aspect of impulsivity is related to deficits in executive function and is therefore rooted in working memory (Hinson, Jameson, & Whitney, 2003; Whitney, Jameson, & Hinson, 2004).

Compulsive buying provides a relevant demonstration of response inhibition problems.

Individuals with normal or high levels of response inhibition recognize when they cannot afford to buy something. These individuals therefore exercise self control in their shopping and avoid the consequences of making an impulsive purchase. Individuals with abnormal response inhibition may be so quick to act that they do not take the time to think about the negative consequences of excessive spending. Alternatively, they may simply lack the mental control to stop themselves from acting even if they *know* they cannot afford a purchase.

It's often useful to frame response inhibition as a working memory function. Inhibitory control problems often manifest themselves as an inability to maintain attention. This provides an important route for measuring response inhibition for two reasons. First, it is unlikely that individuals will endorse that they lack self control (or even that they are aware of it), but many individuals will be willing to endorse that they are easily distracted. Secondly, this frames response inhibition problems in terms of well understood working memory processes. In particular, the focus on attention problems allows us to understand response inhibition as a more general problem in executive control which has been heavily studied in experimental decision making paradigms.

1.3 Current Measures

Personality psychology typically employs one of three scales when measuring impulsivity:

- Barratt's Impulsivity Scale (BIS; Patton, Stanford, & Barratt, 1995)
- Zuckerman's Sensation Seeking Scale (SSS; Zuckerman, 1994)
- Eysenck's Impulsivity Scale (EIS; Eysenck, Pearson, Easting, & Allsopp, 1985)

These measures possess unique strengths and weaknesses. More recently, a new scale was developed that coalesced these three separate measures and others into a unified scale by using a large data-driven exploratory factor analysis. This scale, the Urgency Premeditation

Perseverance and Sensation Seeking Impulsive Behavior Scale (UPPS) will be discussed here as well (Whiteside & Lynam, 2001).

The first and most commonly cited measure, the BIS, is a 34-item questionnaire. Google scholar currently lists more than 450 citations for the most recent revision of the BIS (Patton, Stanford, & Barratt, 1995). From a two-factor impulsivity standpoint, the BIS singlehandedly defined the measurement of response inhibition problems. It has been used extensively in neuropsychological research and is noteworthy for providing a way of relating personality and experimental approaches. Furthermore, the BIS is important for connecting response inhibition problems to cognitive control mechanisms rooted in working memory and specifically executive control. For these reasons, many people in cognitive psychology and the neurosciences see the BIS as a very important tool for relating experimentally derived theories to practical real world problems.

Unfortunately, the BIS contains some serious flaws that have not been addressed. First and foremost among these is an unstable factor structure. In the 1994 article, the BIS was factor analyzed using principal components and varimax rotation. The resulting factor solution revealed three dimensions of impulsivity: Non-Planning, Attention and Motor. Subsequent factor analyses have been rare (Miller, Joseph, & Tudway, 2004; Someya, Sakado, Seki, Kojima, Reist, Tang, & Takahashi, 2001; Spinella, 2007) and weakly support the factor structure of the BIS. Unpublished data on confirmatory factor analyses in a sample of 613 Washington State University undergraduates demonstrates a very unstable factor structure in which the overall model fit is exceedingly poor (CFI = 0.6) unless large amounts of correlated errors and cross-loadings are introduced. Furthermore, the BIS utilizes a four category Likert-type scale. With so few options, interval level assumptions are violated and techniques such as factor analysis are problematic at best (Flora & Curran, 2004). Therefore, these factor analytic results should be interpreted with caution. These shortcomings should not be taken as a criticism of the original authors intents but rather as a reminder that

psychometric theory has come a long way since 1994, therefore an update is needed.

Zuckerman's work on the SSS has been extremely influential, especially in social psychology research. The SSS contains 40 items which measure reward seeking impulsivity. The SSS contains four factors including thrill-and-adventure seeking, boredom susceptibility, experience seeking and disinhibition. Responses on the SSS are elicited using a two-option forced choice format. Participants are given two seemingly diametrical options and indicate their preference by circling the one they most agree with. Final scores on the SSS are calculated by summing the frequency of responses that indicate sensation seeking (Zuckerman, 1994).

One of the major advantages of the SSS is that it intentionally relates the expression of personality to biological bases (Zuckerman & Kuhlman, 2000). This is noteworthy because the SSS does not merely describe personality but helps explain where personality comes from. Increasingly, cognitive psychology is attempting to relate mental phenomena back to brain function and the SSS has been doing so for some time.

Like the BIS, the SSS is also beginning to show its age and contains a number of issues that need to be addressed. Gray and Wilson (2007) conducted a factor analysis of the SSS after updating the scale to a Likert-type response format. Their analysis suggest that the reliability of the SSS is questionable, that some items seriously threaten construct validity and that the wording of the scale is somewhat out of date (e.g., references to the jet set, swingers, etc.). Zuckerman (2007), in response to these criticisms, argues that the validity of the scale is in its extensive use and its established relationship with risky behaviors.

Zuckerman (2002) recommended his own modifications to the SSS and these changes were included in the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ). Although this questionnaire is designed to measure personality from a larger five-factor perspective, it contains a subscale that assesses both sensation seeking and impulsivity. The sensation seeking dimension evolved directly out of the SSS while the impulsivity dimension captures response inhibition type behaviors. The addition of the impulsivity dimension was theoretically nec-

essary but was not particularly successful. A recent evaluation of the ZKPQ demonstrated that the items on the impulsivity scale (i.e., response inhibition) fit poorly into the overall model (Aluja, Rossier, García, Angleitner, Kuhlman, & Zuckerman, 2006). For these reasons, both the SSS and its descendant, the ZKPQ, should be used with hesitation when examining the reward seeking facet of impulsivity.

The last commonly utilized measure of trait impulsivity is the EIS. Now in a seventh revision, this scale contains 54 yes-no items grouped into three factors: Impulsivity, Venturesomeness and Empathy. Impulsivity maps on directly to response inhibition and venturesomeness relates to reward seeking behavior. The empathy factor has been largely abandoned in recent years due to reliability concerns and a lack of construct validity. It has therefore been excluded from the present study.

One of the major advantages of the EIS is that it has been measuring impulsivity from a two factor standpoint for 20 plus years. While both the BIS and SSS are more narrowly focused on the subcomponents of response inhibition and reward seeking, the EIS has drawn both components together into a single assessment. Thus, studies that have utilized the EIS have helped to demonstrate evidence for a two-factor approach to impulsivity. Like the SSS, the EIS is based on the assumption that personality traits are derived from biological processes. Furthermore, Eysenck's personality research was among the first theories to extend across social and cultural groups (Barrett, Petrides, Eysenck, & Eysenck, 1998).

In spite of its definite value, the EIS has not been refined in recent years and has some major psychometric issues. Like the other scales, it utilizes a categorical response format to indicate either agreement or disagreement with a prompt (e.g., Do you quite enjoy taking risks?). Scores are summed so that the total score reflects the number of agreements. Because the scale was revised 20 years ago, items are becoming outdated (e.g., Are you happy when you are with a cheerful group and sad when the others are glum?). Recently, exploratory factor analyses were conducted on the EIS and the general factor structure was maintained

(Caci, Nadalet, Baylé, Robert, & Boyer, 2003). However, these results should be interpreted with extreme caution as the authors did not address the issue of categorical variables in their analysis.

Whiteside and Lyman (2001) recognized the lack of a unified measure of impulsivity and sought to modernize impulsivity assessment by combining the measures into one questionnaire. The result of this endeavor is the UPPS, a four factor impulsivity measure containing 45 items. A number of methodological problems during scale development may ultimately limit the utility of the UPPS. First, the development of the scale was based entirely on an exploratory factor analysis of 20 subscales from nine separate impulsivity measures. The only apparent criterion for scale inclusion was that it come from some theory of impulsivity and not necessarily an accepted theory. This atheoretical approach is problematic because the resulting scale will be completely determined by the data and all conclusions pertaining to what the scale ought to measure are decided by chance correlation rather than careful construction. Without a strong theoretical orientation to drive the scale construction, the resulting factor structure can be determined by spuriously arising factors that do not relate to the intended construct of impulsivity. In this context, the urgency and lack of perseverance factors are not well accounted for by the impulsivity literature. This is primarily because the authors utilized subscales from the NEO-PI-R and EAS-III which emerged as dominant factors. Neither scale has empirical support as an impulsivity measures.

A second major criticism is that the resulting factor structure failed to fit adequately when subjected to a confirmatory factor analysis (Magid & Colder, 2007). This likely occurred because the original exploratory factor analysis was actually principal components rather than a true common factor model. Principal components forces the factors to account for all of the variance in the model including measurement error. Thus, when transitioning to a common factor model as in confirmatory factor analysis, the loadings between items and factors can weaken when error variance is accounted for. Finally, the UPPS has only marginal

validity. Whiteside and Lyman (2005) demonstrated that their scale can predict antisocial personality disorder and borderline personality disorder traits, but it failed to correlate with either alcohol problems or gambling behavior. Taken as a whole, these criticisms suggest that the UPPS is neither a reliable nor valid alternative to the older measures of impulsivity which it intended to replace.

I want to reemphasize that the EIS, SSS and BIS have made unique and valuable contributions to our understanding of impulsivity and decision making. However, the point of my criticism is to illustrate that meaningful improvements are past due and that development of a new measure will result in many of the advantages of these scales without their known psychometric limitations.

1.4 Risk Seeking and Response Inihibition Scales

I set out to develop a new scale for assessing impulsivity that addressed the theoretical and statistical limitations of current measures. The result of this work was a two-factor measure of impulsivity named the Risk Seeking and Response Inhibition Scale (RSRIS). The RSRIS was developed using item response theory (IRT). This approach has a number of advantages including greater reliability at the item level, stronger dimensionality for factors and a more accurate characterization of measurement error for individuals (Embretson & Reise, 2000, see also Appendix A). A number of important considerations were made during the development phase. First, I generated a number of items that dealt with the two factors of impulsivity at a theoretical level. Good psychometric measures are theoretically driven and I wanted my new measure to represent impulsivity from the perspective of response inhibition and reward seeking. For the reward seeking elements, I was primarily concerned with items that dealt with endorsing an interest in risky activities, financial problems and an aversion to living a boring life. With response inhibition, the intent was to frame impulsivity in terms

of executive control difficulties, such as attention problems, long term planning problems, a dislike of waiting, etc. For each factor, 34 items that represent a broad range of impulsive statements were generated.

A second objective of the new scale was to represent impulsivity as two factors with strict unidimensionality between each factor. Unidimensionality is important for measurement in order to accurately distinguish between similar constructs. For example, the authors of the BIS acknowledge three separate subscales that are unique but correlate strongly. In many experiments, the subscales themselves are ignored and only the total scores are examined. This is cause for concern because an individual's total score results from the summing of theoretically diverse items associated with what should be conceptually different constructs. In other words, it is possible for two individuals to get the same total score but endorse completely different behaviors. Alternatively, strictly unidimensional scales will not correlate with each other and are explicitly interpreted as unique constructs. An individual's score can be compared directly with another individual's score without uncertainty about why their scores are different or similar.

In addition to fitting a two-factor structure, I was also interested in the style used to frame the items in the questionnaire. Since the BIS is the most widely utilized measure of impulsivity, I adopted its style when generating items to provide face validity and a format that would be familiar to current researchers. Questions are thus termed in self-reflective statements that an individual can either agree with or disagree with (e.g., I dislike waiting, or I want to live an adventurous life). Due to the problems with a limited number of response options in the BIS, I chose a seven point Likert-type scaling, anchored from strongly agree to strongly disagree. This provides enough points so that interval level analyses can be utilized if other researchers so desired.

The last major consideration when developing the scale was that the end product should be relatively brief so that it can be easily administered as part of larger experimental studies.

From an IRT psychometric framework, shorter scales are preferred and justified by strong reliability at the item level. This goes against the classical assumption that more is always better and has been shown to be a better approach to measurement (Embretson & Reise, 2000). Therefore, the scale should contain enough strong items to maintain a high amount of reliability but a concise scale is preferred to a marginally more reliable and longer scale.

This work resulted in a 14 item scale that contained two unidimensional factors with seven items each (see Appendix D for the final scale).

1.5 Validation

A new scale is valuable to the extent that it predicts both impulsive real world outcomes and impulsive behaviors in the lab. This dissertation attempted to validate the RSRIS by subjecting it to two separate tests of validity. First, the RSRIS was utilized to predict scores on three separate outcome measures, namely alcohol consequences, compulsive buying and gambling problems.

Alcohol problems can develop from a variety of circumstances, but it has been repeatedly demonstrated that impulsivity predicts alcohol problems (Simons & Carey, 2006). The relationship is likely mediated by motives for drinking and overall alcohol use, but for the purposes of validation I restricted analysis to a direct relationship (Martens, Neighbors, Lewis, Lee, Oster-Aaland, & Larimer, 2008). From a two-factor standpoint, reward seeking should be more strongly related to alcohol problems than response inhibition (Magid & Colder, 2007).

Compulsive buying behavior has also been reliably related to trait impulsivity and specifically response inhibition problems (Billieux, Rochat, Rebetez, & Van der Linden, 2008). Clinically compulsive buyers demonstrate high levels of impulsivity (Mueller, Mueller, Albert, Mertens, Silbermann, Mitchell, & de Zwaan, 2007) and impulsivity is positively related

to college students with credit card debt (Pirog & Roberts, 2007).

Lastly, gambling problems are predicted by higher levels of impulsivity (Slutske, Caspi, Moffitt, & Poulton, 2005; Nower, Derevensky, & Gupta, 2004). There is still some debate as to the relative value of each factor of impulsivity, specifically whether reward seeking can predict pathological gambling (Hammelstein, 2004; Blaszczynski, Steel, & Mcconaghy, 1997). On the other hand, response inhibition has been reliably related to gambling problems (Wohl, Matheson, Young, & Anisman, 2008; Slutske, Caspi, Moffitt, & Poulton, 2005; McDaniel & Zuckerman, 2003; Vitaro, Arseneault, & Tremblay, 1999). Therefore, I expected that the response inhibition measures would predict higher gambling problems but that a relationship between reward seeking and gambling may or may not emerge.

To examine the validity of the RSRIS, correlation matrices were constructed in which the total scores from the four impulsivity measures (RSRIS, BIS, EIS and SSS) were related to the three outcomes. This provides both predictive validity (i.e., correlations with the problem behaviors) and demonstrates concurrent validity (correlations between the predictors). Furthermore, a lack of correlation among different types of measures (e.g., risk-seeking vs response inhibition) provides evidence for discriminant validity.

A second form of validity is the ability to predict laboratory behaviors. This is particularly important as measures of impulsivity are often used to lend external validity to the results of experimental manipulations in cognitive studies. The second study examined how the RSRIS predicts behavior in two common decision making paradigms, the Iowa Gambling Task (IGT) and the Balloon Analogue Risk Task (BART). Structural equation models were constructed as a way to test relationships between the RSRIS and these two tasks.

The IGT is one of the most widely cited experimental measures of decision making (Bechara, Damasio, Damasio, & Anderson, 1994; Bechara, Tranel, Damasio, & Damasio, 1996; Bechara, Damasio, & Damasio, 2000). In this task, participants are asked to make decisions by selecting cards from the top of four hypothetical decks. Each card results in a

guaranteed win but periodically contains a loss. Two of the decks are considered good decks because they result in small steady wins over time. The other two decks are considered bad decks and result in large initial wins but even larger losses over time.

The IGT has been used extensively in neuropsychological studies and basic experimental research (Dunn, Dalgleish, & Lawrence, 2006). Previous studies have attempted to correlate performance with both facets of impulsivity with limited success. For instance, Petry utilized a variety of scales but was not able to establish significant correlations with IGT performance (Petry, 2001). However, it should be noted that she did not report raw correlations but correlations with factor scores. Others have demonstrated success in predicting IGT performance using response inhibition (Zermatten, der Linden, Jermann, & Bechara, 2005) and sensation seeking (Breslin, Sobell, Cappell, Vakili, & Poulos, 1999).

One of the problems unique to the IGT is that the scoring often does not reflect actual performance accurately. Task performance is typically quantified by averaging the number of good decks that have been chosen in a block of trials. This is an ill-defined way of characterizing performance because the good decks only become clear after the participant has been playing for a while. In order to get around this, I will use a more complex methodology for teasing apart IGT performance by using the estimates from a model of IGT performance (Busemeyer & Stout, 2002). This model, called the expectancy-valence learning model, estimates three parameters that correspond to participant's attention to losses, consistency in choices and the degree to which their choices are based on recent information. These parameters are respectively labeled attention to losses, consistency and recency. In particular, consistency and recency should be related to inhibitory control and attention to losses should correspond to motivation or reward seeking (Busemeyer & Stout, 2002). This model provides a way of decomposing performance into domains that, if accurate, can be predicted by personality traits from the RS and RI scales.

The BART is a more recent measure of decision making that has garnered success for

its face validity and for its ability to predict real world outcomes like smoking, substance abuse and psychopathology (Hopko, Lejuez, Daughters, Aklin, Osborne, Simmons, & Strong, 2006; Hunt, Hopko, Bare, Lejuez, & Robinson, 2005; Lejuez, Aklin, Jones, Richards, Strong, Kahler, & Read, 2003). During the task, participants are asked to fill hypothetical balloons by instructing the computer to add air each time a button is pressed. Each additional pump earns a set amount of money but also runs the risk of bursting the balloon. Participants therefore try to pump the balloon as full as possible but must stop pumping the balloon before it bursts.

The BART was developed explicitly as a method for measuring risk taking behavior (Lejuez, Read, Kahler, Richards, Ramsey, Stuart, Strong, & Brown, 2002). The BART also contains a strong component of response inhibition where a participant must stop pumping the balloon or suffer the consequence of a burst balloon. For this reason, the BART has modestly but reliably correlated with both reward seeking behavior as measured by the SSS and response inhibition as measured by the BIS (Lejuez, Read, Kahler, Richards, Ramsey, Stuart, Strong, & Brown, 2002). Unlike the IGT, the BART does not have a cognitive model to estimate various subcomponents of the task. In fact, looking at the average number of pumps per intact balloon provides the only reasonable measure of risk-taking behavior (Hunt, Hopko, Bare, Lejuez, & Robinson, 2005). This is an important limitation because the average number of pumps likely represents a combination of the reward seeking and response inhibition facets of impulsivity. In order to get around this limitation, on approximately half the trials the amount of money per pump was increased from \$0.05 to \$0.25. A difference between the average number of pumps in the low vs. high reward condition should reflect the motivational component of risk taking and this should allow me to partial out the remaining variance as a measure of response inhibition.

Part of the goal for the second study was to determine if the RSRIS can reliably tease apart what specifically is being measured by the BART and the IGT. Although both tasks

correlate modestly with both facets of impulsivity, the strength of these relationships and the theoretical differences between the two tasks should become clearer with a more refined approach to measuring impulsivity. Therefore, I predicted that the RSRIS will be able to accurately discriminate between the underlying processes of these two tasks.

Chapter 2

Study 1

2.1 Method

The primary aim of the first study was to validate the RSRIS against other measures of impulsivity and to validate the RSRIS by predicting outcome measures associated with alcohol, compulsive shopping and gambling.

As stated before, impulsivity can be subdivided into response inhibition and reward seeking. Response inhibition can be measured using the BIS, the Impulsiveness subscale of the EIS (IMP) and the response inhibition (RI) scale. I therefore expected to see correlations among these total scores. Reward seeking can be measured through the SSS, the Venturesomeness subscale of the EIS (VENT) and the risk seeking (RS) scale. The presence of meaningful correlations among the RSRIS and the older corresponding measures will provide evidence for concurrent validity. It is likely that some degree of correlation may be expected between response inhibition measures and reward seeking measures but these correlations should be small, which would provide evidence for discriminant validity.

Demonstrating correlations among the various impulsivity measures and the measures of problem behaviors will give evidence for predictive validity. I predicted that response

inhibition scales (BIS, IMP and RI) will correlate with both compulsive buying and possibly gambling and alcohol problems. Alternatively, I predicted that the reward seeking scales (SSS, VENT and RS) would correlate with alcohol problems and gambling but not compulsive buying.

2.1.1 Measures

RSRIS – The Risk Seeking and Response Inhibition Scales contain two factors composed of 7 items each. The questions are self-reflective (e.g., I often have trouble paying attention). To indicate their preference, participants give a score of 1-Strongly Agree to 7-Strongly Disagree. Factor scores are estimated using an empirical Bayesian approach and are given as θ estimates which are similar to Z-scores.

BIS – Barratt's Impulsivity Scale contains 34 items (four of which are fillers) divided into three factors. The 30 useful items are routinely summed into a total score. The questions are worded as self-reflective statements (e.g., I make up my mind quickly) and are scored on a four point scale ranging from 1 - Rarely/Never to 4 Almost Always/Always.

SSS – Zuckerman's Sensation Seeking Scale contains 40 items and four factors. Like the BIS, scores are commonly summed into a total score. Each item gives two alternative choices of statements and the participant selects which statement they agree with.

EIS – Eysenck's Impulsivity Scale contains a total of 54 items divided into three subscales. It is not recommended to use the total score for the entire scale but to only look at the individual factors. Furthermore, the empathy factor is often left out of analyses. The items are worded so that a participant can respond either yes or no (e.g., Do you quite enjoy taking risks?). Factor scores are calculated by summing the 'yes' responses.

YAACQ – The Brief Young Adult Alcohol Consequences Questionnaire was specifically developed for use in college age students (Read, Kahler, Strong, & Colder, 2006). The brief version of the YAACQ contains 24 items which fit into a unidimensional factor of alcohol problems. This scale is both valid and reliable (Kahler, Strong, & Read, 2005) as well as sensitive to changes over time (Kahler, Hustad, Barnett, Strong, & Borsari, 2008). Participants endorse yes/no statements about the consequences of their drinking (e.g., While drinking, I have said or done embarrassing things). 'Yes' items are summed up into a total score that represents the severity of participant's problems with drinking.

CBI – Compulsive buying was measured using the Compulsive Buying Instrument (Edwards, 1993). This scale contains 13 items that can be divided into four factors. Items are written as self-reflective statements (e.g., I hate to go shopping) and responses are solicited from a five-point Likert-type scale ranging from strongly agree to strongly disagree. There are some issues with the factor structure but total scores can be utilized successfully (Manolis & Roberts, 2008). Importantly, this scale measures the continuum of compulsive buying better than the other commonly used Compulsive Buying Scale (Faber & O'Guinn, 1992) which was intended for more clinical applications (Manolis & Roberts, 2008).

SOGS – Gambling problems were measured using the South Oaks Gambling Screen - Revised. The SOGS was originally developed to address pathological gambling criteria as set forth in the Diagnostic and Statistical Manual of Mental Disorders - Third Edition (Lesieur & Blume, 1987). The 20-item version of the SOGS addresses a variety of gambling behaviors by asking participants a series of yes/no questions (e.g., Have people criticized your gambling) as well as variable response questions with multiple options. Scoring consists of counting the number of yes responses and, in the case of multiple choice responses, above a critical cutoff. A score between 0 and 20 is obtained and higher scores indicate a greater

likelihood of pathological problems. The SOGS is well validated and has been found to still be reliable (Strong, Lesieur, Breen, Stinchfield, & Lejuez, 2004).

2.1.2 Participants

A total of 225 (154 females, 71 males) participants were recruited from the WSU Psychology Department Human Subjects Pool. The sample ranged in age from 18 years to 36 years (M = 20 years). The distribution of ethnicity is displayed in Figure Figure 2.1 and the distribution of class standing is displayed in Figure Figure 2.2. The demographics generally represent the composition of the WSU Subject Pool.

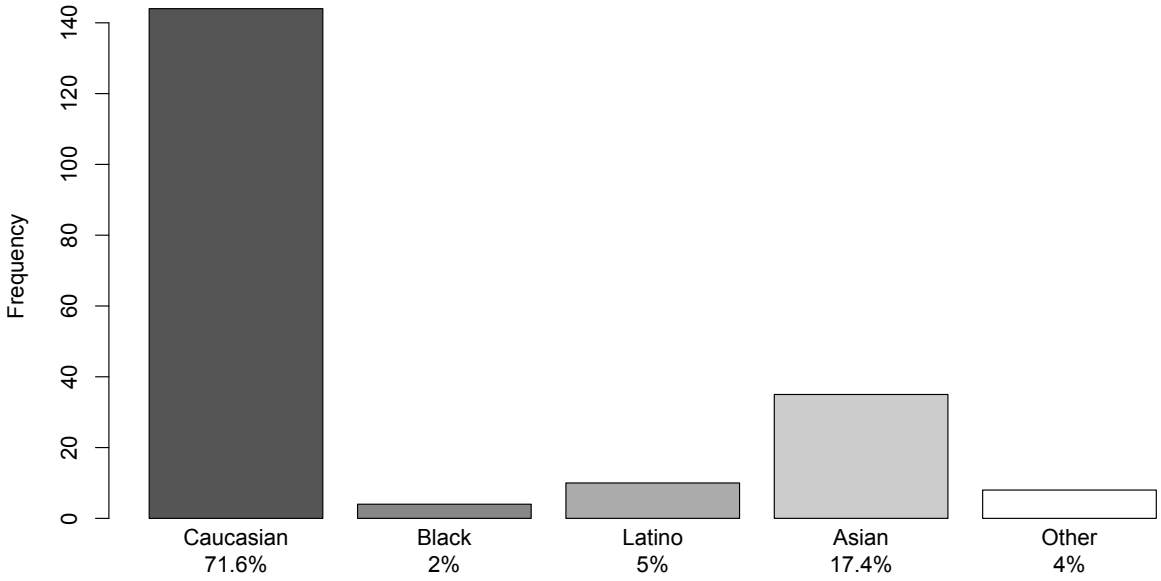

Figure 2.1: Ethnicity Distribution for Study One. Y-axis indicates frequency, proportions are listed under each group's label.

2.1.3 Procedure

Participants completed the questionnaires in a single research session. At the beginning of the session, participants read a brief form that provided directions for the experiment and

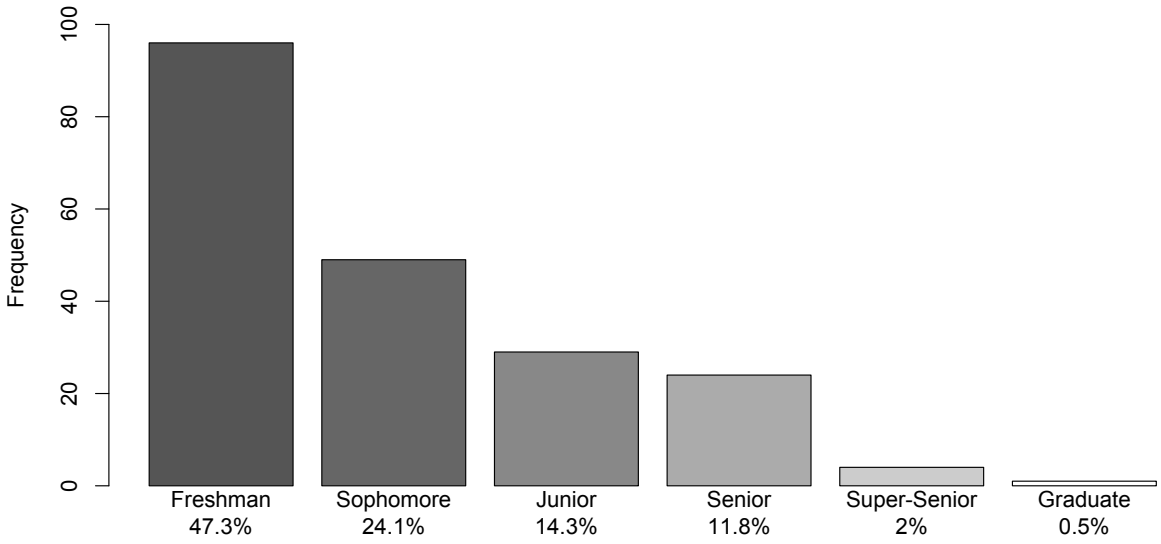

Figure 2.2: Distribution of class standing for Study One. Y-axis indicates frequency, proportions are listed under each group's label.

also served as a way of obtaining informed consent. All the data were collected anonymously to encourage honest responding from the participants. Following consent, questionnaires were presented to the participants using a computerized response system at individual workstations. Questions were presented one at a time and participants indicated their choice by clicking on the corresponding response using the mouse. Participants were informed that they could go back and change responses as necessary and that they could choose to not respond to questions if they were uncomfortable.[1] After participants had completed the questionnaires, they were fully debriefed and thanked for their participation.

[1]This did, unfortunately, result in rather high levels of missing data for some individuals, I'll return to this later (see section 2.2). In the future I would recommend not including a 'no response' option but instead allowing participants to continue with some sort of prompt (e.g., Are you sure you would like to skip this question?)

2.2 Analysis

Before any analysis, the dataset was screened for missing data, aberrant patterns of scoring and descriptive properties. All analysis were carried out in the R programming environment (R Development Core Team, 2007). Overall, there was a small proportion of missing data (3.3%) across all questionnaires and participants. Approximately 22 individuals had more than 10% missing data but were retained for future analysis as their pattern of missing data suggested that it was not *missing completely at random (MCAR)* but somewhat systematically (Schafer & Graham, 2002).

Debriefing revealed that many individuals left data missing on the the gambling problems questionnaire, the SOGS, because they simply did not gamble. To account for this, all missing responses on the SOGS were assumed to indicate no gambling and/or drinking behavior. Further examination of the missing data revealed that certain questionnaires were also prone to missing data, especially the SSS which was missing 7.6% of cells. Anecdotally, a number of participants indicated that questions on this measure were confusing, which may account for the higher proportion of missing data.[2]

The remaining missing values were imputed using multivariate imputation by chained equations (MICE; van Buuren & Oudshoorn, 2000; Schafer & Olsen, 1998). MICE imputes values in a process by which each missing value is replaced using predictions from similar, complete data. This process reiterates until convergence is reached and the missing value does not change from one iteration to the next. One of the hallmarks of multiple imputation is that unlike other missing data techniques it retains a certain degree of randomness which reflects the somewhat unknown nature of missing data. It does this by producing not one but several sets of complete data on which any subsequent statistics are analyzed and the

[2]I would like to point out that the RSRIS as a whole contained only 0.1% missing data. This is probably due to the fact that it is shorter than the other questionnaires, presented first in the experiment and contains a Likert-type response format that many students are familiar with.

results of these statistics pooled together. This reflects what the complete data could have looked like accounting for both the best estimate from the non-missing data as well as some randomness. For this experiment, five imputed datasets were utilized in all subsequent analyses.

Table 2.1: Descriptive Statistics for Study One, $N = 225$

Measure	Mean	SD	Min	Max	Range	Skew	Kurtosis
RI	-0.08	1.08	-3.38	3.5	6.88	0.16	0.25
BIS	61.58	8.91	42	88	46	0.36	0.06
IMP	6.57	3.88	0	17	17	0.43	-0.34
RS	-0.06	0.89	-2.49	2.64	5.13	0.12	-0.21
SSS	19.29	6.2	4	33	29	-0.17	-0.44
VENT	10.2	3.18	1.4	16	14.6	-0.48	-0.59
CBI	34.87	11.17	13	65	52	0.33	-0.62
YAACQ	7.13	5.42	0	23	23	0.71	-0.21
SOGS	1.07	1.87	0	12	12	2.87	9.89

After dealing with missing data, the resulting scores were analyzed descriptively. Table 2.1 shows the means, standard deviations, range, skewness and kurtosis for the data.[3] A couple of points are worth noting. First, the overall means for the RI and RS scales are fairly similar to the calibration sample ($M = 0$, $N = 613$), but the RS did show slightly less variability than before. Second, the means of the BIS, SSS and IMP scales were all significantly lower than those seen in the RSRIS calibration sample when compared using non-parametric Welch T-Tests (all $p's < 0.01$). Third, the means for the YAACQ and CBI were similar to reports from other recent studies (Kahler, Hustad, Barnett, Strong, & Borsari, 2008; Manolis & Roberts, 2008) although the CBI was higher than what was reported when it was originally developed(Edwards, 1993). Last, as expected scores on the SOGS were rather low and this resulted in high levels of skewness and kurtosis. However, these scores are typical for non-clinical samples (Holt, Green, & Myerson, 2003).

[3]Histograms are also available in Appendix E

Table 2.2: Concurrent and Discriminant Validity Matrix

Measure	RI	BIS	IMP	RS	SSS	VENT
RI	1.00					
BIS	0.59	1.00				
IMP	0.41	0.68	1.00			
RS	0.15	0.35	0.45	1.00		
SSS	0.18	0.37	0.41	0.62	1.00	
VENT	0.03	0.14	0.17	0.59	0.68	1.00

2.2.1 Concurrent & Discriminant Validity

Following the descriptive analysis, a correlation matrix (Table 2.2) was constructed in order to address discriminant and concurrent validity. In this table, concurrent validity is demonstrated by large correlations in the upper left quadrant for the RI and the lower right quadrant for the RS. As expected, the RI scale correlated highly with the BIS ($r = .59$) but unfortunately only correlated moderately with the IMP ($r = .41$). Both the BIS and IMP correlated strongly with each other ($r = .68$). On the other hand, the RS scale correlated well with both the SSS ($r = .62$) and the VENT($r = .59$) although it did not correlate to the same degree as those two measures correlated with each other ($r = .68$). Generally, this provides good evidence that the RS and RI scales both possess at least a fair degree of concurrent validity.

The lower left quadrant displays the discriminant validity between the scales. As expected, the RI and RS show very good discriminant validity with each other. This is unsurprising as they are both clearly unidimensional measures by design. The RI also demonstrates good discriminant validity with the SSS and the VENT. Alternatively, the RS shows a moderate discrimination with the BIS but less than desirable discrimination with the IMP. In fact, the RS correlates more highly with the IMP ($r = .45$) than the RI scale correlated with the IMP. This would be problematic except that it appears that the IMP does not discriminate from either the SSS or the VENT as well. It is therefore likely that the IMP scale does

not relate strictly to response inhibition, but rather contains a fair degree of overlap with reward seeking type behaviors as well. In this case, the IMP represents a broad approach to measuring impulsivity rather than a factor-defined approach.

2.2.2 Predictive Validity

With good concurrent validity demonstrated and some measure of discriminant validity, the focus now turns to the correlations with the behavioral outcomes. Table 2.3 contains the correlations that represent the predictive validity of the scales.

First, attention should be directed to the three leftmost columns which indicate the response inhibition type scales. As hypothesized, the RI, BIS and IMP predict both shopping problems and alcohol problems but not gambling. The RI does appear to perform more poorly than either the BIS or IMP scales in the YAACQ but performs moderately well in the CBI. All the correlations are fairly small, accounting for less than 10% of the variance in the outcome scales. The one exception to this rule is the IMP which accounts for 16% of the variance in the YAACQ.

The correlations in the three rightmost columns reveal a similar picture to what was hypothesized. The hypothesis that reward-seeking type scales can predict alcohol and gambling problems was partially upheld and these scales did not predict compulsive buying. The RS performed more poorly than the SSS in predicting alcohol problems but was able to predict them at a low level. The RS was the only measure beside the IMP that was able to predict any measure of gambling behavior, even though this correlation was very small. Referring back to the descriptive analysis this is not terribly surprising. The SOGS was developed as a diagnostic scale and a score of two or more indicates probable pathological gambling. However applying it to a non-clinical population produces a floor effect that severely attenuates any relationships. Interestingly, a subset of the 49 individuals who met the diagnostic

criteria showed a significant correlation between the RS and the SOGS ($r = .46$) but no correlation with the IMP ($r = .15$)[4]. This provides marginal evidence that the RS is 'best' predictor of gambling behavior.

Table 2.3: Predictive Validity Matrix

Measure	RI	BIS	IMP	RS	SSS	VENT
CBI	0.26	0.32	0.29	0.08	0.06	-0.09
YAACQ	0.23	0.30	0.40	0.22	0.33	0.01
SOGS	0.12	0.13	0.21	0.19	0.10	0.09

In summary, study one revealed that the RS and RI scales performed sufficiently well although there is definite room for improvement. The overall pattern across all the scales suggests that the relationship between impulsivity as a personality trait and problem behaviors is weak. This may be especially true for a sample of college students who may not represent the intended sample. It is possible that stronger relationships would emerge in clinical samples in contrast to the college sample presented in this study.

[4]See section E.4 for scatter plots that help to illustrate the problem.

Chapter 3

Study 2

3.1 Method

In the first study, the RS and RI scales were validated using concurrent, discriminant and predictive validity. Predictive validity was limited to behavioral outcomes assessed through self-report questionnaires. The present study extends that validation by examining two common measures of experimental decision making. In order to accomplish this, structural equation modeling was employed to examine predicted pathways between the latent constructs (i.e., Risk Seeking or Response Inhibition) and the dependent variables from experimental tasks.

Structural equation modeling requires a strong theoretical base on which the models need to be built. For this reason, it was necessary to establish a priori hypotheses about how the models were tested. With the BART, I hypothesized that the RS should predict the difference score between pumps in the low and high reward conditions. Furthermore, this difference score should also predict a portion of the variance in the global average for number of pumps. The RI was hypothesized to predict the global average for number of pumps. An individual with a high average of pumps per balloon should have lower inhibitory control.

It's also possible that the RS can predict global averages for the pumps, however I expected that this relationship should be fully or partially mediated by the difference between the high and low reward conditions.

In the case of the IGT, I expected that the RS would predict only the attention to loss component of IGT performance which is motivationally based. The RI should predict the recency and consistency components of the IGT which are more cognitive forms of impulsivity.

3.1.1 Tasks

IGT - The Iowa Gambling Task is a simulated card game in which participants are required to choose the top card off one of four decks. For this study, the task will be entirely based on the original computer implementation (Bechara, Damasio, Damasio, & Anderson, 1994; Bechara, Damasio, & Damasio, 2000). After selecting the card, a participant always wins a set amount of money that depends on which deck was chosen. For decks A and B, the amount won is consistently $100. Alternatively, decks C and D yield consistent gains of $50. In the beginning of the task, decks A and B result in wins with no accompanying losses but eventually begin to incur large losses (e.g $1250) that offset the large gains. Decks B and C contain much smaller losses and result in an overall net gain over time. Participants initially are attracted to the high gains of decks A and B but must learn over time to reverse their preference towards the lower but consistent gains of decks C and D. The outcomes for the entire gambling task are based off of a set script.

BART - The Balloon Analogue Risk Task is a simulated game in which participants fill a hypothetical balloon full of air. This study will use a version of the BART that is almost identical to the original implementation except for the manipulation of the amount won in the high and low payoff conditions (Lejuez, Read, Kahler, Richards, Ramsey, Stuart, Strong,

& Brown, 2002). For each pump that is added, the participant will earn a small amount of hypothetical money. In the low payoff condition this will be $0.05 and in the high payoff condition this will be $0.25. As participants continue to pump the balloon, their earnings are added into a small pot of money. At any point, the participant can redeem this pot of money and add it to their total earnings. However, if they continue to pump the balloon and the balloon pops, they do not receive the pot of money. The balloons will randomly pop between 1 and 128 pumps. This simulates a real life situation of diminishing gains where each additional pump yields more money but increases the risk of popping the balloon as it fills up.

3.1.2 Participants

In total, 229 participants (151 Females, 71 Males) were recruited from the WSU Psychology Department Human Subjects Pool. The distribution of ethnicity is displayed in FigureFigure 3.1 and the distribution of class standing is displayed in FigureFigure 3.2. The composition of the sample was similar to that in study one.

3.1.3 Procedures

Groups of up to 10 students participated in each experimental session. Each session began with a brief overview of the experiment and the obtainment of consent. All the data for study 2 were collected through computer responses. Instructions for the task were presented both during the consent process and in structured practice trials during the experiments. After completion of a brief demographic survey, participants filled out the RS and RI Scales. Participants then moved on to the experimental tasks, first completing the IGT followed by the BART. Once finished with the BART, participants were fully debriefed and dismissed. Each session took approximately 30-40 minutes from start to finish.

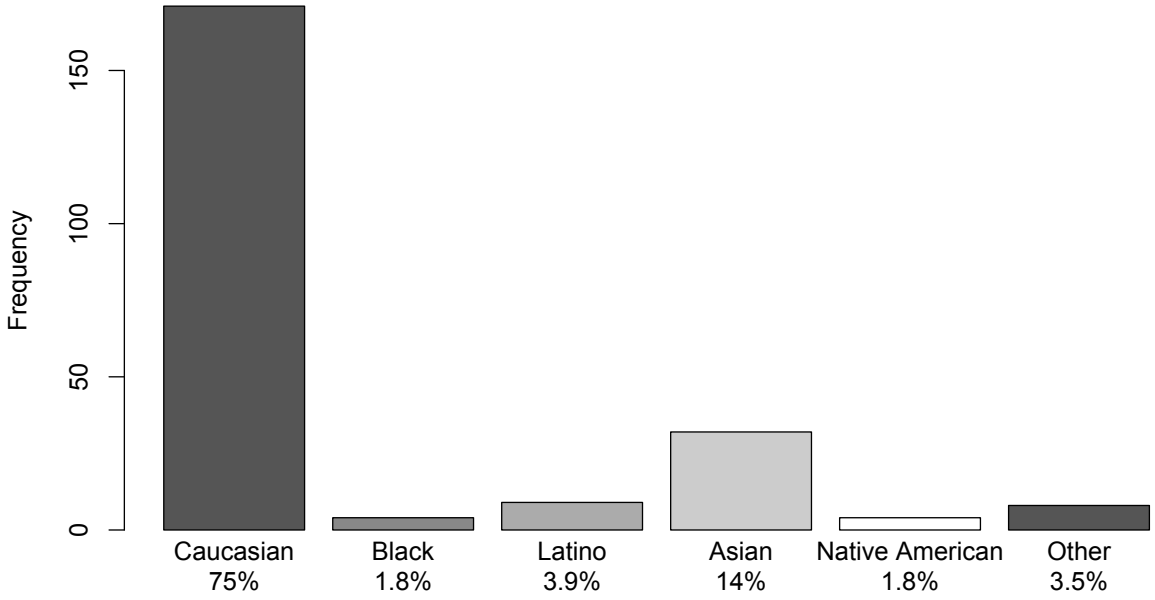

Figure 3.1: Ethnicity Distribution for Study Two. Y-axis indicates frequency, proportions are listed under each group's label.

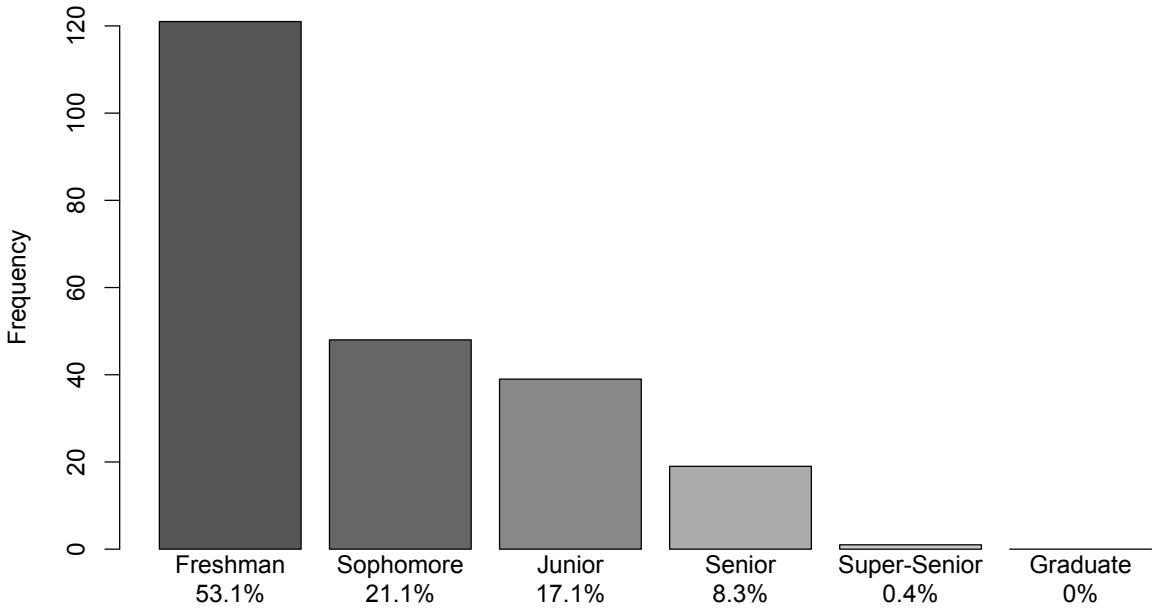

Figure 3.2: Distribution of class standing for Study Two. Y-axis indicates frequency, proportions are listed under each group's label.

3.2 Analysis

Analysis for the study proceeds as follows. I begin with an analysis of the results from the IGT and BART using a means-based (i.e., *t*-test and ANOVA) approach. Following this, I discuss a series of structural models that were constructed to look at the relationship between experimental performance and the RS and RI scales. All basic analysis were carried out in the R programming environment and the structural models were analyzed using M-Plus (Muthén & Muthén, 2006).

3.2.1 Iowa Gambling Task

To understand performance on the IGT, two separate approaches can be adopted. First, it is possible to look at the group as a whole and analyze how participants learn to shift preference from the bad, disadvantageous decks to the good, advantageous decks. The second approach is to decompose task performance using an expectancy-valence model that describes the extent to which participants base their decisions off of three parameters. These parameters include attention to losses, consistency and recency.

Table 3.1: IGT Choice Proportions
IGT Advantageous Choice Proportion. 1 Block = 20 Trials.

Block #	Mean	SD	Range
1	0.40	0.15	1.00
2	0.50	0.16	0.95
3	0.53	0.17	1.00
4	0.55	0.20	1.00
5	0.51	0.22	1.00
Overall	0.50	0.11	0.67

As a whole group, performance is depicted in Figure 3.3, Figure 3.4 and Table 3.1. In general, performance starts with a preference for the disadvantageous decks and shifts towards the advantageous decks as trials progress. The hypothesis that performance improves as a

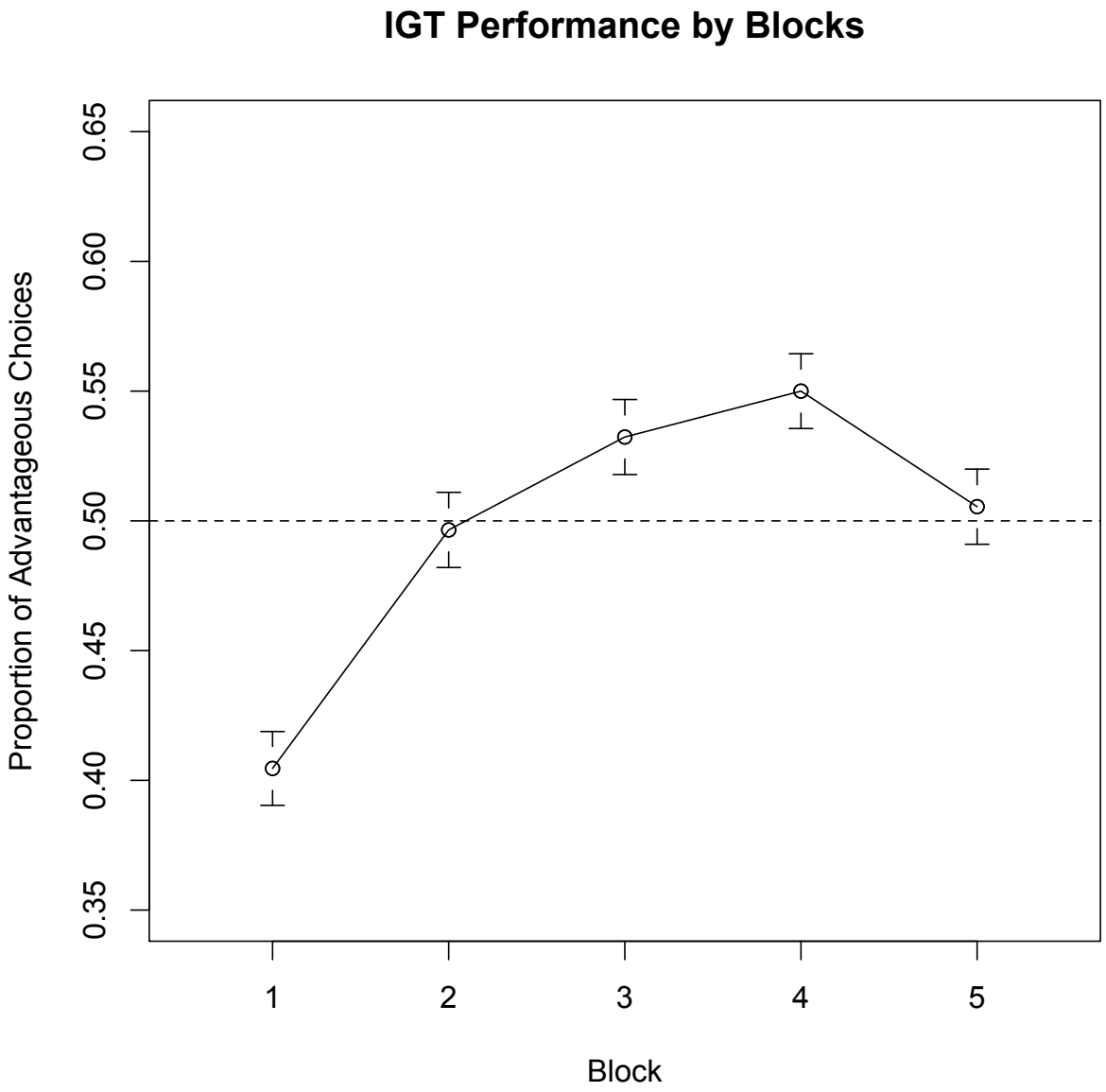

Figure 3.3: Block performance on the IGT. Error bars indicate the 95% confidence interval. The dotted line indicates no preference for either good or bad decks.

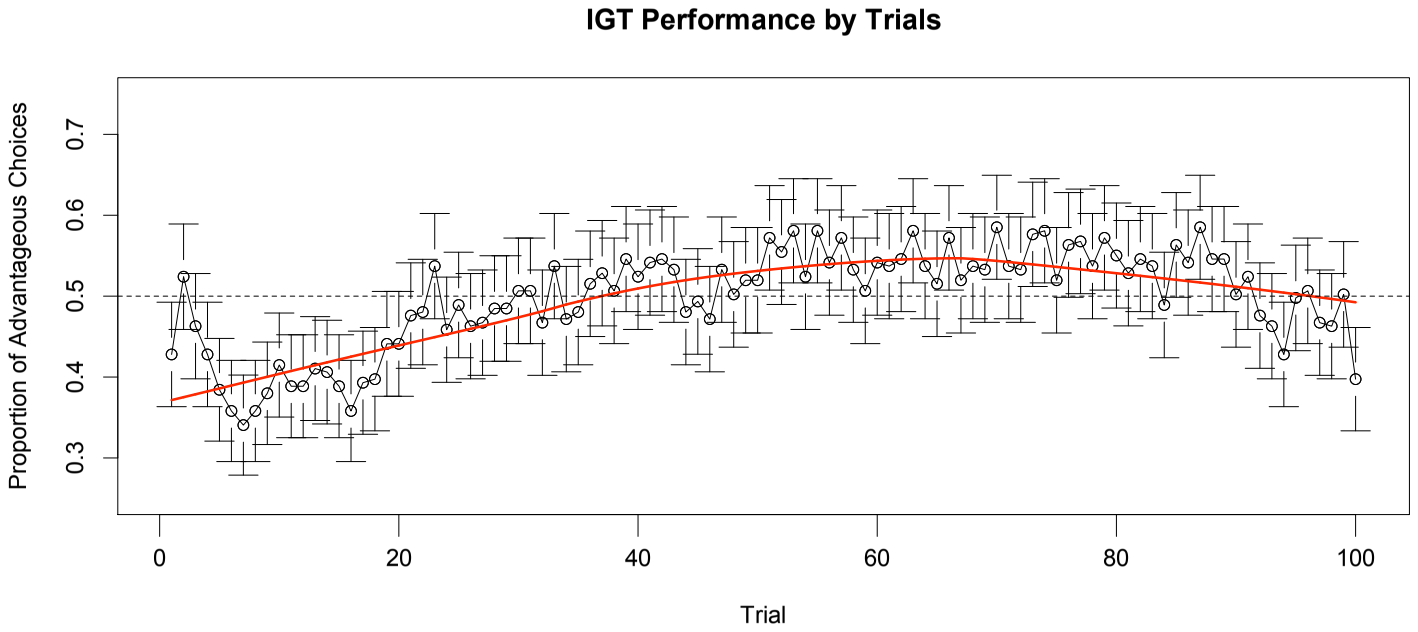

Figure 3.4: Trial Performance for the IGT. Error bars indicate the 95% confidence interval. A lowess smoother has been added to approximate the trend in performance. The dotted line indicates no preference for either good or bad decks.

function of trial block was tested using a repeated measure ANOVA. There was a significant effect of trial block ($F(4, 912) = 28.319, p < 0.05, \eta^2 = 0.11$). Referring again to Figure 3.3, this indicates that as the task progressed participants made significantly more advantageous choices. However, although advantageous choices are generally increasing, participants are overall performing quite poorly ($M = 0.5$ for all blocks combined). Participants only show a significant preference for advantageous choices in blocks three and four as evidenced by a 95% confidence interval. Thus, participants do appear to be *leaning* towards the advantageous decks but as a group there is not a clear preference for the advantageous decks.

Following the means analysis, parameter estimates were obtained for each participant using the Expectancy-Valence model (Busemeyer & Stout, 2002). Descriptive statistics for the Expectancy-Valence model are displayed in Table 3.2. The means for the Recency and Attention to Loss parameters are similar to previous reports, although the consistency parameter is higher and shows more variation (Busemeyer & Stout, 2002). As an interesting note, both Attention to Loss and Recency show negative kurtosis. Further examination reveals that the model appeared to truncate a number of scores at either 0 or 1 which resulted in a somewhat inverted-normal distribution. This was not a problem in the Consistency parameter.

Table 3.2: Descriptive Statistics for the IGT Expectancy-Valence Model Parameters

Measure	Mean	SD	Min	Max	Range	Skew	Kurtosis
Recency	0.38	0.44	0.00	1.00	1.00	0.61	-1.51
Attention To Loss	0.39	0.31	0.00	1.00	1.00	1.02	-0.21
Consistency	0.72	2.03	-5.00	5.00	10.00	0.03	0.86

In addition to parameter estimates, it is also possible to calculate Wilk's G^2 statistic for the Expectancy-Valence Model and a Baseline model in which choices are independently determined outside of prior experience. A G^2 difference score can be calculated for each individual and a simple comparison of this score allows one to determine if the cognitive

learning model predicts choices better than a model that assumes no learning. An alternative approach to assessing model fit involves calculating the Bayesian Information Criteria for each model. This analysis yields similar results. The Expectancy-Valence Model performed better than a baseline model for 53% of participants ($M = 3.63, SD = 17.36$). This means that for close to half of the participants, the Expectancy-Valence model did not provide a good fit. This is a much lower proportion than reported in many studies, however the sample size for this study is much larger ($229 > 33$) than that used in the original model evaluation. A likely explanation is that these individuals simply never demonstrate a learning strategy. This is confirmed if you look at many of these individual's response patterns. Furthermore, with the group on average performing close to chance (i.e., 50% advantageous choices) the lack of model fit is not surprising.

3.2.2 Balloon Analogue Risk Task

For the BART, four scores were calculated for each participant. First, a global adjusted average of the number of pumps per balloon was tallied. This average included the number of pumps across all conditions, but only for balloons in which the participant earned money and the balloon did not pop. The total number of bursts that occurred across all conditions was calculated. Third, difference scores were calculated across the experimental conditions by subtracting the average number of pumps in the low condition from the average in the high condition (i.e., $M_{dif} = M_{High} - M_{Low}$)[1]. Last, the analogous difference score was also calculated for the number of bursts by each experimental condition.

Table 3.3 displays the descriptive statistics for these four types of scores. For the average pumps and total bursts, these numbers are similar to what has been demonstrated in prior research with some obvious differences likely due to higher number of trials and the

[1]Difference scores were chosen instead of a ratio in order to approximate a gaussian distribution. Ratio scores demonstrated high kurtosis

experimental manipulation.

Table 3.3: BART Descriptives
Descriptive statistics for the Balloon Analogue Risk Task

Measure	Mean	SD	Min	Max	Range	Skew	Kurtosis
Average Pumps	36.99	13.97	9.03	81.33	72.31	0.52	0.3
Total Bursts	12.69	4.67	2	27	25	0.5	0.34
Difference in Pumps	-1.94	9.47	-32.6	24.95	57.55	0.06	0.82
Difference in Bursts	-0.39	2.79	-7	7	14	-0.16	-0.02

The difference scores can be interpreted as follows. A positive difference score indicates that the individual was more sensitive to the rewards in the high payoff condition or more inhibited during the low payoff condition. A negative difference score indicates higher responding in the low payoff condition. Figure 3.5 displays histograms for the difference scores using the adjusted average of pumps.

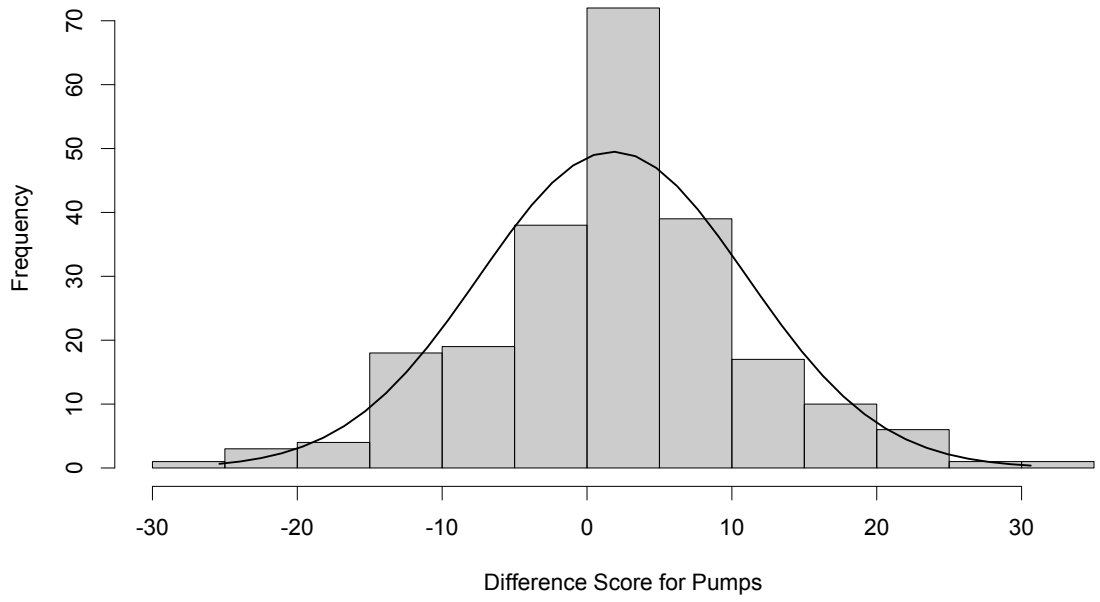

Figure 3.5: Histogram of the Difference Scores for Pumps

A simple t-test on the difference scores demonstrates that participants were significantly more conservative in the high condition than in the low condition ($t(228) = -3.11, p < 0.05, CI_{95} = -3.18 to - 0.71$). However, Figure 3.5 also shows that there is a good deal of variability and that while generally participants were more conservative in the high condition, the experimental manipulation produced greater conservatism in some individuals while more risky behaviors in others.

3.2.3 Structural Equation Modeling

Before fitting the structural models, the data for the RSRIS were analyzed descriptively as seen in Table 3.4. There was moderate skew and kurtosis ($skew < 3, kurtosis < 10$), which indicates that the data were approximately normal. Similar analyses were carried out on the descriptives for the IGT and the BART (see Table 3.2 and Table 3.3).

Table 3.4: RSRIS Descriptives
Descriptive statistics for the RSRIS items

Item	Mean	Sd	Skew	Kurtosis
RS1	5.39	1.08	-1.29	2.37
RS2	4.28	1.31	-0.18	-0.74
RS3	5.72	1.07	-1	0.92
RS4	4.06	1.46	0.07	-0.64
RS5	4.6	1.35	-0.6	-0.03
RS6	3.73	1.37	-0.02	-0.68
RS7	4.91	1.35	-0.95	0.39
RI1	4.17	1.55	-0.15	-1.03
RI2	4.5	1.53	-0.55	-0.68
RI3	3.9	1.58	-0.18	-1.09
RI4	4.28	1.49	-0.28	-0.81
RI5	3.44	1.39	0.17	-0.81
RI6	4.28	1.49	-0.5	-0.71
RI7	3.68	1.28	0.1	-0.18

Each model was estimated using weighted least squares since the RS and RI scales utilize ordered categories. This estimation is different from the normal maximum likelihood

estimation and is recommended for IRT models. Because of this estimation technique, the global fit parameters are different from what normally applies. All models were assessed for global fit using the following criteria. The comparative fit index (CFI) should be above 0.90 and preferably above 0.95. The root mean square error of approximation (RMSEA) should be below 0.10. Finally, the chi-square should ideally be nonsignificant but this criterion is fairly flexible as chi-squares will almost always be significant with large or even moderate sized samples (Kline, 2005).

An initial model was fitted that only included the RS and RI scales. This model is depicted in Figure 3.6 and is essentially a confirmatory factor analysis with ordered categorical indicators. Fit indices (see Table 3.5) indicate that the model fits adequately well.[2]

The second model is depicted in Figure 3.7. In this model, the IGT parameters are being predicted by the RS and RI factors. Again, model fit was good as seen in Table 3.5. The Consistency and Recency parameters were not predicted by the RI factor($p > .05$). The Attention to Loss was significantly predicted by the RS factor ($z = -10.585, p < 0.05$). This indicates that individuals higher in Risk Seeking tended to pay less attention to the losses during the IGT. The R^2 for this effect is approximately 10%.

The third and final model is shown in Figure 3.8. Here, pumps and the difference scores for high/low conditions are being predicted by the RI and RS scales respectively. The difference scores for high/low conditions also predicted the global average of pumps. Models for both pumps and bursts were tested, but only the model with pumps is shown due to superior fit both statistically as well as theoretically.

Because all parameters in this model were significant ($p < 0.05$) there are three paths that need explanation. First, risk-seeking predicts the global average of pumps. This is expected as risk is explicitly suggested as a component of the BART's name although it

[2]The RMSEA is higher than it should be, however this is likely due to a relatively small sample for the WLS estimator. Furthermore, in the calibration sample (N=613) the RMSEA for the same model was .085

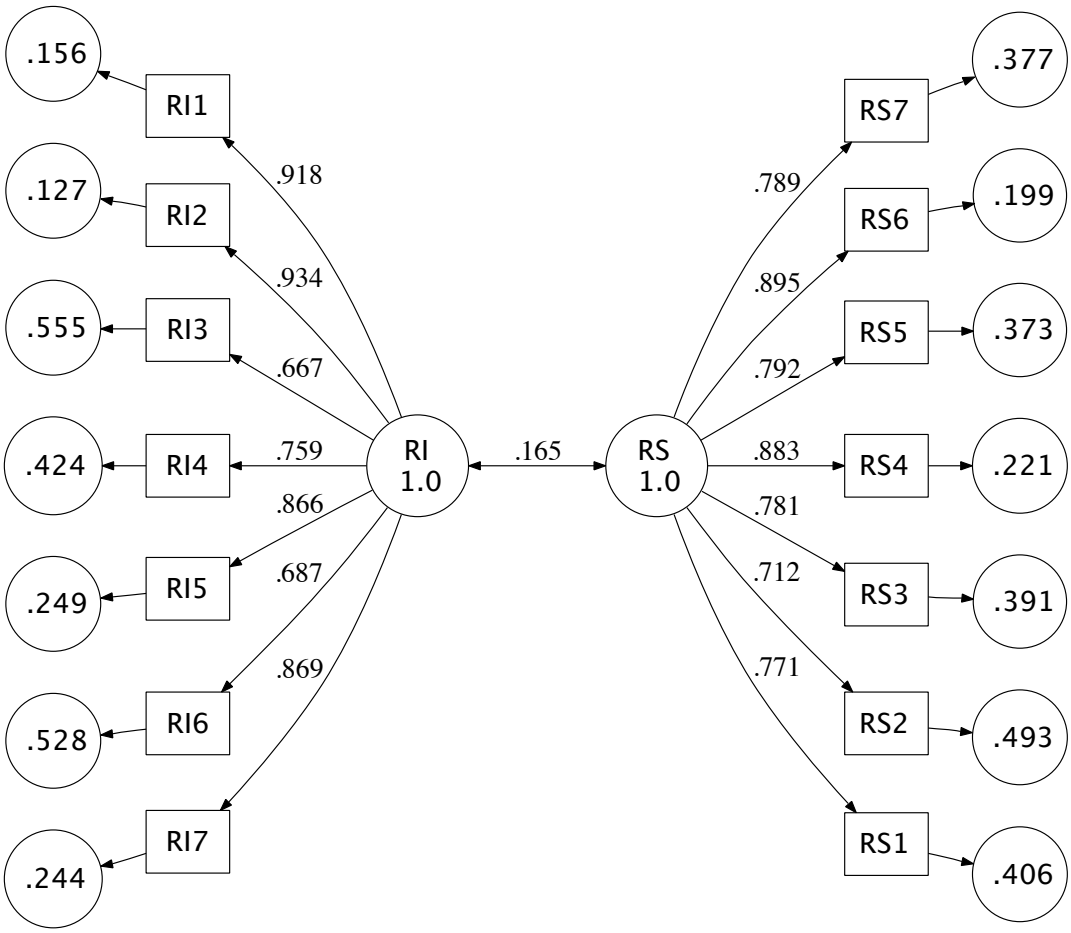

Figure 3.6: Confirmatory Factor Model for the RI and RS scales. All parameters are standardized with regards to both X and Y.

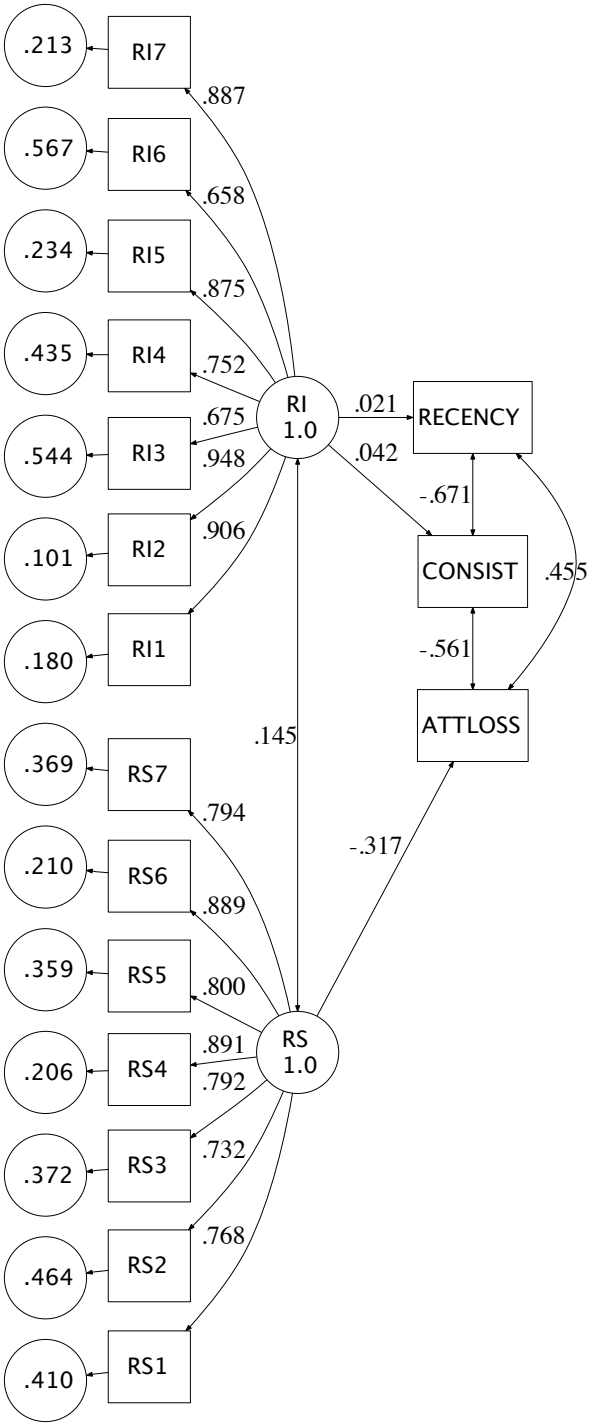

Figure 3.7: Structural Model for the RI and RS scales predicting IGfT Expectancy-Valence Parameters. All parameters are standardized with regards to both X and Y.

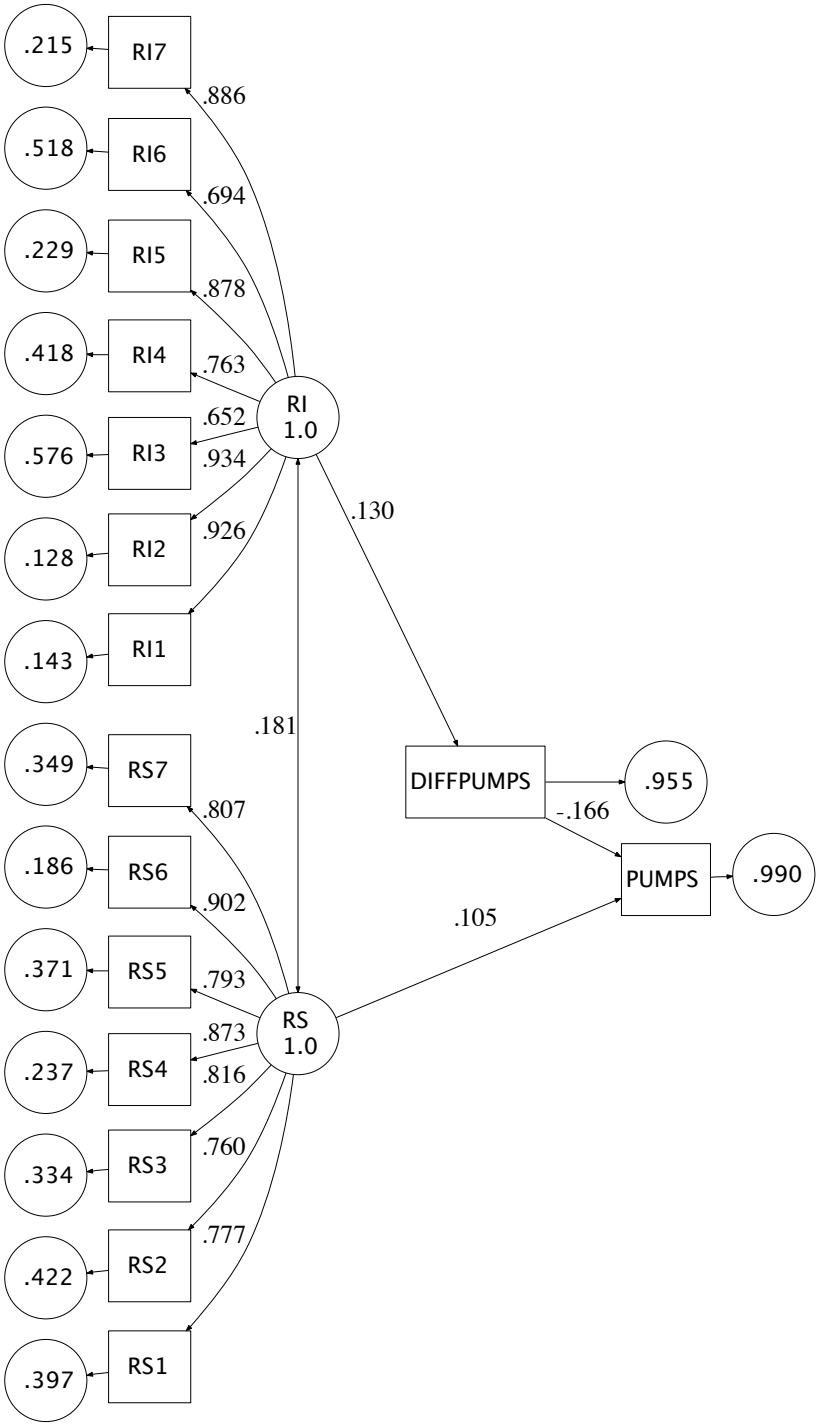

Figure 3.8: Structural Model for the RI and RS scales predicting BART Pumps and Difference Scores for Pumps (DIFFPUMPS). All parameters are standardized with regards to both X and Y.

was originally hypothesized that response inhibition would be a better predictor. This is a positive relationship which indicates that participants who endorsed higher levels of risk pumped the balloon more than others. The second important path shows a positive relationship between response inhibition and difference scores. In other words, individuals higher in response inhibition have a higher difference score. Recall that higher, positive difference scores indicate more pumps on the high condition than the low condition. Finally, the global average of pumps can be predicted by the difference score. This suggests that the response inhibition factor is indirectly related to the global average of pumps through its effects on the difference in the high and low condition. This indirect effect is significant (standardized coefficient $= -.022, p < .05$) but not meaningful. Thus the general form of the model is similar to what was hypothesized, but the effects of the RI and RS were reversed.

Table 3.5: Summary of Model Fit Statistics

Model	χ^2	df	p	CFI	RMSEA	Description
1	363.56	76	.000	.970	.129	Two Factor RSRIS
2	607.74	115	.000	.971	.137	RSRIS & IGT
3	484.15	102	.000	.967	.128	RSRIS & BART

Chapter 4

Discussion

Study one provided reasonable evidence for concurrent validity with both the RS and RI scales, as they correlated highly with the scales that were considered to measure a common facet of impulsivity. The one exception to this was the Impulsivity factor of the EIS which correlated more highly with the RS than the RI. I will return to this curiosity later. Discriminant validity was demonstrated through low correlations between the reward-seeking type scale and the response-inhibition type scales.

The qualification for the validity of the RS and RI scales is due to the trouble that they had with predictive validity in both studies. In all cases, the RI showed lower correlations with the measures of problem behavior than either the BIS or the IMP. The RS also performed worse than the SSS when predicting alcohol problems, but did perform better than the other scales when predicting gambling behaviors for individuals who did endorse gambling problems. Overall, the best measure of broadly-defined impulsivity is the IMP factor from the EIS. One explanation for this finding is that the IMP measures a more general form of impulsivity than was addressed by either the RI or RS scales. This is a problem for a two-factor theory of impulsivity because a unitary measure of impulsivity should not perform as well as more specifically defined measures. Of course, there could be a problem

with specifying impulsivity as two factors. Recall that this definition was derived from two lines of experimental research. As shown by study two, experimental measures of impulsivity may not be the most reliable source for determining personality factors.

Both scales performed only moderately well in predicting the experimental tasks in study two. In the BART, both scales predicted only a small fraction of behavior. Furthermore, the model that ultimately fit was not the model that was originally expected. So while significant, these paths were hardly meaningful. There are a number of possible explanations for this outcome. The most parsimonious explanation is that the constructs measured by the RS and RI scales are not meaningfully related to the construct measured in the BART. While a disconnect between the constructs is likely a strong cause, I think that this is not the sole explanation. Another likely reason for the weak relationship between the BART and the RSRIS is a problem of reliability. In this model, the BART was treated as a manifest or observed variable while the RS and RI factors were latent variables. Thus, the RS and RI factors represent their respective constructs with the error variance statistically removed. The outcomes on the BART, however, are assumed to be measured without error. Naturally, this assumption is not strictly met but the hope is that there is enough meaningful variance in the manifest variables that the assumption can be safely violated. The problem, is that if there is an excess of error variance in the measure then any meaningful relationship would be attenuated. In this case, it seems unlikely that the RS scale, which is a clear measure of risk, does not relate more strongly to any variable measured by the Balloon Analogue *Risk* Task. Again, the problem may likely be one of excessive error variance.

In regards to what can be extrapolated from the BART, it appears that the high and low payoff manipulation influenced participants self-control. Specifically, an individual endorsing higher levels of response inhibition was more likely to pump the balloon more in the high payoff condition. Although people were more conservative in the high condition, the level of that conservatism is influenced by their response inhibition trait. Second, individuals who

are risk seeking were more likely to have a higher global average across both conditions. Put another way, the risk seeking factor does not seem to influence their sensitivity to the high/low manipulation but rather influences the mean number of pumps across both conditions.

For the IGT, the situation was less fortunate. The RI scale was unable to reliably predict any of the parameters from the IGT but the RS scale did reliably relate to participant's attention to losses. This makes good theoretical sense as the RS can be seen as a motivational construct. However, the lack of relationship between the RI scale and the IGT is paradoxical. The consistency parameter from the IGT is specifically referenced as relating to boredom or distractibility. Participants high in consistency are cognitively focused while those low in consistency are easily distracted or bored with the task. The lack of relationship between the RI (which contains questions such as 'I am easily distracted') to a parameter that measures inconsistency due to distraction seems odd. A possible explanation for this poor performance is a little more straightforward. One of the features of the Expectancy-Valence model is that model fit can be diagnosed through the G^2 statistic. In this sample, the Expectancy-Valence model provided a poor fit for 47% of participants. That lack of fit cannot be ignored, suggesting that either the model is ill-specified or that it simply does not work with this sample. I would argue that it is actually a combination of these two explanations. The classic IGT has never worked particularly well with WSU samples (Wirick, 2006) and has performed poorly in other samples as well (Glicksohn, Naor-Ziv, & Leshem, 2007). The expectancy-valence learning model requires that participants eventually perform advantageously or at least consistently. This may be a misspecification because many participants appear to respond to other factors that are not assessed by a purely cognitive model of performance. This has been confirmed qualitatively during debriefing where participants are asked about the IGT. Many participants indicate that their responding is due to idiosyncratic behavior that tries to predict the pattern of the payoffs as it relates to the different decks. This

behavior occurs in spite of the fact that they are told at the beginning of the experiment that each deck is completely independent of each other. A conclusion may therefore be that many participants fail to 'get' the IGT. For the remaining 53% of participants, a learning strategy does appear to accurately model their behavior.

So should the RS and RI scales be adopted as replacements for the older scales? Again the answer should be qualified as maybe. It appears that the RS and RI scales perform moderately well for their limited length. In a situation where an experimenter lacks the time to utilize a longer measure of impulsivity, the RS and RI scales may suffice. However, it appears that the predictive validity of these scales is somewhat lacking and therefore they need revision before they can be adopted as replacements for the broader, older measures of impulsivity.

An even larger issue when considering the results of this study is what it tells us about the ability to predict outcome behaviors using personality. In the best case, the IMP subscale was able to predict alcohol problems with $r = 0.4, r^2 = .16$. While 16% of the variance is considered a medium effect size, this leaves a lot of room for other factors to explain problem behaviors. Furthermore, many measures of impulsivity are used as mediators between experimental results and real-world outcomes. Only accounting for 16% of the variance in the outcomes means that even the best impulsivity measure provides only a modest analogue for real world validity. One of the explicit goals for developing the RSRIS was that it would provide a more accurate way of predicting impulsivity and would therefore be more appropriate as a mediating bridge between experimental and real world outcomes. This goal was simply not met. Therefore, a vacuum still remains between the experimental findings and their validity in the real world.

Related to the issue of the validity of impulsivity measures is the problem that may be inherent in how this data is collected. All of the measures utilized in study one were self-report. As has been demonstrated numerous times, the validity of self-report measures

is directly affected by the participant's knowledge of themselves and by their willingness to share that knowledge. Anecdotally, I can confirm that most people are willing to share that they are impulsive. Each time that I begin to describe my research I am oftentimes met with the comment 'You should study me, I am extremely impulsive' or some other derivative remark. The question then becomes how much are people aware of their own impulsivity? This question, to my knowledge, has not yet been answered. It is therefore possible that future attempts to measure impulsivity should focus on more subtle forms of assessment. This may result in a loss of face validity, but may increase the construct validity of these measures.

4.1 Limitations

There are some limitations of this study that need to be addressed. Irrespective of the outcome, different approaches may have produced different results. In the case of study one, it may have been appropriate to utilize more outcome measures than just the CBI, YAACQ and the SOGS. These were chosen because they represented prominent measures in each of their respective fields. However, the SOGS in particular had strong floor effects that precluded any meaningful interpretation. It may have been advantageous to use a wider spectrum of measures to provide stronger evidence for the validity of the scales.

The use of a conventional, college sample may have also produced sample dependent error. Predictive outcome validity may have been demonstrated more strongly through the use of clinical samples who are known to demonstrate problem behaviors rather than college students where problems were self-reported. This was a decision made for convenience and it was an issue that I would have had difficulty getting around. It should be noted that there were reasonable amounts of problem behavior reported. Therefore, any limitation due to self-report would be an issue of reliability rather than the absence of problem behaviors.

Third, the experimental tasks may also have been problematic. Both the IGT and the BART are not easily decomposed. The IGT is not only a learning task, but also a reversal task, a preference task that likely depends to some degree on emotional evaluations. This may characterize some of the aspects of impulsivity but it complicates validity because it is hard to determine which of these factors or combination of factors is explaining performance in an individual. Similarly, the BART may also contain more than just preference for risk. The experimental manipulation which was intended to modify participants preference for risk, instead was influenced by their response inhibition. Thus, the task is not as clearly defined as expected.

Last, there may be some inherent limitations from the development of the RS and RI scales. Although there is evidence for their reliability, they may not be reliably measuring the intended constructs. Furthermore, during their development a tradeoff was made between more reliability and brevity. This tradeoff may not have been justified in the long run and it may have contributed to their less than stellar performance in this dissertation.

4.2 Future Directions

The predictive validity demonstrated in study one suggests that none of the measures, old or new, performed exceedingly well. There is still a need for a new measure of impulsivity. I have already explained the need for a strong theoretical approach when developing a new measure and I would reiterate that theories of impulsivity should guide future scale development. Were I to continue the development of the RS and RI scales I would revisit the theoretical understanding of impulsivity through qualitative interviews with clinicians, experimenters and other subject matter experts. The RS and RI scales as they exist now are still very preliminary and need extensive revision before they are ready for business. Going back to the theory would be the best way to continue their development.

Study two also demonstrates the need for better experimental methodologies. The IGT in particular currently enjoys a sort of demigod status as a measure of decision making. Unfortunately, nearly half of all participants fail to adopt an advantageous strategy even in a non-neuropsychological sample. This is true using either the typical choice data or when considering a cognitive model like expectancy-valence learning. The BART as well appears to have many more factors contributing to behavior than suggested by the authors. It should not be treated as a pure measure of risk taking, and its design, while definitely interesting, provides more face validity than real construct validity. In both of these tasks, the structural models also suggest a high degree of unexplained error variance. Thus new experimental decision making tasks should focus on not only validity issues but also issues of reliability.

4.3 Conclusions

This dissertation provided an approach to studying the validation for a new measure of trait impulsivity. The results suggest that marginal validity was demonstrated but that improvements are needed before the RSRIS can be utilized in further research. In particular, validity needs to be improved in predicting the results from experimental research and demonstration that the RSRIS can predict outcomes from real-world behaviors and not just self-report questionnaires.

These studies also suggest that improvements are needed in the understanding of what real world impulsivity looks like and how impulsivity is measured in the laboratory. In the former case, self-report measures of problem behaviors may not capture elements that can be explained by personality or may themselves be invalid or unreliable. Experimental tasks may also be unreliable or could be identifying phenomena that are interesting in the lab but irrelevant for real life.

Last, there is still a need for a refined measure of impulsivity. Study one demonstrated that the older measures of impulsivity, while better than the RSRIS, are still lacking when it comes to predicting real world problems. The criticisms of their psychometric properties still apply and this dissertation demonstrates that their validity is still marginal. So while the RSRIS is not the replacement, that does not mean that a replacement is not needed. Therefore, future research should focus on developing an empirically sound measure of impulsivity that deals with all of the statistical shortcomings of the current measures and the theoretical shortcomings of the RSRIS.

References

Aluja, A., Rossier, J., García, L., Angleitner, A., Kuhlman, M., & Zuckerman, M. (2006). A cross-cultural shortened form of the ZKPQ (ZKPQ-50-cc) adapted to English, French, German, and Spanish languages. *Personality and Individual Differences*, *41*(4), 619–628.

Barrett, P., Petrides, K., Eysenck, S., & Eysenck, H. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences*, *25*(5), 805–819.

Bechara, A., Damasio, A., Damasio, H., & Anderson, S. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1-3), 7–15.

Bechara, A., Damasio, H., & Damasio, A. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, *10*(3), 295–307.

Bechara, A., Dolan, S., & Hindes, A. (2002). Decision-making and addiction (Part II): Myopia for the future or hypersensitivity to reward. *Neuropsychologia*, *40*(10), 1690–1705.

Bechara, A., Tranel, D., Damasio, H., & Damasio, A. (1996). Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cerebral Cortex*, *6*(2), 215–225.

Billieux, J., Rochat, L., Rebetez, M., & Van der Linden, M. (2008). Are all facets of impulsivity related to self-reported compulsive buying behavior? *Personality and Individual Differences*, *44*(6), 1432–1442.

Blaszczynski, A., Steel, Z., & Mcconaghy, N. (1997). Impulsivity in pathological gambling: The antisocial impulsivist. *Addiction*, *92*(1), 75–87.

Breslin, F., Sobell, M., Cappell, H., Vakili, S., & Poulos, C. (1999). The effects of alcohol, gender, and sensation seeking on the gambling choices of social drinkers. *Psychology of Addictive Behaviors*, *13*, 243–252.

Busemeyer, J., & Stout, J. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the bechara gambling task. *Psychological Assessment*, *14*(3), 253–262.

Caci, H., Nadalet, L., Baylé, F., Robert, P., & Boyer, P. (2003). Cross-cultural study of the impulsiveness-venturesomeness-empathy questionnaire (IVE-7). *Comprehensive Psychiatry*, *44*(5), 381–387.

Dawe, S., Gullo, M. J., & Loxton, N. J. (2004). Reward drive and rash impulsiveness as dimensions of impulsivity: Implications for substance misuse. *Addictive Behaviors*, *29*(7), 1389–1405.

Dawe, S., & Loxton, N. J. (2004). The role of impulsivity in the development of substance use and eating disorders. *Neuroscience and Biobehavioral Reviews*, *28*(3), 343–351.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience and Biobehavioral Reviews*, *30*(2), 239–271.

Edwards, E. (1993). Development of a new scale for measuring compulsive buying behavior. *Financial Counseling and Planning*, *4*, 67–84.

Embretson, S., & Reise, S. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Assoc Inc.

Evenden, J. (1999). Varieties of impulsivity. *Psychopharmacology*, *146*(4), 348–361.

Eysenck, S., Pearson, P., Easting, G., & Allsopp, J. (1985). Age norms for impulsiveness, venturesomeness and empathy in adults. *Personality and Individual Differences*, *6*(5), 613–619.

Faber, R., & O'Guinn, T. (1992). A clinical screener for compulsive buying. *Journal of Consumer Research*, *19*(3), 459–469.

Fellows, L. K., & Farah, M. J. (2005). Dissociable elements of human foresight: A role for the ventromedial frontal lobes in framing the future, but not in discounting future rewards. *Neuropsychologia*, *43*(8), 1214–1221.

Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491.

Franken, I. H., van Strien, J. W., Nijs, I., & Muris, P. (2008). Impulsivity is associated with behavioral decision-making deficits. *Psychiatry Research*, *158*(2), 155–163.

Glicksohn, J., Naor-Ziv, R., & Leshem, R. (2007). Impulsive decision-making: Learning to gamble wisely? *Cognition*, *105*(1), 195–205.

Gray, J. M., & Wilson, M. A. (2007). A detailed analysis of the reliability and validity of the sensation seeking scale in a uk sample. *Personality and Individual Differences*, *42*(4), 641–651.

Hammelstein, P. (2004). Faites vos jeux! another look at sensation seeking and pathological gambling. *Personality and Individual Differences*, *37*(5), 917–931.

Hinson, J. M., Jameson, T. L., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology-Learning Memory and Cognition*, *29*(2), 298–306.

Holt, D. D., Green, L., & Myerson, J. (2003). Is discounting impulsive? Evidence from temporal and probability discounting in gambling and non-gambling college students. *Behavioural Processes*, *64*(3), 355–367.

Hopko, D., Lejuez, C., Daughters, S., Aklin, W., Osborne, A., Simmons, B., & Strong, D. (2006). Construct validity of the Balloon Analogue Risk Task (BART): Relationship with mdma use by inner-city drug users in residential treatment. *Journal of Psychopathology and Behavioral Assessment*, *28*(2), 95–101.

Hunt, M. K., Hopko, D. R., Bare, R., Lejuez, C. W., & Robinson, E. V. (2005). Construct validity of the Balloon Analog Risk Task (BART) - Associations with psychopathy and impulsivity. *Assessment*, *12*(4), 416–428.

Kahler, C., Hustad, J., Barnett, N., Strong, D., & Borsari, B. (2008). Validation of the 30-day version of the Brief Young Adult Alcohol Consequences Questionnaire for use in longitudinal studies. *Journal of Studies on Alcohol*, *69*(4), 611–615.

Kahler, C., Strong, D., & Read, J. (2005). Toward efficient and comprehensive Measurement of the alcohol problems continuum in college students: The Brief Young Adult Alcohol Consequences Questionnaire. *Alcoholism: Clinical and Experimental Research*, *29*(7), 1180–1189.

Kline, R. (2005). *Principles and practice of structural equation modeling*. The Guilford Press.

Lejuez, C., Aklin, W., Jones, H., Richards, J., Strong, D., Kahler, C., & Read, J. (2003). The Balloon Analogue Risk Task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology*, *11*(1), 26–33.

Lejuez, C., Read, J., Kahler, C., Richards, J., Ramsey, S., Stuart, G., Strong, D., & Brown, R. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84.

Lesieur, H., & Blume, S. (1987). The South Oaks Gambling Screen (SOGS): A new instrument for the identification of pathological gamblers. *American Journal of Psychiatry*, *144*(9), 1184–1188.

Magid, V., & Colder, C. (2007). The UPPS Impulsive Behavior Scale: Factor structure and associations with college drinking. *Personality and Individual Differences*, *43*(7), 1927–1937.

Manolis, C., & Roberts, J. (2008). Compulsive buying: Does it matter how it's measured? *Journal of Economic Psychology*, *29*(4), 555–576.

Martens, M., Neighbors, C., Lewis, M., Lee, C., Oster-Aaland, L., & Larimer, M. (2008). The roles of negative affect and coping motives in the relationship between alcohol use and alcohol-related problems among college students. *Journal of Studies on Alcohol and Drugs*, *69*(3), 412–419.

McDaniel, S., & Zuckerman, M. (2003). The relationship of impulsive sensation seeking and gender to interest and participation in gambling activities. *Personality and Individual Differences*, *35*(6), 1385–1400.

Miller, E., Joseph, S., & Tudway, J. (2004). Assessing the component structure of four self-report measures of impulsivity. *Personality and Individual Differences*, *37*(2), 349–358.

Mueller, A., Mueller, U., Albert, P., Mertens, C., Silbermann, A., Mitchell, J., & de Zwaan, M. (2007). Hoarding in a compulsive buying sample. *Behaviour Research and Therapy*, *45*(11), 2754–2763.

Muthén, L., & Muthén, B. (2006). Mplus user's guide. *Los Angeles, CA: Muthén & Muthén*, *2006*.

Nower, L., Derevensky, J., & Gupta, R. (2004). The Relationship of Impulsivity, Sensation Seeking, Coping, and Substance Use in Youth Gamblers. *Psychology of Addictive Behaviors*, *18*(1), 49–55.

Patton, J., Stanford, M., & Barratt, E. (1995). Factor Structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, *51*(6), 768–774.

Petry, N. (2001). Substance abuse, pathological gambling, and impulsiveness. *Drug and Alcohol Dependence*, *63*(1), 29–38.

Pirog, S., & Roberts, J. (2007). Personality and credit card misuse among college students: The mediating role of impulsiveness. *The Journal of Marketing Theory and Practice*, *15*(1), 65–77.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Read, J., Kahler, C., Strong, D., & Colder, C. (2006). Development and preliminary validation of the Young Adult Alcohol Consequences Questionnaire. *Journal of studies on alcohol*, *67*(1), 169–177.

Reynolds, B., Richards, J. B., & de Wit, H. (2006). Acute-alcohol effects on the Experiential Discounting Task (EDT) and a question-based measure of delay discounting. *Pharmacology Biochemistry and Behavior*, *83*(2), 194–202.

Roberti, J. (2004). A review of behavioral and biological correlates of sensation seeking. *Journal of Research in Personality*, *38*(3), 256–279.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological methods*, *7*(2), 147–177.

Schafer, J., & Olsen, M. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545–571.

Simons, J. S., & Carey, K. B. (2006). An affective and cognitive model of marijuana and alcohol problems. *Addictive Behaviors*, *31*(9), 1578–1592.

Slutske, W., Caspi, A., Moffitt, T., & Poulton, R. (2005). Personality and problem gambling: A prospective study of a birth cohort of young adults. *Archives of General Psychiatry*, *62*(7), 769–775.

Smillie, L., & Jackson, C. (2006). Functional impulsivity and reinforcement sensitivity theory. *Journal of Personality*, *74*(1), 47–84.

Someya, T., Sakado, K., Seki, T., Kojima, M., Reist, C., Tang, S., & Takahashi, S. (2001). The Japanese version of the Barratt Impulsiveness Scale, 11th version (BIS-11): Its reliability and validity. *Psychiatry and Clinical Neurosciences*, *55*(2), 111–114.

Spinella, M. (2007). Normative data and a short form of the Barratt Impulsiveness Scale. *International Journal of Neuroscience*, *117*(3), 359–368.

Strong, D., Lesieur, H., Breen, R., Stinchfield, R., & Lejuez, C. (2004). Using a Rasch model to examine the utility of the South Oaks Gambling Screen across clinical and community samples. *Addictive Behaviors*, *29*(3), 465–481.

Swann, A. C., Bjork, J. M., Moeller, F. G., & Dougherty, D. M. (2002). Two models of impulsivity: Relationship to personality traits and psychopathology. *Biological Psychiatry*, *51*(12), 988–994.

van Buuren, S., & Oudshoorn, C. (2000). *Multivariate Imputation by Chained Equations: MICE V1.0 Users's Manual*. Leiden, The Netherlands: TNO Prevention and Health, Public Health.

Vitaro, F., Arseneault, L., & Tremblay, R. (1999). Impulsivity predicts problem gambling in low SES adolescent males. *Addiction*, *94*(4), 565–575.

Whiteside, S., & Lynam, D. (2001). The Five Factor Model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, *30*(4), 669–689.

Whiteside, S., Lynam, D., Miller, J., & Reynolds, S. (2005). Validation of the UPPS impulsive behavior scale: A four-factor model of impulsivity. *European Journal of Personality*, *19*(7), 559–574.

Whitney, P., Jameson, T., & Hinson, J. (2004). Impulsiveness and executive control of working memory. *Personality and Individual Differences*, *37*(2), 417–428.

Williams, J., & Taylor, E. (2006). The evolution of hyperactivity, impulsivity and cognitive diversity. *Journal of the Royal Society Interface*, *3*(8), 399–413.

Wirick, A. (2006). *Reversal Learning and Somatic Markers in the Gambling Task*. Master's thesis, Washington State University.

Wohl, M., Matheson, K., Young, M., & Anisman, H. (2008). Cortisol rise following awakening among problem gamblers: Dissociation from comorbid symptoms of depression and impulsivity. *Journal of Gambling Studies*, *24*(1), 79–90.

Zermatten, A., der Linden, M. V., Jermann, F., & Bechara, A. (2005). Impulsivity and decision making. *The Journal of Nervous and Mental Disease*, *193*(10), 647–650.

Zuckerman, M. (1994). *Behavioral expressions and biosocial bases of sensation seeking*. Cambridge, New York: Cambridge University Press.

Zuckerman, M. (2002). Zuckerman-Kuhlman Personality Questionnaire (ZKPQ): An alternative five-factorial model. In B. de Raad, & M. Perugini (Eds.) *Big five assessment*, (pp. 377–396). Seattle: Hogrefe and Huber Publishers.

Zuckerman, M. (2007). The Sensation Seeking Scale V (SSS-V): Still reliable and valid. *Personality and Individual Differences*, *43*(5), 1303–1305.

Zuckerman, M., & Kuhlman, D. M. (2000). Personality and risk-taking: Common biosocial factors. *Journal of Personality*, *68*(6), 999–1029.

Appendix A

Comparison of Item Response Theory and Classical Test Theory Approaches

Item response theory (IRT) and classical test theory (CTT) represent two different approaches for scaling tests and dealing with psychometric issues. In brief, CTT focuses generally on the scale as a whole entity while IRT focuses on the scalability of the items. The cumulative properties of a scale suggest that if all the items are scalable under IRT then the test as a whole should perform well. In this comparison, I will focus on the following key factors in comparing IRT and CTT, especially highlighting the perceived benefits of an IRT approach. First, I will begin by discussing the assumptions of IRT and CTT. Second, I will discuss how IRT and CTT approach scale dimensionality. Third, I will deal with the issue of scale length and reliability. Finally, I will finish by highlighting how these approaches can be used in a complementary fashion.

To begin, IRT and CTT make very different assumptions about what it means to *measure* constructs. CTT starts from the assumption that tests are able to measure constructs with equal accuracy across all levels of the trait continuum. This means that if I am measuring a hypothetical construct like clinical depression, my scale works equally well for both severely

depressed individuals and individuals who lack any depressive symptoms. Naturally, this assumption is difficult to uphold because by giving a measure of *clinical depression* I am specifically targeting my scale towards individuals with some level of depression. In other words, I only care if my measure can distinguish non-depressed individuals from depressed individuals. In contrast, IRT implicitly assumes that I will be more accurate at labeling some levels of the trait continuum than others. When I am evaluating my clinical depression scale, I am not only able to determine how well it measures the average individual, but also how well it measures individuals with varying levels of depression.

A second assumption of CTT is that the responses for each item are equally good or are able to equally distinguish between various levels of a construct. One way of thinking about this statistically is that the factor loadings (or more realistically the item-total correlations) should all be similarly high. Certain IRT models do not make this claim and recognize that some items may simply be better than others and that these items ought to be given more weight. More stringent IRT models may also require that items be equally good at discrimination, but it does not merely assume this, it requires that the statistical model demonstrates this property.

The third major assumption of CTT is that the total score reflects the latent construct with a fair degree of accuracy. IRT models do not necessarily make this assumption and usually try to estimate the latent construct by considering the full pattern of responses rather than just the sum of each item. In this way, IRT gives an estimate for the individuals' level of latent construct and an estimate of how much error is expected due to inconsistent response patterns. This can be useful for determining how well each individual has been scaled by the test.

Moving on from the assumptions, IRT and CTT utilize very different approaches when looking at scale dimensionality or how a scale describes the latent construct. Current IRT theories require that scales be strictly unidimensional or in other words that there is no

cross loaded items (from a factor analytic perspective). CTT scales do not require this and it is common for a total score to be comprised of a number of individual factors that are added together haphazardly. Part of the reason that this occurs is that CTT approaches often utilize exploratory factor analysis where the psychometrician tries to obtain factors that are good enough. It is usually assumed that a given scale will have multiple subfactors in order to achieve a high enough item count for reliability. Exploratory factor analysis also can produce problematic results because it subtly encourages post-hoc identification of how a latent construct ought to be measured. I will return to this problem later. In IRT, assumptions about unidimensionality must be strictly tested. The unidimensionality assumption assures that when dealing with a latent construct only a single unitary aspect of that construct is being measured. A note of caution should be inserted at this point. This assumption requires that you have a *very clear idea* about what your latent construct is. It is, after all, possible to formulate a unidimensional scale that measures an unimportant factor very well.

The third major difference between CTT and IRT is the issue of scale length. I hinted earlier that CTT prefers long scales in order to increase reliability. CTT, by nature, assumes that the same trait must be measured over and over again in order to reliably estimate the true level of the construct. Most of the reliability statistics associated with CTT reflect this idea including Cronbach's alpha which will always increase with a greater number of items. IRT does not prefer short scales but rather prefers scales that are composed of good items. Often times this leads to shorter scales that are more accurate. Thus an IRT approach would prefer ten items that are very good predictors rather than 40 items that are a mix of good and moderate predictors. From a pragmatic standpoint this allows for decreased participant fatigue and can be utilized to construct multiple scales that have been equated for reliability.

Lastly, I would like to discuss how IRT and CTT can be used in a complementary way. The overarching goal of both approaches is to develop valid and reliable measures. Further-

more, a good measure that has been developed under IRT will also look like a good measure when tested under CTT (although the opposite may not be true). For this reason, once an IRT measure has been developed and/or refined, it can be used within the context of CTT to explore larger theoretical issues such as validity and multi-construct statistical models. This is especially important because IRT does not currently have its own set of methods for dealing with these issues. For instance, CTT has a large body of research describing how to explore the relationships among different constructs using structural equation modeling. An IRT scale should perform very well at a measurement level and can then be used to build structural models that interact with other constructs. IRT would also work exceedingly well in the case of path models because the score from an IRT measure is essentially a latent estimate that has accounted for measurement error. Essentially, when considering IRT as a measurement approach, it is important to remember that it only concerns itself with the scale and item level and does not consider the broader context. It will however deal with the scale and item level from a much more structured and refined approach than current CTT methods and this is where IRT really shines.

Appendix B

Development of the Risk-Seeking and Response-Inhibition Scales

Generally, the development of the RSRIS followed a four step process that will be elaborated here. This appendix is a condensed version of the scale development that was presented in a poster at the Washington State University Wiley Research Exposition. I will try to offer some explanation for why analysis decisions were made at each step.

1. Item Creation
2. Item Selection using Mokken Scale Analysis
3. Scale Refinement using Samejima's Graded Response Model
4. Final Scale Evaluation

B.1 Item Creation

There were a number of concerns that needed to be addressed when I created items for the new measure of impulsivity. First, I needed to develop a conceptual framework from which to write items. Secondly, I needed to consider the format of the items and their responses.

Finally, the items had to be written from either a general personality perspective or from a more specific assessment of target behaviors. I will discuss the second and third topics first as they helped to explain the larger conceptual framework.

The general format of the items for my new measure of impulsivity was patterned off of the Barratt's Impulsivity Scale. The main reason for this was simply to provide some face validity as the BIS is considered *the* measure of impulsivity. This decision also made it easy to use a seven point response system that varied from Strongly Agree to Strongly Disagree. [1] Seven or more response options makes it easier to ensure that the response options are treated as true intervals rather than just ordered categories. For IRT, this is not necessary, but in order to use structural models this is important. In the end, the items were written as self-reflective statements that the participant could either agree or disagree with (e.g., I am easily distracted).

A decision also needed to be reached as to whether the items should reflect endorsement of specific vs. general behaviors and attitudes. Using specific behaviors such as 'I have trouble controlling my spending' is attractive because it establishes better criterion validity than statements that are more generalized such as 'I like to take risks'. Using specific behaviors has a number of important drawbacks though. First, it forces the researcher to narrow impulsivity down to a set of behaviors that may or may not share a common underlying psychological construct (this presents a dimensionality problem). It also requires that the participant endorse what could be considered socially inappropriate behaviors and therefore might introduce response bias. For these reasons, I chose to use more generalized statements in the hope of capturing a more general personality construct rather than a set of specific problematic behaviors.

[1] I should mention that I made one poor decision at this point. As a middle point, I had a number of options including statements like 'neutral', 'undecided' or 'neither agree nor disagree'. Of course I chose the worst possible option - undecided. This became evident in the Item Characteristic Curves during the IRT analysis. I subsequently changed the middle option to neutral and the problem has resolved itself.

This leads to the overall conceptual framework that was used to develop the new scales. Theoretically I was trying to measure the two factors of impulsivity described by empirical research. In this way, a pool of items was written to reflect the various aspects of sensation seeking or reward seeking behavior. In particular, items asked about participants attitudes towards risk, avoiding boredom and seeking out new experiences. A separate pool of items was written to address response inhibition or impulse control problems. These items asked participants about their attention, distractibility, patience, racing thoughts and snap decisions. For a complete list of items that were generated, refer to Appendix C.

B.2 Item Selection

All the items were administered to a group of 613 participants in Fall of 2006. In addition to the new items, participants also completed several other questionnaires. Initially, items were screened for basic descriptive properties such as mean, standard deviation, skewness and kurtosis. Most of the items were fairly normally distributed with one exception; a single item in which no participant endorsed strongly agree.

At this point, a CTT approach would run an exploratory factor analysis to identify an acceptable factor structure. Remember that I had made the decision a priori to have two unique factors. Rather than utilize an exploratory analysis, I therefore chose to utilize a non-parametric IRT technique called Mokken Scale Analysis. Essentially, Mokken Scale Analysis constructs unidimensional scales by searching for items with similar covariance until a specified cutoff. It is an exhaustive approach that will continue to construct scales until all the items have been placed into a scale. A huge selling point for this technique is that it automatically eliminates items that overlap with many scales. Thus the resulting scales are completely unidimensional.

Like factor analysis, it is important to examine the scales that are constructed to make

sure that you are in fact measuring the construct you set out to scale. The results of the analysis indicated that I had three strong scales: One scale of 22 items from the risk seeking pool, a second scale with three items from the response inhibition pool and a third scale with eight items from the response inhibition pool. The three item scale was discarded because it was a) too short and b) contained items that intuitively did not relate to the central construct of response inhibition. The other two scales were retained for further analysis.

Mokken Scale Analysis also contains methods for assessing two other assumptions of IRT scales: monotonicity and non-intersection. In particular, the monotonicity assumption must be met and the items in both scales met this assumption without issue. Nonintersection is important for determining whether to use a strict Rasch IRT model that assumes the items are equally good predictors. This assumption is often hard to meet and I had previously planned to use a non-Rasch model. It was therefore not surprising when non-intersection was not demonstrated.

B.3 Scale Refinement

IRT excels at taking scales and providing copious amounts of information about the properties of the individual items. This is useful for scale development because it provides a tool for determining which items to include in a final scale and which to discard. In particular, researchers are often interested in determining two things about items. First, what range of the latent continuum does an item accurately measure. This is referred to as the item's *difficulty*. Secondly, how well does an item *discriminate* between various levels of a latent trait. Discrimination is especially important for determining which items are more accurate at measuring a latent construct and this was the parameter that was primarily used in determining which items to retain for the final scales. Table B.1 and Table B.3 display the discrimination(α) and difficulty(β) parameters for each scale.

Table B.1: Item Parameters for the RS scale

Item	β_1	β_2	β_3	β_4	β_5	β_6	α
RS2	-3.075	-2.408	-1.630	-1.317	-0.159	1.412	2.015
RS3	-4.157	-2.154	-1.182	-0.697	0.494	1.903	1.740
RS4	-3.934	-2.817	-1.738	-1.060	0.276	1.638	1.645
RS5	-3.299	-2.009	-0.805	-0.375	0.826	2.294	1.847
RS6	-2.901	-1.988	-1.215	-0.726	0.426	2.003	1.514
RS7	-3.942	-2.767	-2.028	-1.341	-0.254	1.051	1.806
RS8	-3.000	-1.749	-0.286	0.244	1.387	2.876	1.538
RS9	-3.862	-3.464	-2.378	-1.813	-0.432	1.109	1.557
RS11	-2.617	-1.364	-0.513	0.141	1.318	2.258	1.764
RS12	-2.511	-1.457	-0.623	-0.032	0.967	1.771	2.453
RS15	-2.496	-1.131	-0.302	0.343	1.414	2.578	1.537
RS17	-3.338	-2.031	-0.794	-0.157	1.203	2.711	1.220
RS18	-2.334	-1.008	-0.194	0.409	1.532	2.474	2.128
RS19	-2.957	-1.792	-1.172	-0.672	0.759	1.929	1.751
RS20	-3.951	-2.796	-2.001	-1.316	0.134	1.674	1.625
RS21	-3.606	-2.062	-0.891	-0.391	0.696	2.140	1.374
RS28	-4.586	-2.860	-1.039	-0.161	1.540	3.418	1.082
RS30	-5.220	-3.475	-1.659	-0.611	0.929	2.949	1.126
RS31	-3.592	-2.487	-1.044	-0.395	0.970	2.685	1.464
RS32	-4.863	-3.564	-2.449	-1.519	0.091	1.748	1.395

Table B.2: Item Parameters for the RI scale

Item	β_1	β_2	β_3	β_4	β_5	β_6	α
RI1	-2.260	-0.874	-0.294	-0.148	1.005	1.906	2.738
RI2	-1.177	1.560	3.082	3.302	4.288	6.062	1.027
RI3	-2.766	-1.406	-0.603	-0.290	0.826	1.832	2.412
RI4	-2.828	-1.032	-0.127	0.377	1.901	3.111	1.172
RI5	-3.156	-1.414	-0.477	0.080	1.206	2.530	1.623
RI9	-2.527	-0.891	0.281	0.945	2.158	3.871	1.633
RI15	-3.853	-1.615	-0.555	-0.031	1.756	3.889	1.167
RI19	-3.351	-1.842	-0.418	0.815	2.183	4.096	1.347

The rule-of-thumb criterion for deciding between what constitutes a good item and a bad item is typically whether the α parameter is greater than 1.0. This was the criteria that I used and as you can see all of the items in each scale exceeded this standard. However, refer to item RI2. The item's difficulty parameters suggest that it is only measuring individuals who have higher levels of response inhibition problems. This would not matter for a typical IRT analysis, but it creates an extremely skewed item distribution which makes it difficult to use in CTT analysis. This item was therefore discarded using this rationale.

B.4 Final Scale Evaluation

With two sets of good items, a decision needed to be made. Since the RS scale contained a large number of good items, and the RI scale contained a small number, I could either maintain two scales of unequal length or par down the RS scale to match the seven items in the RI scale. As in CTT, more items means more information (or reliability) and the question then became how much information would be lost by going to a smaller scale. To determine this, I first needed to examine how much information I was starting with and how much would be lost if I only used the seven best items from the RS scale. To do this, Total Information Plots were obtained as seen in Figure B.1.

At this point, a lot of rationalization happened and for some reason I opted to drop the RS scale down to the seven best items. This was a mistake, clear and simple. The rationale for this mistake is understandable however. First, I wanted to keep the scales as short as possible. This was to be a rapid measure so that it could be used in the context of large experiment protocols without taking too much time. Secondly, the RS scales still provided more information than the RI scale even when it's length was dropped from 22 to 7 items. Lastly, no one likes an unbalanced scale (of course no one likes invalid scales either...).

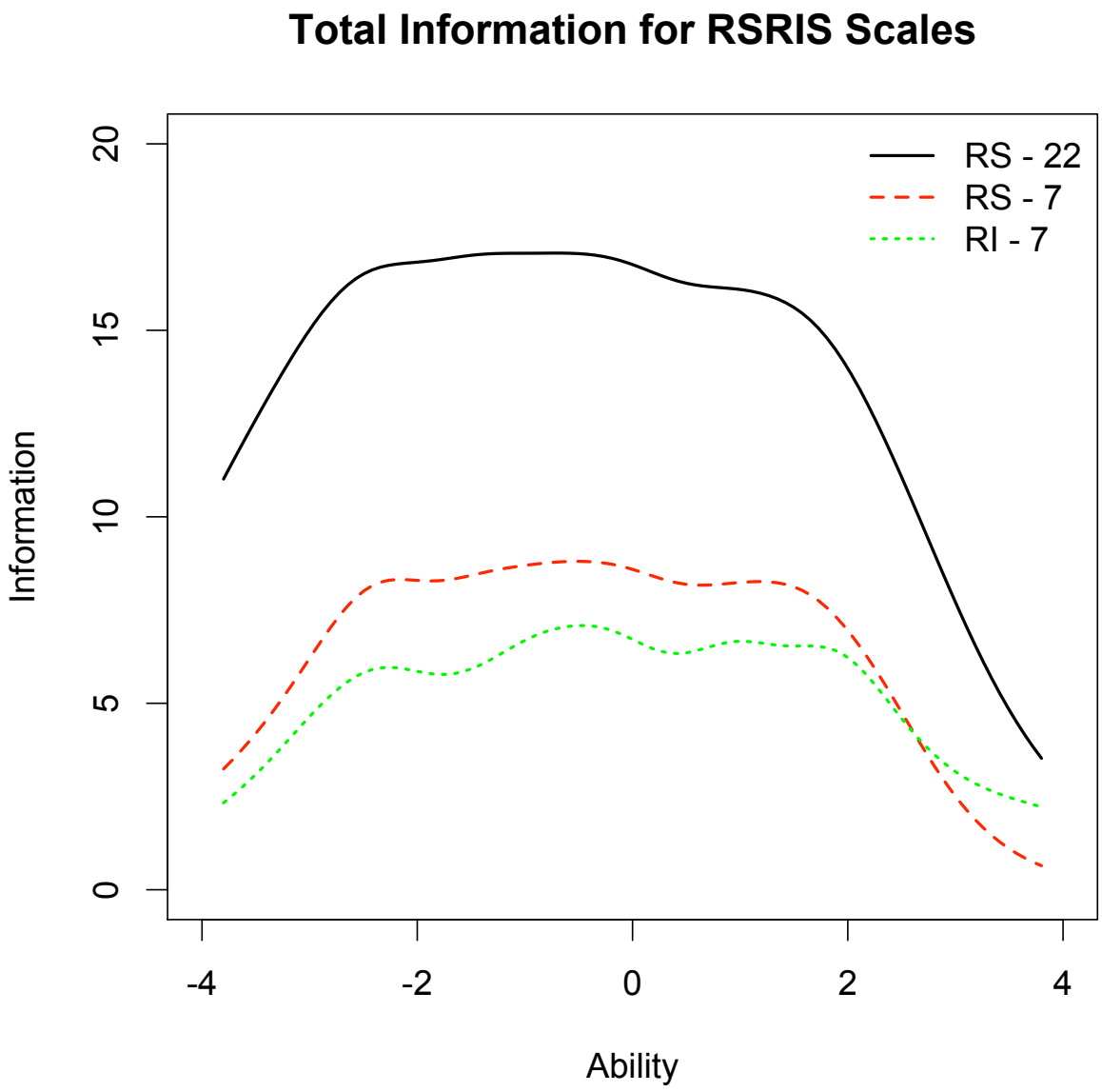

Figure B.1: Total Test Information for the RI scale (7 items) and the RS scale (22 & 7 Items)

To put into perspective the magnitude of this decision, let me share some numbers. When estimating factor scores, in addition to a θ estimate, the expected standard error of that estimate is also given. This standard error corresponds directly to the reliability of the measurement for each individual. Using the 22-item RS scale, the average standard error in the calibration sample (N=613) was 0.241. Using the 7-item RS scale, the standard error rises to 0.326. Do a little math and this suggests 1.35 times more error in the shorter form. This was giving up a lot of reliability.

Overall, the information obtained from both the RS and RI scales (at seven-item length) is not 'bad'. A score called the separation index can be calculated by using the formula $\frac{Var[\hat{\theta}]-Var[\epsilon]}{Var[\hat{\theta}]}$. This scores is analagous to Cronbach's α. The separation index for both these scales is a very high 0.99. This consistency is extremely high because all of the items are individually reliable. This demonstrates the utility of the IRT approach. Of course, this says nothing of the validity of the scales, but that is for another study.

In conclusion, the development of these scales utilized some of the most modern techniques available to psychometrics. There were some large errors made, but the scales were still quite reliable.

Appendix C

Complete list of the original RSRIS Items

This appendix contains the original 68 items from the Response Inhibition and Risk Seeking Scales.

Response Inhibition Items

1. I often have trouble paying attention.
2. I can focus on tasks when I need to.
3. I am easily distracted.
4. I drift off during long conversations.
5. I have difficulty tuning out distractions.
6. I have trouble slowing my mind down sometimes
7. I tend to get carried away easily
8. I have trouble keeping track of things
9. I have a really good attention span
10. I sometimes answer people before they are finished talking
11. I am quick to make decisions

12. I like to follow my initial instinct
13. I dislike waiting
14. I try not to let my thoughts stray when I am working
15. I sometimes lose track of what I am doing
16. I sometimes make mistakes because I work too fast
17. I usually take my time before I make a decision
18. I often get impatient
19. I am able to focus better than most people
20. I think that most people move too slowly
21. I often wish others would hurry up
22. I have trouble keeping my hands still
23. I often find myself twitching my leg or playing with my hands to keep busy
24. I need to keep busy or I get bored
25. I dont like spending a lot of time on one task
26. I prefer to multitask rather than working on one task at a time
27. I think it is important to pace yourself when you work
28. I have trouble sitting still for long periods of time
29. I feel that I sometimes have problems with self-control
30. I sometimes make mistakes because I act without thinking
31. I usually make choices based on what seems easy or quick
32. I prefer a quick answer to a long thought out response
33. I dont mind making a few mistakes as long as I work as quickly as possible
34. I try not to make quick decisions

Risk Seeking Items

1. I am a cautious person.
2. I enjoy taking risks from time to time.
3. I often do things on the spur of the moment.
4. I like to live in the moment.
5. I try to avoid taking risks.
6. I can be a 'wild' person sometimes.
7. I want to live an adventurous life
8. I like to play it safe
9. I like seeking out new experiences
10. I enjoy meeting people from all walks of life
11. I want to live fast and hard
12. I am always looking for a new thrill
13. I dont like people who are stuck in their ways
14. I like to drive fast
15. I am an adrenaline junkie
16. I think it would be fun to skydive
17. I like to keep changing things in my life
18. I live on the edge of life
19. I like a little danger every once in a while
20. I like events that make me feel alive
21. I dont get intimidated by new experiences
22. I really like it when I get to try new things
23. I prefer to stick to what I know works best rather than try new things
24. I really dont want to live a boring life
25. I would be happy with living a simple life

26. I dont really worry about consequences
27. I really worry about getting hurt
28. I think it is really important to play it safe
29. I think that a lot of people take too many risks
30. I am wary of new experiences
31. I find it better to be cautious in life
32. I like to try out a wide variety of experiences
33. I buy things on impulse
34. I like instant gratification

Appendix D

Final RSRIS Scale

Items 1–7 are part of the *Risk Seeking Scale*. Items 8–14 are part of the *Response Inhibition Scale*. Items with a * are reverse coded.

1. I enjoy taking risks from time to time.
2. I try to avoid taking risks.*
3. I want to live an adventurous life.
4. I want to live fast and hard.
5. I am always looking for a new thrill.
6. I live on the edge of life.
7. I like a little danger every once in a while.
8. I often have trouble paying attention.
9. I am easily distracted.
10. I drift off during long conversations.
11. I have difficulty tuning out distractions.
12. I have a really good attention span.*
13. I sometimes lose track of what I am doing.
14. I am able to focus better than most people.*

Appendix E

Supplementary graphs

E.1 Histograms for Study 1 – Response Inhibition

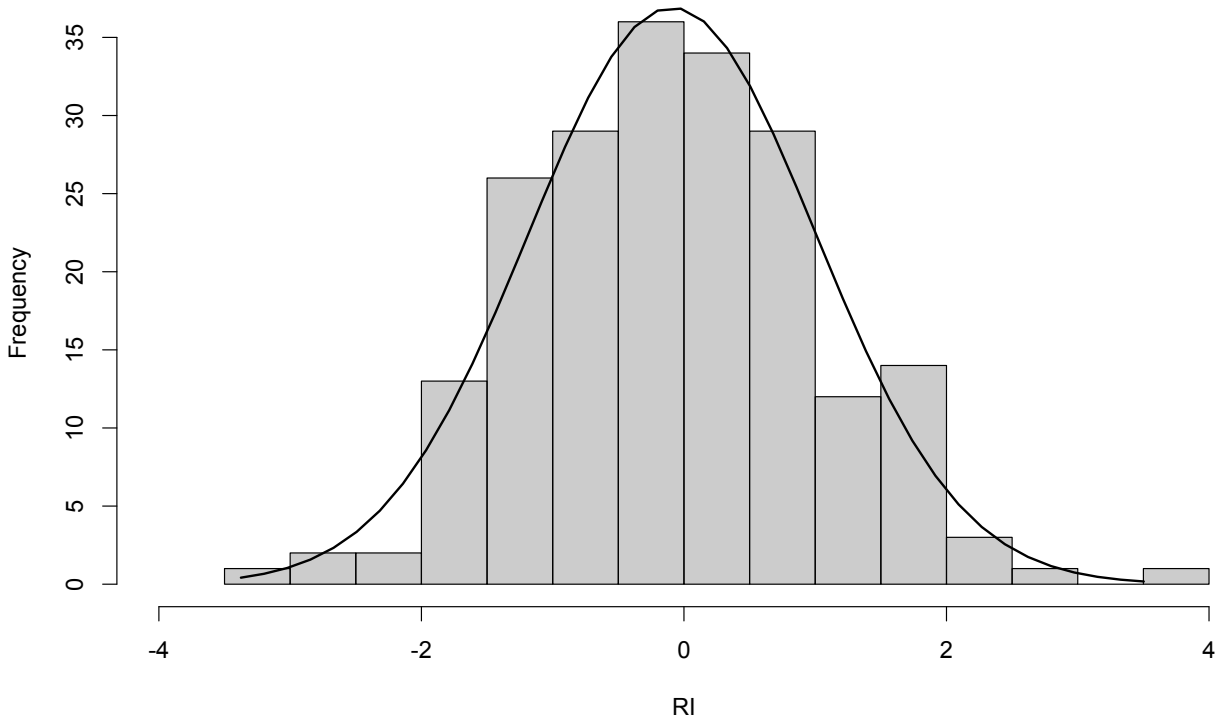

Figure E.1: Histogram for RI Scale

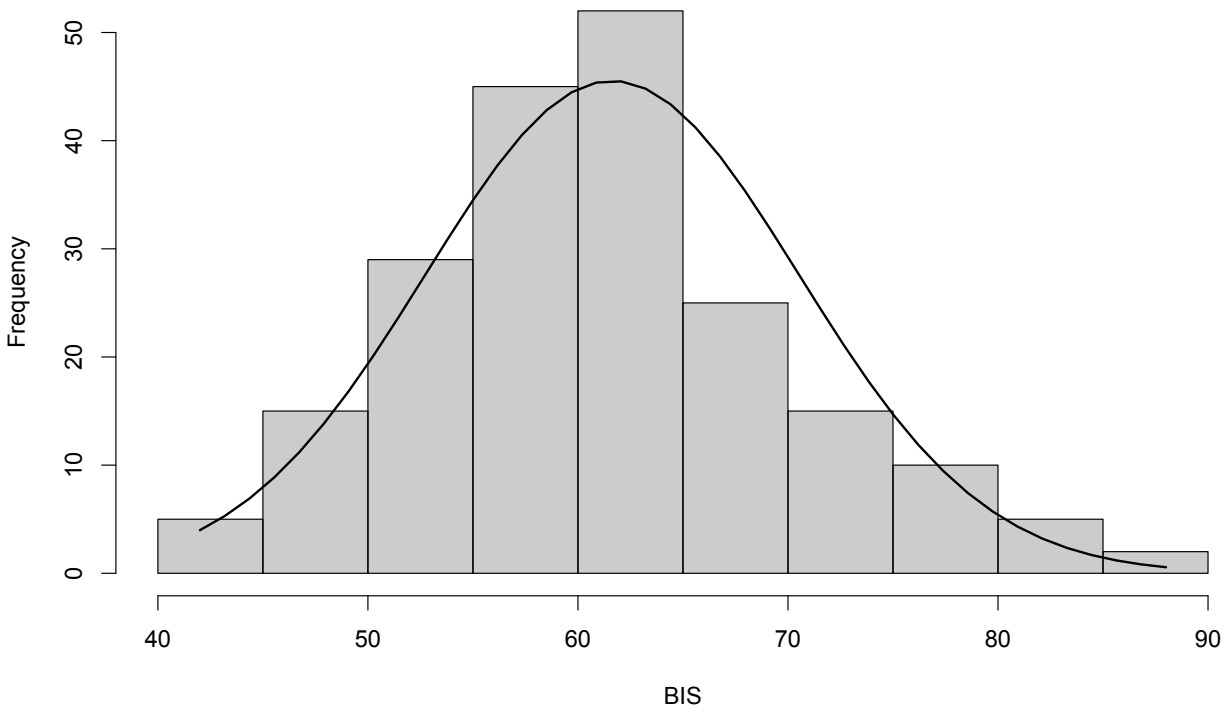

Figure E.2: Histogram for BIS

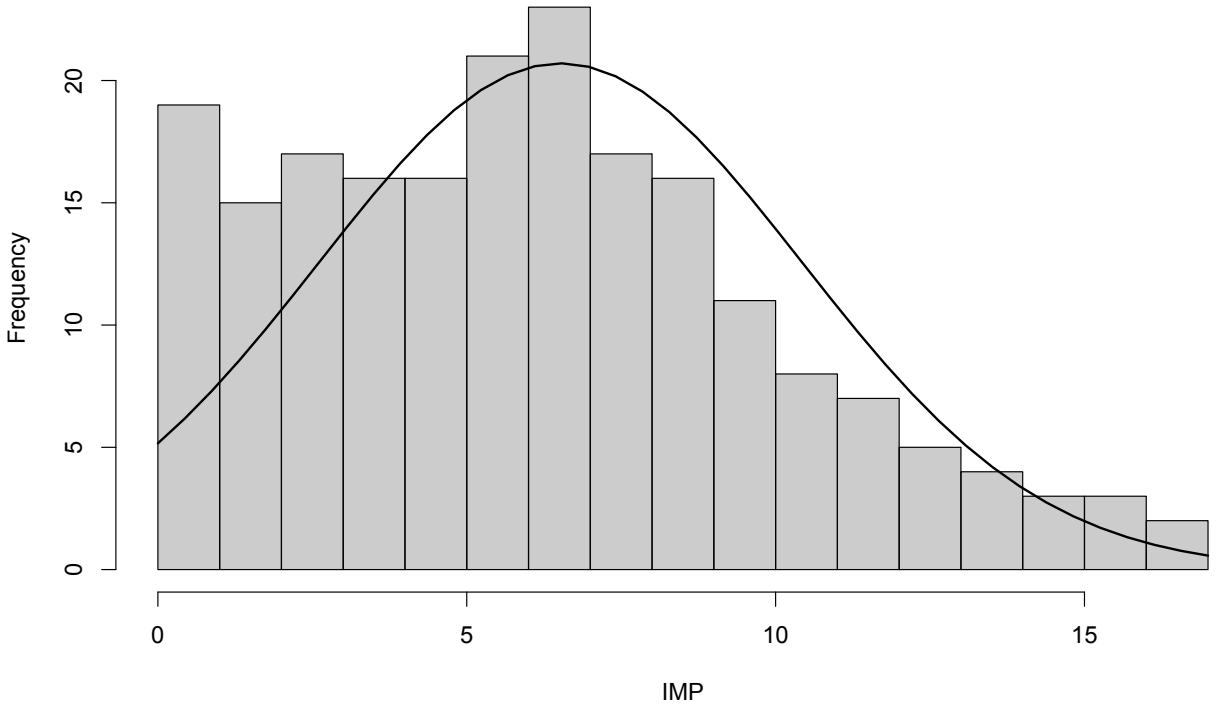

Figure E.3: Histogram for EIS Impulsivity

E.2 Histograms for Study 1 – Reward Seeking

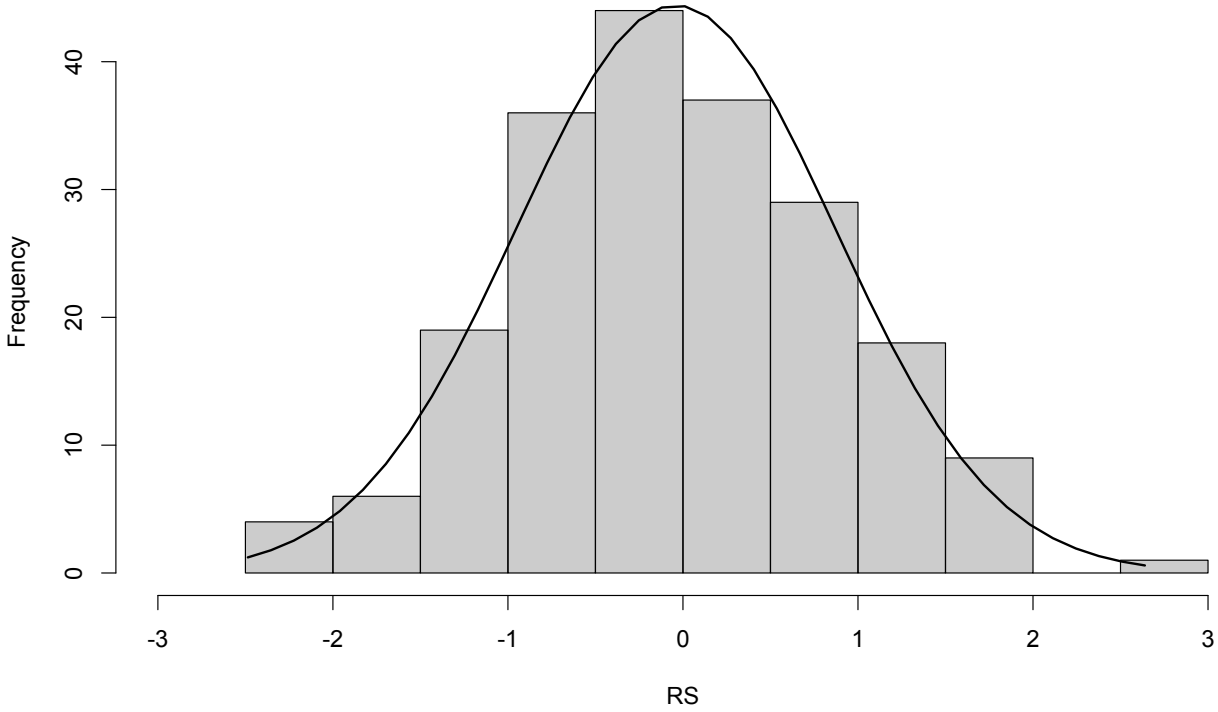

Figure E.4: Histogram for RS Scale

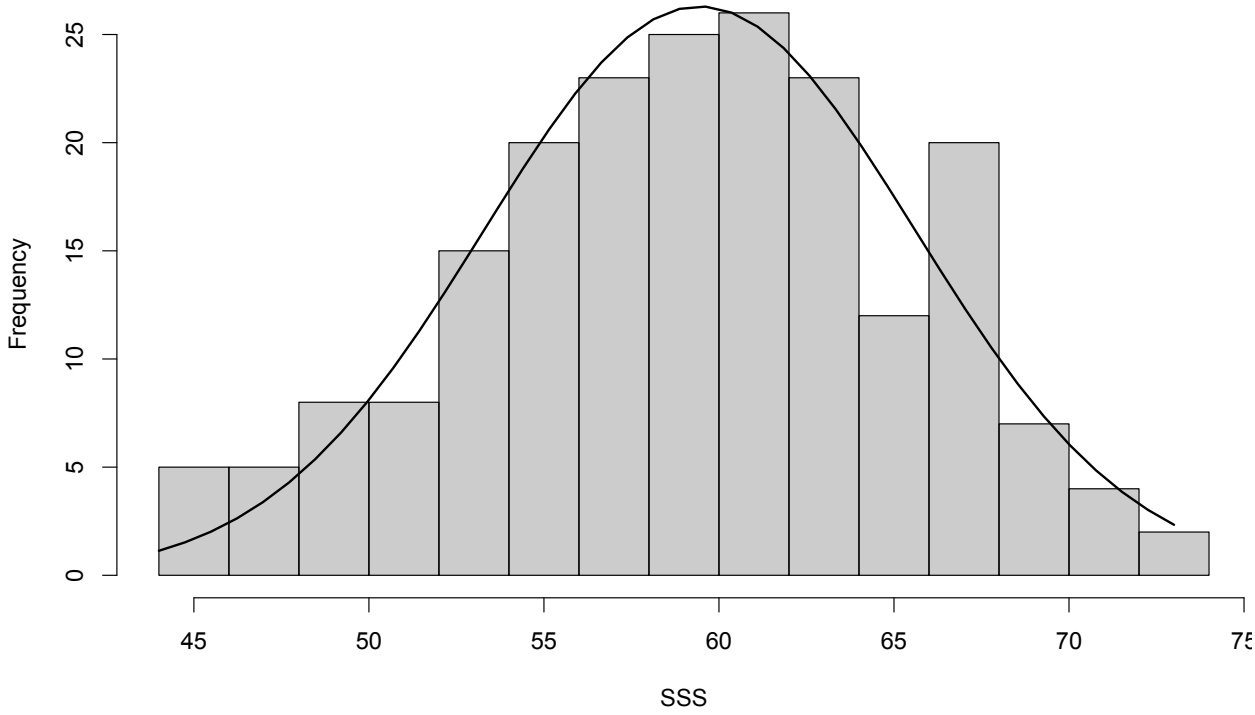

Figure E.5: Histogram for SSS

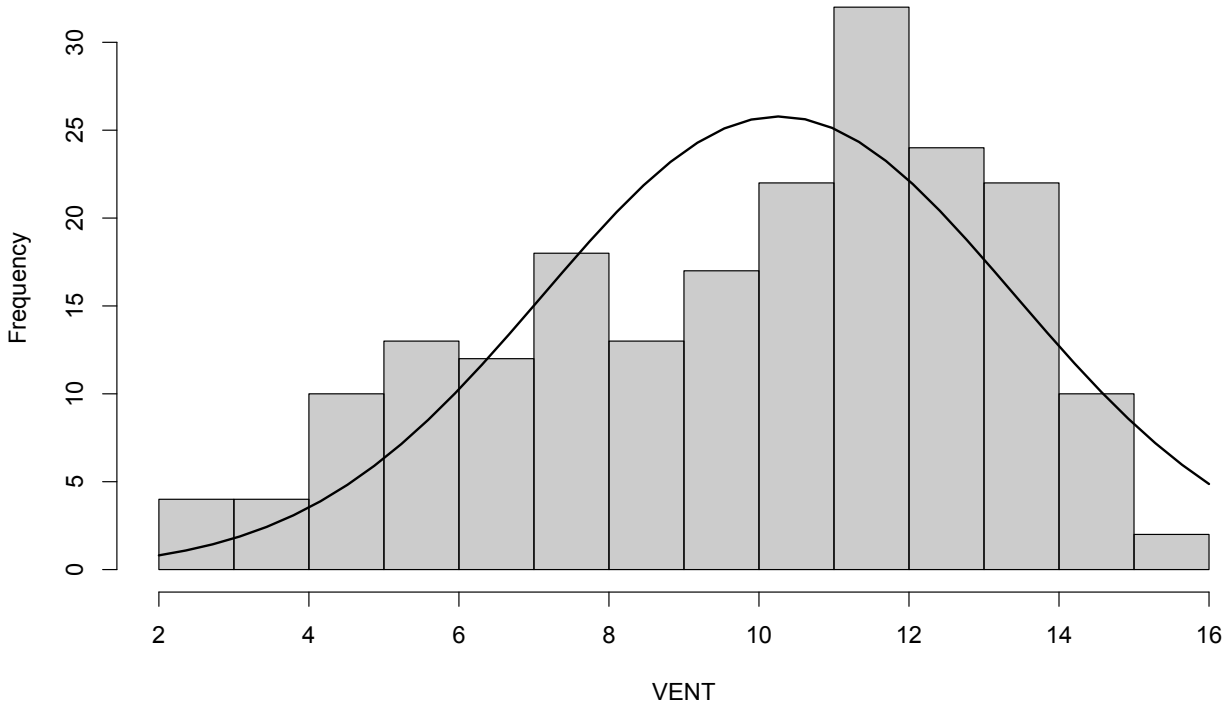

Figure E.6: Histogram for EIS Venturesomeness

E.3 Histograms for Study 1 – Problem Behaviors

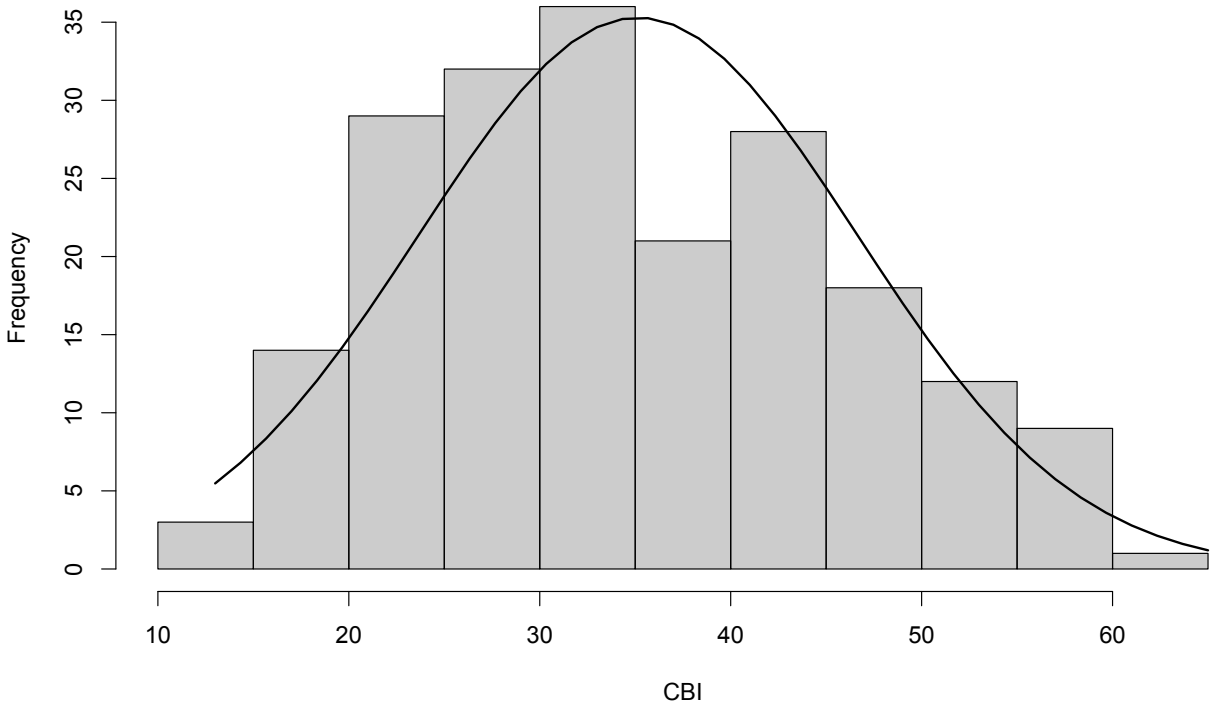

Figure E.7: Histogram for CBI

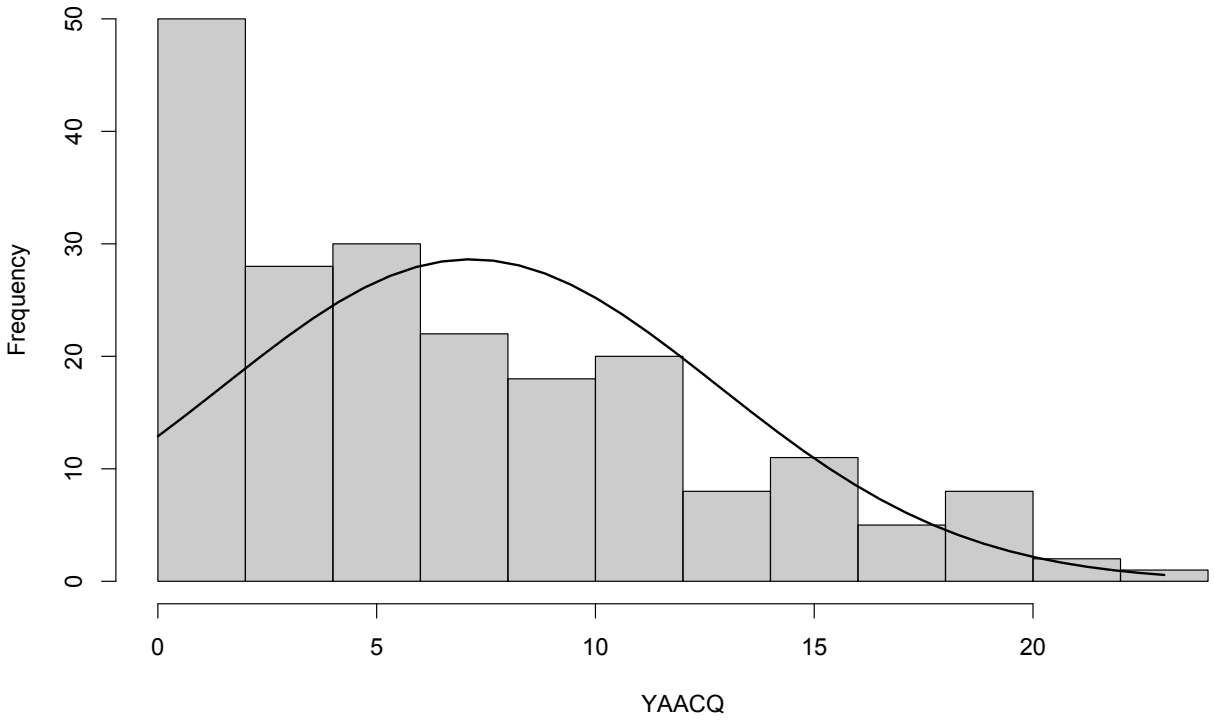

Figure E.8: Histogram for YAACQ

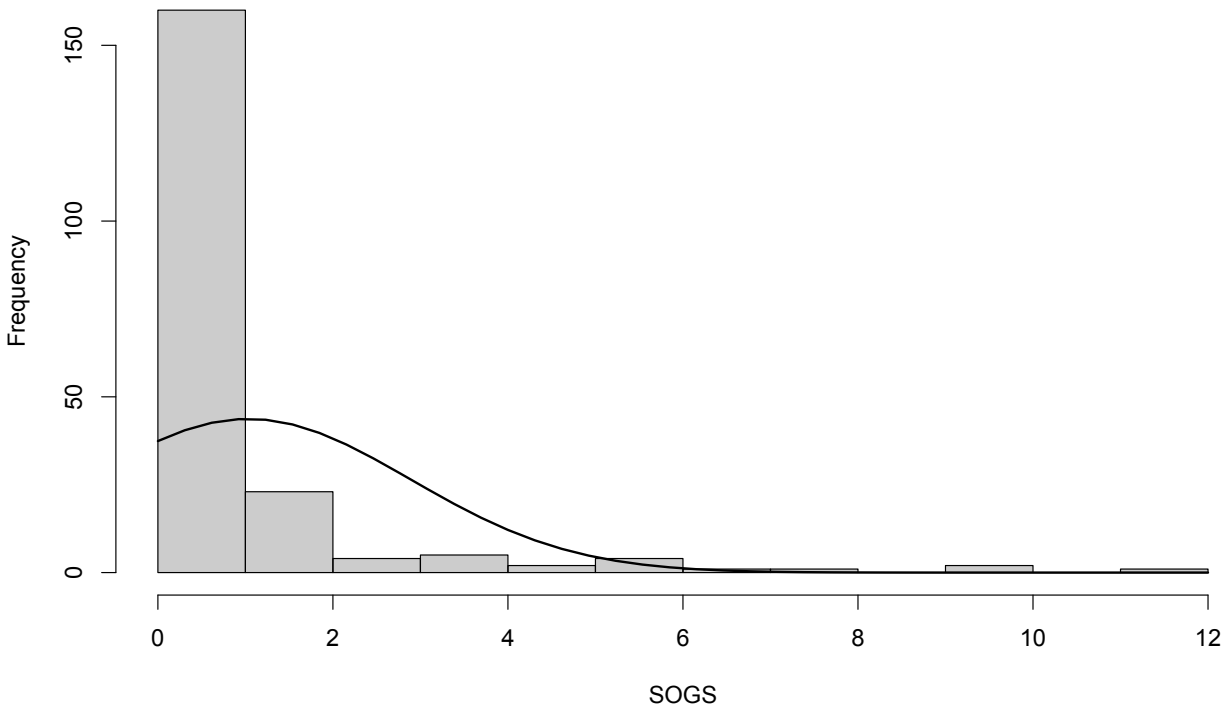

Figure E.9: caption

E.4 Correlation Between RS and SOGS

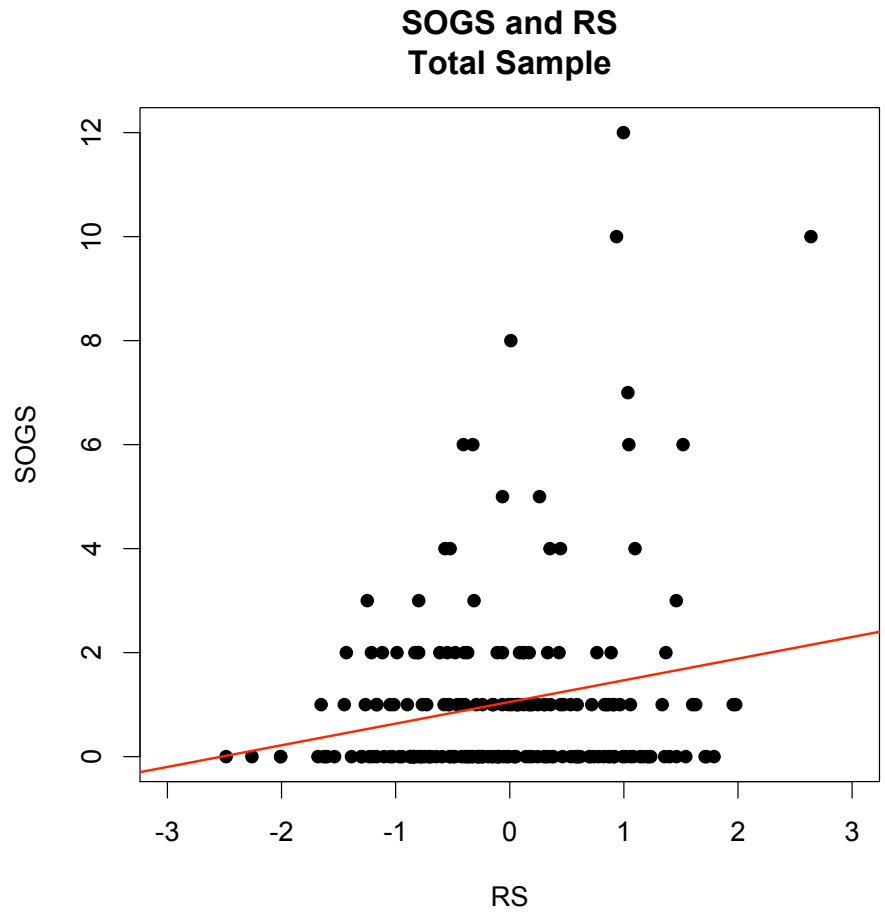

Figure E.10: Scatterplot depicting the correlation between the SOGS and the RS in the full unrestricted sample. Note the floor effect for the SOGS.

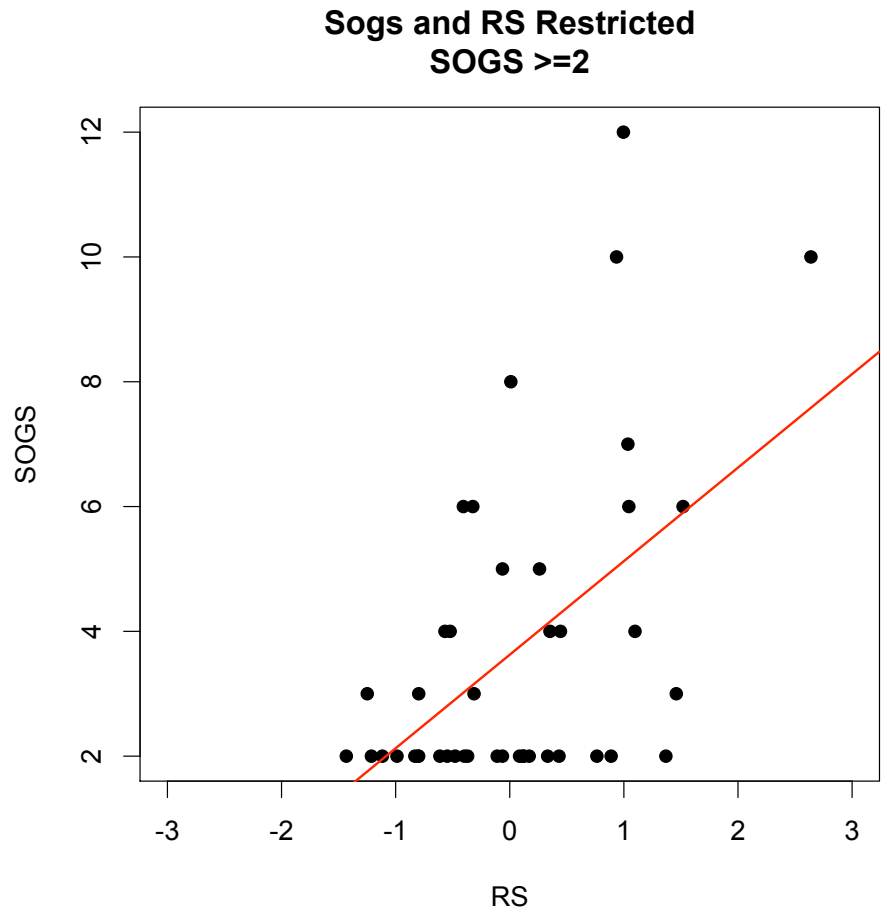

Figure E.11: Scatterplot depicting the correlation between the SOGS and the RS when the sample is restricted to only include people who meet criteria for possible pathological gambling.

E.5 IGT Performance

Figure E.12 and Figure E.13 display IGT performance in two separate samples collected at Washington State University. In both cases, performance is much lower than what is reported by other authors with much smaller samples. This raises some concerns about what 'Bad' performance on the IGT means. Some authors suggest that performance less than 50% indicates probable frontal lobe dysfunction which would apparently include most of these participants.

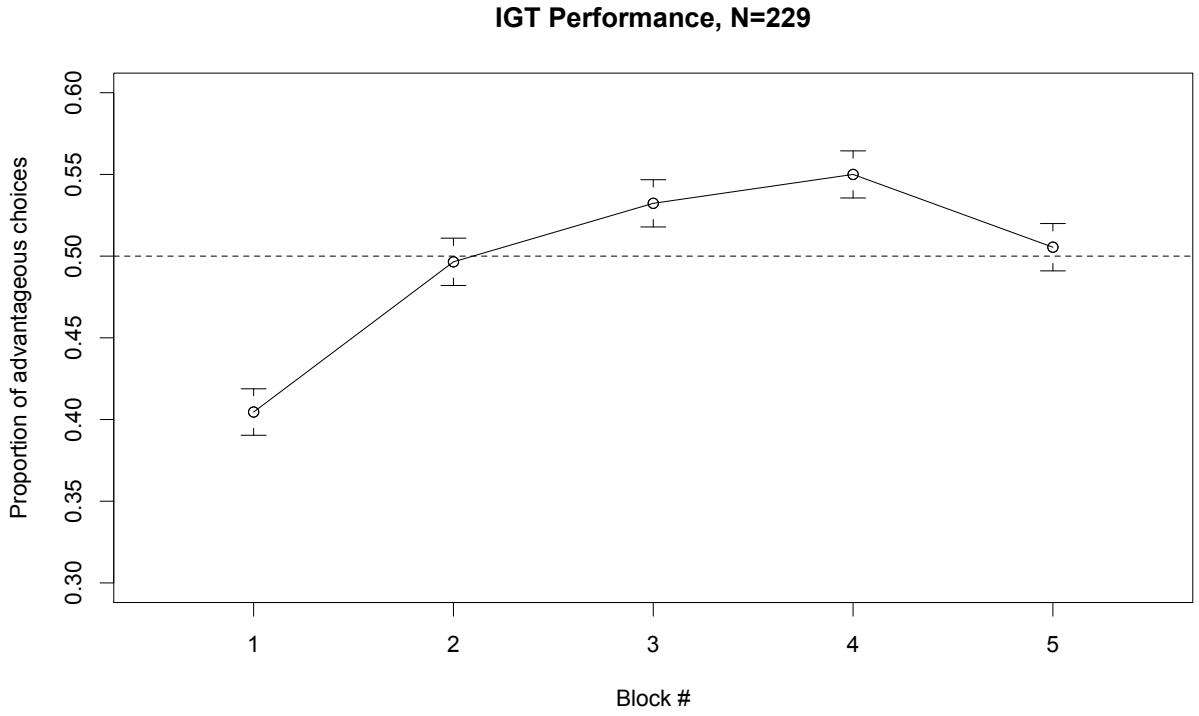

Figure E.12: IGT Performance in the current study collected in early 2009. Error bars indicate the 95% confidence interval. This data includes only the standard IGT.

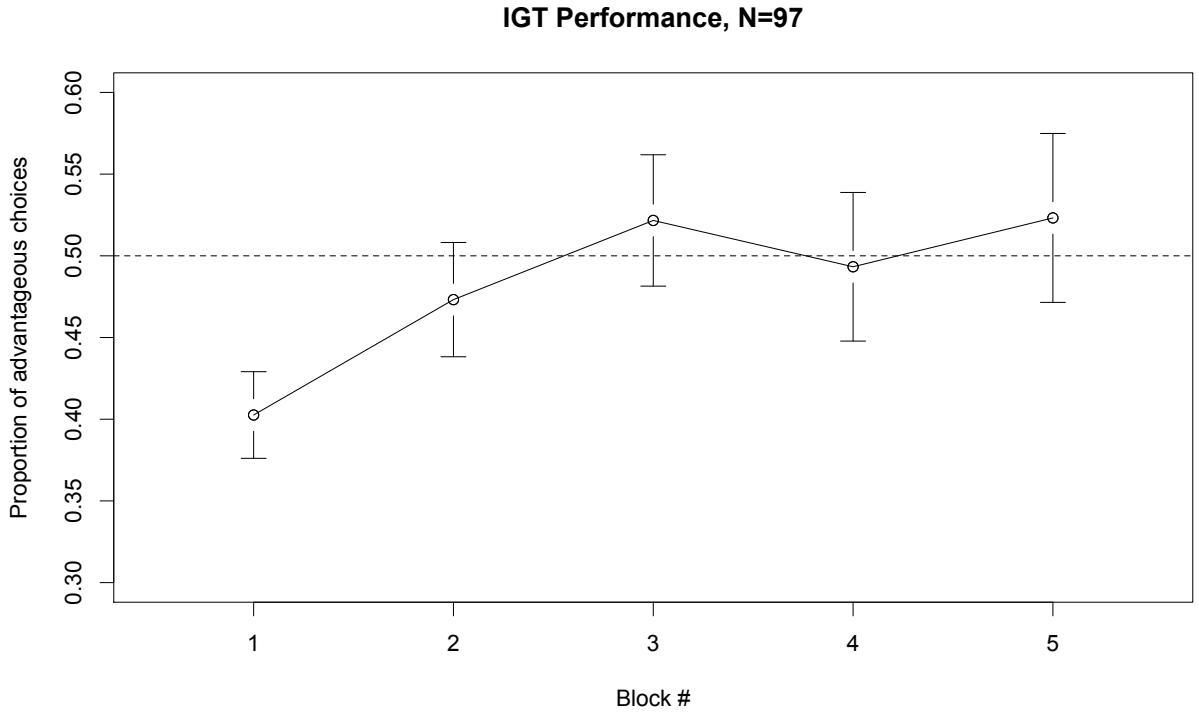

Figure E.13: IGT Performance in the author's masters thesis collected in Fall 2005 and Spring 2006. Error bars indicate the 95% confidence interval. This data includes both the standard IGT and a modified IGT with a late reversal.

Appendix F

A brief IRT analysis of the scales in Study One

F.1 Impulsivity Measures

Although these analyses are extremely preliminary, they do tell a very interesting story. To accomplish this, the BIS, SSS, IMP and EIS were subjected to a rudimentary IRT analyses using all of the items available for each of the scales. No assumptions of unidimensionality, monotonicity or nonintersection were examined but they aren't important for demonstrating my point. Total information plots were obtained to look at each of the scales ability to predict impulsivity traits across the full latent continuum. These plots do not reflect a common factor, so it is likely that impulsivity is defined differently for each of the scales (i.e., the scales are not interchangeable). Essentially, this describes how well each of these scales measures individuals from low levels ($\theta = -4$) to high levels ($\theta = +4$). For those unfamiliar with IRT, ability roughly correlates to z-scores. Information corresponds to the inverse of standard error, so higher information means less error.

Generally speaking, the scales that use Likert-type multiple response options do a better job of measuring the full range of the latent continuum. This is not surprising. What is also

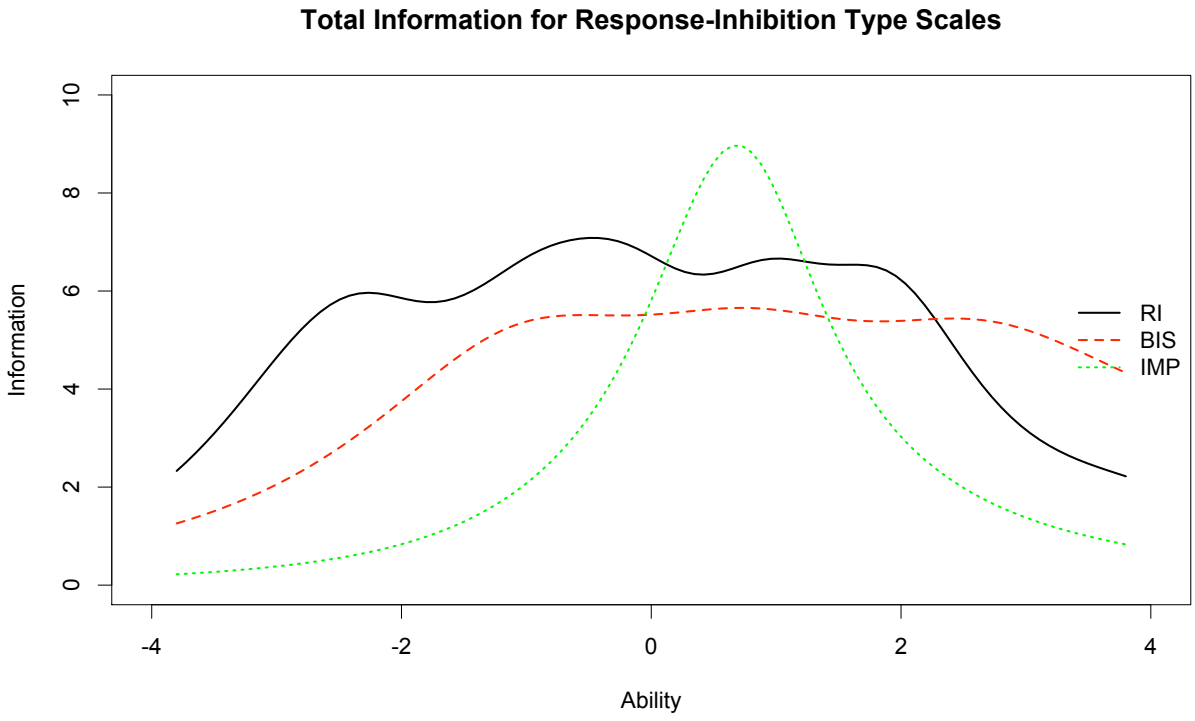

Figure F.1: Comparison of RI, BIS and IMP under IRT parameterization

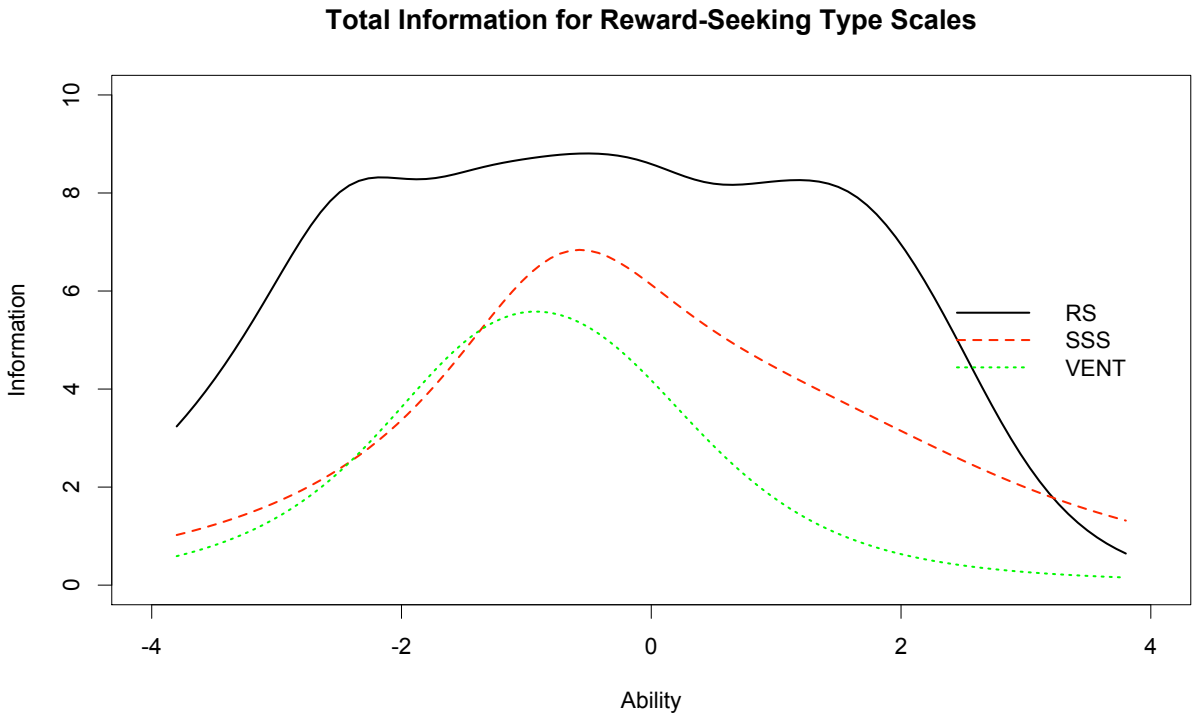

Figure F.2: Comparison of RS, SSS and VENT under IRT parameterization

not surprising is that the measures developed under IRT have more information across the full continuum. Referring to Figure F.1 we can see that the RI and BIS both do a good job across varying levels of impulsivity with the RI edging out the BIS by a slim margin. In Figure F.2, the RS clearly excels at measuring all levels of the latent trait when compared to the SSS or VENT. Lastly, Figure F.3 shows that the IMP has good information but that this information is only high in a very narrow range of the latent continuum. In this case, the IMP is good at measuring people who are slightly more impulsive than average, but not good at measuring people outside of this region. This figure probably accounts for why the RS does a better job of predicting the SOGS for the restricted sample.

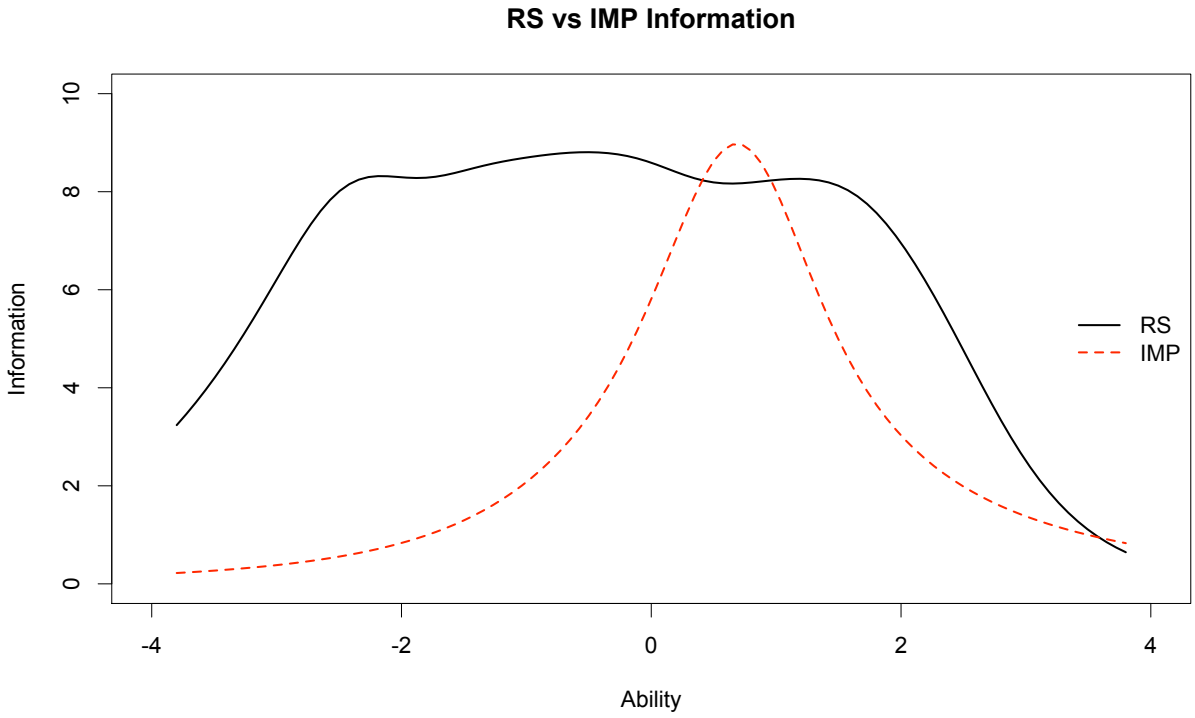

Figure F.3: Comparison of RS and IMP under IRT

F.2 Behavioral Outcomes

The last interesting thing that can be learned is how well the outcome measures perform. The SOGS (see Figure F.4) does an exceedingly good job of discriminating between people with low vs high levels of gambling problems. This means that the SOGS is *NOT* precise when it comes to giving people a score at various levels of the trait continuum but that it is *very* accurate in the range which is diagnostically relevant. A quick note, the y scale is not mistaken, the information is nearly 2000 in that particular range.

The CBI and YAACQ both perform quite well from an IRT perspective as seen in Figure F.6 and Figure F.5. Both scales do a particularly good job of identifying people from 0 to +4. This means that the scales were designed to measure higher levels of problem behaviors and that they will not do a good job for individuals who have fewer problems.

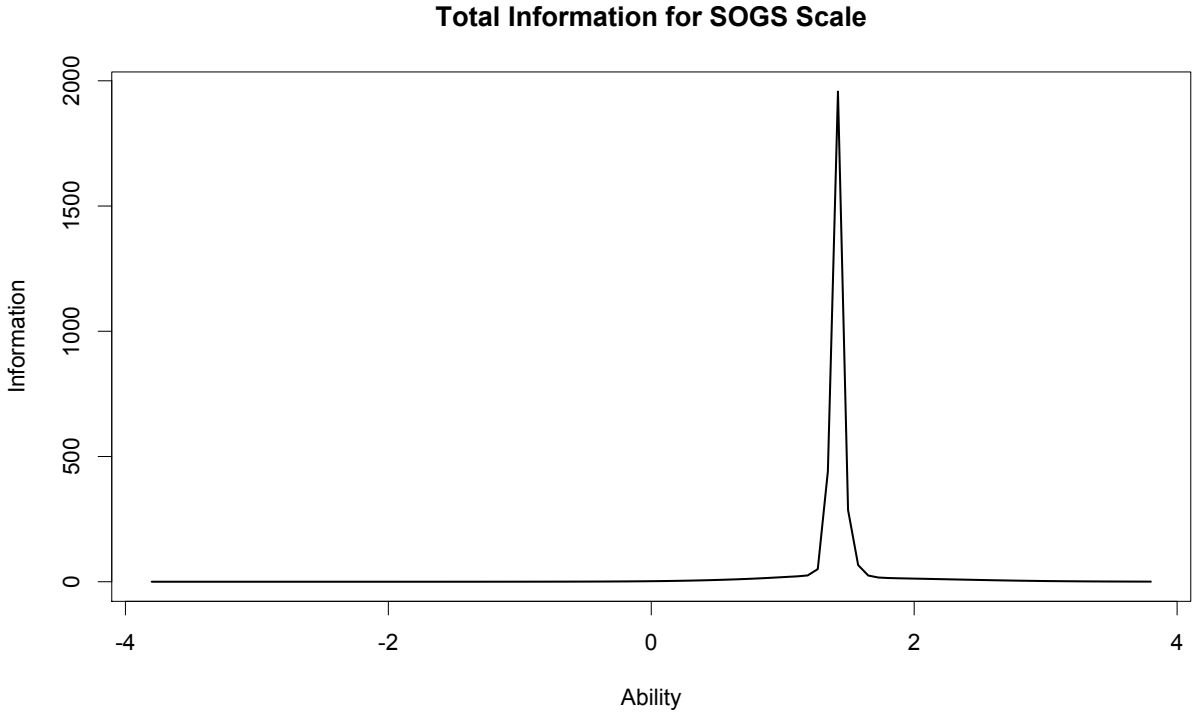

Figure F.4: SOGS Information Under IRT

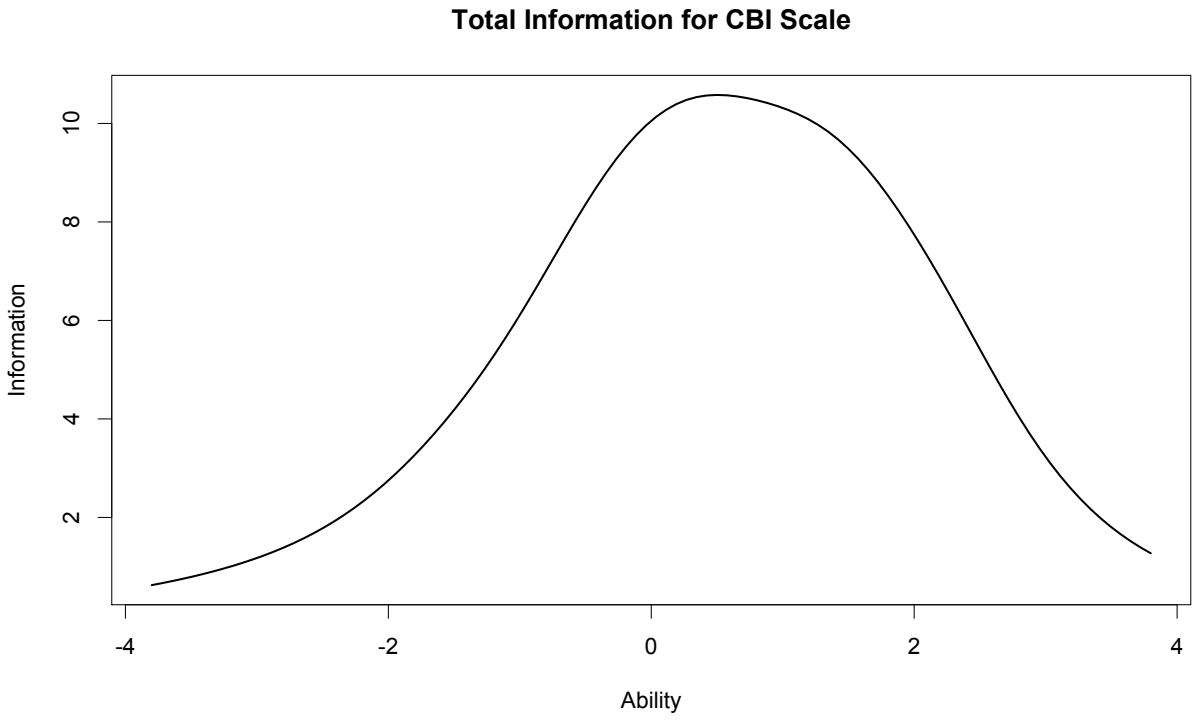

Figure F.5: CBI Information under IRT

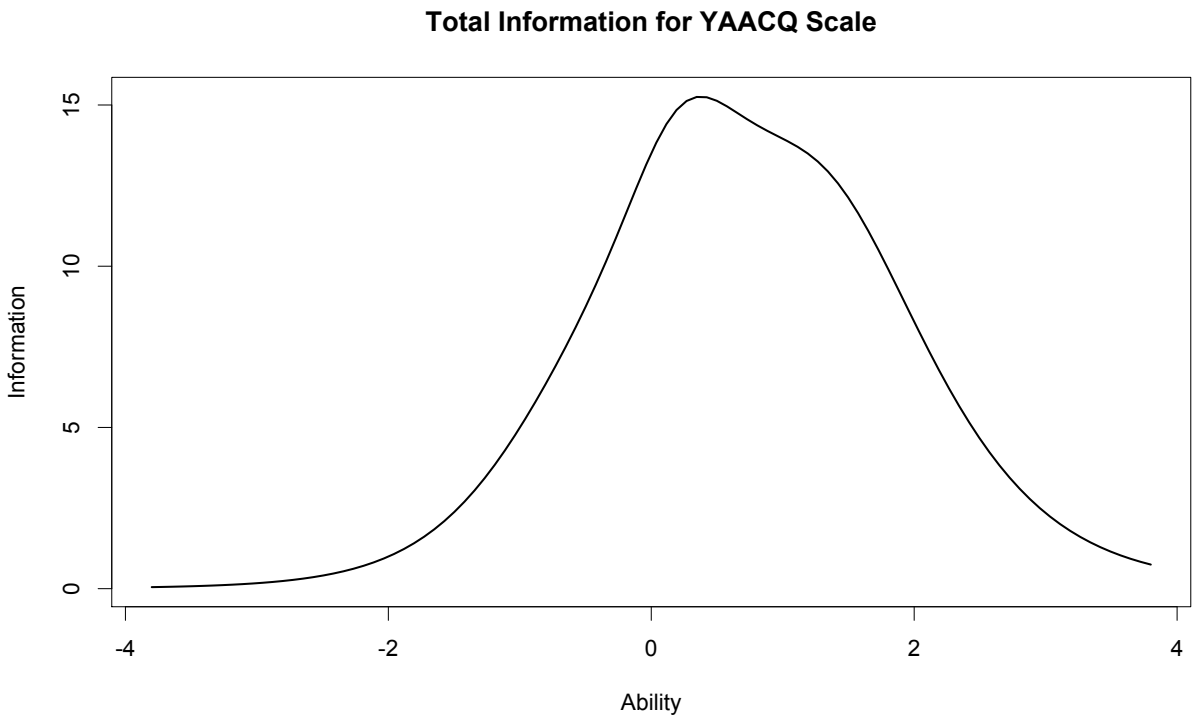

Figure F.6: YAACQ Information under IRT