THE DEVELOPMENT AND VALIDATION OF THE AUBURN PSYCHOLOGY

TERM TEST (APTT)

Except where reference is made to the work of others, the work described in this thesis is my own or was done in collaboration with my advisory committee. This thesis does not include proprietary or classified information.

Dale L. Smith

Certificate of Approval:

Martha Escobar
Assistant Professor
Psychology

Lewis Barker, Chair
Professor
Psychology

Adrian Thomas
Associate Professor
Psychology

Stephen L. McFarland
Acting Dean
Graduate School

DEVELOPMENT AND VALIDATION OF THE AUBURN PSYCHOLOGY TERM TEST (APTT)

Dale L. Smith

A Thesis

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Master of Science

Auburn, Alabama
August 7, 2006

DEVELOPMENT AND VALIDATION OF THE AUBURN PSYCHOLOGY

TERM TEST (APTT)

Dale L. Smith

____

Signature of Author

____

Date of Graduation

THESIS ABSTRACT

DEVELOPMENT AND VALIDATION OF THE AUBURN PSYCHOLOGY

TERM TEST (APTT)

Dale L Smith

Master of Science, August 7, 2006
(B.S., Olivet Nazarene University, 2001)

94 Typed Pages

Directed by Lewis Barker

The construction and investigation of the psychometric properties of the Auburn Psychology Term Test (APTT), a yes-no test designed to measure psychology knowledge, is described in this paper. The relationships between this instrument and more typical indicators of student performance, including students' ability to identify and define psychology vocabulary items, and students' introductory psychology course grade, was significant. Strong alternate form reliability with a second version of the test was found. A signal detection analysis of test scores showed that students who performed well on the test showed more conservative responding strategies, in that they made slightly more hits and substantially fewer false alarms. The internal properties of this test were also assessed through item analyses and an exploratory factor analysis, which demonstrated that some variance exists in the effectiveness of APTT items, and suggested that the dimensionality of the APTT may be difficult to determine.

ACKNOWLEDGEMENTS

Above all, I thank God, for providing numerous opportunities for success, and my parents, whose encouragement and support were the foundation for my decision to continue my education. I would also like to acknowledge my committee members Dr. Martha Escobar and Dr. Adrian Thomas for their time and integral contributions to this project. Finally, I would like to thank Dr. Lewis Barker, whose insight and guidance has made this project not only possible, but enjoyable.

Style manual used: Publication Manual of the American Psychological Association
(5th edition)

Computer software used: Microsoft Word 2000®

Microsoft Excel 2000®

SPSS 11.5®

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I. INTRODUCTION

A taskforce of the American Psychological Association recently addressed the need for assessment of Psychology major achievement (Halonen et al., 2002). This taskforce established numerous outcomes defining this achievement, ranging from technological literacy to sociocultural awareness. Research reported here will focus on the first stated assessment outcome, developing a knowledge base of the basic ideas, perspectives, and concepts in psychology. While the taskforce reviewed and concluded that a number of methods showed strong potential for in-class assessment, they warned against concentrating solely on classroom indices. Addressing the need for assessment outside of the classroom, they suggested that only use of an assessment center and locally developed tests showed strong potential for this purpose.

A number of obstacles exist in the development and administration of such locally developed tests. Problems include determining what actually constitutes student ability, differences in course selection among majors, and the expense and time involved in developing a tool that adequately and objectively assesses what they have achieved. In addressing the development of such a test, the task force lists a number of student achievement goals, the first of which involves demonstrating "familiarity with the major concepts, theoretical perspectives, empirical findings, and historical trends in psychology" (Halonen et al., 2002, ¶ 1). Such a test should also assess a student's ability to use psychology's "concepts, language, and major theories" adequately (Halonen et al., 2002, ¶ 1). Other goals involve the ability to apply psychological knowledge, think critically about psychology, and adhere to psychology's core values. Because of the wealth of relevant student outcomes, the APA warns against the use of only one or two measures in assessing majors.

The research reported here focuses on the development of a test called the Auburn Psychology Term Test (APTT). This test assesses a student's knowledge of psychology vocabulary, including key terms, people, theories, and perspectives. The test is based on the premise that the ability to recognize, identify, and use the language of psychology underlies the development of more complex thinking and application skills within the discipline. In brief, students taking the APTT are presented with a list of 100 terms, 50 key terms and 50 foils, and asked to specify which terms belong in each category.

Given that the domain of psychology is comprised of numerous key terms, people, theories, and perspectives, and that some are more important than others, the first task was to identify key terms comprising the core of relevant psychology knowledge. Using several introductory psychology textbooks, 50 key terms representing fifteen different content areas in psychology common to introductory textbooks were selected such as learning and personality (see Appendix A for complete list). Griggs, Bujak-Johnson, and Proctor (2004) have recently addressed discrepancies between key terms across introductory textbooks, finding that 455 terms are in over half of the 44 current introductory textbooks, and 155 are in 80 percent or more. The researchers argued that because 6269 glossary terms exist in all introductory textbooks, with 74 percent of these terms appearing in three or fewer textbooks, little similarity can be found in this domain. In many cases the discrepancies between key terms amounts to changes in the way a similar concept is phrased, such as the presence of the concept *bell curve* in one textbook and *normal curve* in another. While the researchers attempted to account for clear synonyms, a thorough analysis of all possible similarities would likely be a rather large undertaking. In the midst of such discrepancies, the prevalence of 155 to 455 terms across textbooks points to a core of such terms that are relatively common to introductory psychology. A number of APTT terms are amongst these core terms. Approximately 60 percent of each version's key terms can be found in at least half of all available introductory psychology textbooks. In an attempt to eliminate the potential confound of participants in the present study having been exposed to different terms during their tenure as psychology students, all participants in experiment one were enrolled in an introductory psychology course being taught by the same professor in the same semester, and all terms used in the creation of the test were present in the required textbook.

Based on material from the same content areas, 50 foils (pseudo-psychological terms) were created, designed to resemble true psychology terms. Foils differed along several dimensions from their key term counterparts. The most common differences were semantic. Such foils were created by modifying existing psychological concepts, clearly changing or reversing their meaning. While some of these changes may have been the result of altering only a few letters, such as "gestation psychology", the resulting meaning significantly differs from any known concept in psychology. In some cases these changes were morphological, and involved adding prefixes or suffixes to existing psychology terms that clearly altered their meaning, such as "unnatural selection." Other foils sound like potential psychology terms but are not found in any psychological literature, thus students could not have been previously exposed to them, including “animalism” and “terminal stasis.” Finally, a few changes were phonetic, in which several letters of a key term were changed, resulting in a term with significant phonetic differences from the original term. An example of such a change is the term “tetragen,” derived from the developmental term “teratogen.” Such changes were applied following the commonplace recommendation that at least two letters be changed when creating such foils. (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001). Both for the sake of simplicity and because this is the way the concepts are generally used in the memory literature, items that subjects have previously been exposed to, key psychology terms in the APTT, will often be referred to as “old” and items to which subjects could have not had previous exposure, or foils, as “new.”

Traditionally, the gold standard for student assessment in higher education has consisted of asking students to recall information in written form, thus demonstrating exactly what, and how much, they know about the topic. Those in higher education realize that a number of limitations exist in using this type of assessment procedure, including the time commitment involved in creating and grading such tests, potential biases involved in grading the multiple possible interpretations of a concept, and the difficulty in sampling from the wealth of material that may have been covered in a course (or courses). Though memory researchers have not wholly agreed on the nature of the relationship between recall and recognition, tests of recognition may be a more efficient method of accessing the knowledge base of a student. The existence of lively debates amongst memory theorists for several decades has resulted in the formulation of a number of models of recognition and memory (i.e. Mandler, 1980; Hintzman, 1988; Gillund & Shiffrin, 1984).

Although the models of early recognition theorists often postulated a single memory process accounting for both recognition and recall (Neath & Surprenant, 2003), most theorists generally contend that separate or additional processes or steps are involved (e.g. Gillund & Shiffron, 1984). While the theoretical frameworks behind these models is beyond the scope of this paper, most of these models are based on studies testing differences between subjects' ability to recognize presented terms as being part of a list that they were previously exposed to, and their ability to recall, or generate, such terms. In a typical study, words are initially presented for very short periods of time, typically measured in milliseconds or seconds, and time from initial exposure to testing is also quite short, most often measured in seconds or minutes.

Several such studies have demonstrated similarities in performance between recognition measures and recall measures in a number of different types of memory tasks (Challis, Velichkovsky, & Craik, 1996), with study trials manipulating level of processing (Asthana & Nigrani, 1984) and with varying study times (Ratcliff & Murdock, 1976). The type of tasks used in these studies, however, differs from the present research in a number of important ways. While levels of processing may be manipulated, and in some cases participants may even be asked to read and manipulate a brief passage (i.e. Shaughnessy & Dinnell, 1999), the level of understanding of the terms is not on par with what happens in a classroom setting.

The second distinction between traditional recognition verses recall research and the present study is the amount of time over which exposure to key terms has occurred. Rather than encountering a term once or twice over a period of seconds, students in a classroom setting may be intermittently exposed to a term over the course of several days, weeks, or months. Few, if any, studies by memory theorists have attempted to measure the relationship between yes/no recognition and recall ability of material whose meaning has been highly emphasized, and which has been presented to subjects over an extended period of time. The nearest equivalent to this use of yes/no recognition tests has been their use in educational testing fields, research conducted by Stanovich and colleagues, and second language research.

The methodology of the APTT was originally modeled after Stanovich's "Print Exposure Checklist" (Stanovich & West, 1989). Stanovich presented participants with lists of real and fabricated authors and publications and tested their ability to discriminate between them. An assessment of the reliability and validity of these tests found them to exceed many traditional literacy measures (Stanovich, 2000). Stanovich found strong relationships between performance on these tests and a number of cognitive abilities related to literacy such as spelling ability, verbal fluency (Stanovich & Cunningham, 1992), orthographic and phonological processing skill (Stanovich & West, 1989), vocabulary size, reading ability (Cunningham & Stanovich, 1997), cultural knowledge (West & Stanovich, 1991), declarative knowledge (Stanovich, West & Harrison, 1995), and real-world reading activity (West, Stanovich & Mitchell, 1993). Even when variability due to general cognitive ability, age, and education was statistically factored out, these strong relationships still existed.

While Stanovich's Print Exposure Checklist inspired the creation of the APTT, yes/no recognition tests have been used to assess student outcomes in the field of language testing for over twenty years (Beeckmans et al., 2001). Use of this methodology evolved from first language testing beginning in the late 1920s in which students identified the words for which they knew the meaning from a checklist of terms. Foils were later added as a defense against possible overestimation (Anderson & Freebody, 1983). Meara and Buxton (1987) adapted this framework to second language testing. The primary focus of such tests was to determine vocabulary size in second language learners.

Second language testing researchers have espoused numerous benefits of the yes/no recognition methodology, the most salient of which concerns the ease of and speed of testing. Kojic-Sabo and Lightbown (1999) praise the format for its ability to test "a large number of words… within a very short period of time" (p. 180). Even critics of term-based vocabulary tests concede that, despite their simplistic structure, such tests can be a better indicator of participant's vocabulary than an in-depth analysis of only a few items (Reed, 2000). This is particularly beneficial in the context of educational testing outside the classroom, where time and resources may be limited, or when the amount of material covered in a class necessitates the use of a more superficial and comprehensive format. Kojic-Sabo and Lightbrown (1999) further laud the tests' efficiency as a measure of vocabulary size in light of several studies finding strong relationships between performance on the test and several other indices of first and second language proficiency. One such study was Anderson and Freebody's (1983) initial research with use of foils, which found that the yes/no format correlated more strongly with actual word knowledge (r = .85), measured through an interview process, than did multiple choice tests over the same words (r = .45). Loring (1995) compared yes/no vocabulary tests consisting of academic words to the Michigan Test of English Language Proficiency and its vocabulary subtest, finding strong correlations (rs = .70 and .68). Meara and

Buxton (1987) also studied the relationship between yes/no recognition tests and several established measures of second language proficiency such as the Cambridge First Certificate Examination, and also found strong correlations between these measures (e.g. $r = .70$). In addition, several studies have suggested that students prefer this testing methodology to other types of language tests (e.g. Cameron, 2002; Kojic-Sabo & Lightbrown, 1999).

Use of the format in second language testing has not been without its critics. Beeckmans et al (2001) outlines several criticisms of the format, most relate to analyzing results of yes/no recognition tests. Due to its unique testing format, there exists a wealth of possible scoring methods and corresponding theoretical frameworks with which to analyze the results of a yes/no recognition test. They contend that an adequate method of scoring must be found to further the analysis of the yes/no testing design. The reason for the difficulty in scoring yes/no recognition tests is that such tests yield four possible outcomes (Green & Swets, 1966), outlined in Figure 1. If an item is a key term and the subject identifies it as such, a "hit" is recorded. If the subject fails to identify the item as a true psychology term this is considered a "miss." If the item is a foil and is correctly identified as a foil, a "correct rejection" is scored, though if the subject incorrectly identifies a foil as a key term, the student has made a "false alarm."

	Response	
	Yes	**No**
Key Term	*Hit*	*Miss*
Foil	*Correct rejection*	*False Alarm*

Figure 1. Responding to yes/no recognition tests.

A full analysis of test performance using this methodology involves more than simply tallying correct responses. Two components must be taken into consideration. The first is sensitivity, which is a participant's actual accuracy or ability to discriminate between old and new items. The second is a participant's response bias, or criterion. An unbiased criterion means that a participant "always selects the alternative with the larger likelihood" (Pastore, Crawley, Berens, & Skelly, 2003; p. 558). A liberal or conservative bias leads a subject to be more likely to answer yes or no, respectively. Those with more liberal biases are more likely to score a hit, but also more likely to make a false alarm, while those with more conservative biases are more likely to correctly reject a foil but also more likely to miss a key term.

Feenan and Snodgrass (1990) caution against simply treating bias as a nuisance variable, stating that it is important to understand this part of the recognition memory process and how it is manifested in test performance. This contention has been supported in memory research that looked at effects of different study times, and thus varying levels of familiarity with, terms on a yes/no recognition test (Ruiz, Soler, & Dasi, 2004). This study indicated that while hit performance increased with study time, correct rejection performance increased even more significantly with study time. Other studies have found significant differences for both hits and correct rejections as familiarity with target material increases (e.g. Ratcliff, Clark, & Shiffrin, 1990). This phenomenon also speaks to the importance of response bias considerations, which will be addressed shortly.

In developing an assessment tool researchers must determine how much of the student's raw score is due to actual sensitivity and how much is a byproduct of responding criterion. The process of finding the most accurate way to measure how much the participant actually knows has led to a number of different models and formulas attempting to effectively measure sensitivity and bias. The following is not intended as a comprehensive analysis of methods of analyzing yes/no recognition tasks, but merely as a brief overview and explanation for the proposed use of several different measures to analyze APTT data.

One such method of analysis involves use of 'thresholds.' Among the earliest threshold models was Blackwell's (1953) high-threshold model, which was derived from early psychophysics (Luce, 1963). This model implies the existence of a threshold for stimulus detection, and suggests that the researcher's primary task is to determine where this threshold lies. The key difference between threshold models and most other models lies in threshold models' rejection of the idea that a continuum exists on which different memory strengths lie. Threshold based models assume that either an item is encoded at study or it is not. Items that are not encoded should therefore be completely unavailable at test (Snodgrass, Volvovitz, &Walfish, 1972). The implications this has for computing scores will be discussed below.

Two common threshold models exist. The simplest is generally known as the *one high threshold* model, which assumes only two possible memory states: recognition and nonrecognition (Snodgrass & Corwin, 1988). According to this model, if an old item exceeds the subject's memory threshold it will be correctly identified; failure to exceed the threshold will result in a "miss." While differing levels of encoding success easily account for differences in hit and miss rates, it would initially seem that threshold theories are unable to account for false alarms. If a participant has never been exposed to an item, it could never have been initially encoded, and it should be incapable of exceeding the memory threshold. Threshold models generally justify the presence of false alarms by stating that when participants do not recognize an item a level of response bias often leads them to guess. As previously mentioned, different responding criterions will result in participants being more or less likely to guess when they do not recognize the item, leading to differing numbers of false alarms.

While the *one high threshold* model's inability to adequately explain qualities of actual data have led to its disuse (see Bayen, Murnane, & Erdfelder, 1996; Murdock, 1974; Snodgrass & Corwin, 1988), another threshold model, known as the *two high threshold* model has received some degree of support (Corwin, 1994; Feenan & Snodgrass, 1990). The *two high threshold* model assumes that two thresholds exist, one separating a state of uncertainty with a state of certainty that the item is a target, and the other separating the uncertain state with a state of certainty that the item is a foil (Corwin, 1994). These two thresholds are assumed to be equal, an assumption supported by research on recognition mirror effects (Snodgrass & Corwin, 1988). The hit rate will include a number of guesses, determined by the participant's response criterion, while the false alarms will consist solely of guesses from the uncertain state.

Threshold theories typically use participants' number of false alarms to determine their true probability of getting a hit. Statistics attempting to measure sensitivity based on threshold theories include P_r, the probability of new or old items exceeding the threshold, which simply subtracts the probability of a false alarm from the probability of a hit, by:

$$P_r = P(h) - P(f). \qquad (1)$$

P*(h), an estimation of the true hit rate, is calculated by dividing Pr by one minus the probability of a false alarm, or:

$$P^*(h) = P(h) - P(f) / 1 - P(f). \qquad (2)$$

Most formulas that attempt to correct for guessing also utilize the framework of threshold theories, assuming that either the participant knows the correct response or simply guesses at random (Huibregtse, Admiraal & Meara, 2002).

Measuring bias using a *two high threshold* framework involves calculating B_r, which Huibregtse et al.(2002) define as "the probability of saying yes to an item when in the uncertain state." B_r is calculated by dividing the false alarm probability by 1 minus P_r , or:

$$B_r = P(f) / [\ 1 - (P(h) - P(f))\ ] \qquad (3)$$

If B_r is equal to 0.5 the participant is said to have a neutral bias, anything above or below 0.5 is considered to be due to liberal or conservative bias, respectively.

The primary criticism of threshold theories is that recognition data often suggest a continuum of memory strength, as all items are not generally assumed to be equally familiar or unfamiliar (Murdock, 1974). For this reason, strength theories, the most prominent of which is signal detection theory, have abandoned the idea of thresholds. Signal detection theory's theoretical origins can be traced back to Fechner and Thurstone's era of psychophysics and their attempts to determine how adept subjects are at distinguishing between stimulus situations containing a signal and noise and those containing only noise (Green & Swets, 1963; Luce, 1963). In applying signal detection theory to recognition memory experiments, the signal is widely considered to be "strength of evidence" (Pastore, Crawley, Berens & Skelly, 2003, p.560), though what actually constitutes the 'noise' component of a recognition task involving memory has recently come under question. Numerous articles since the introduction of signal detection theory have defined noise in terms of cognitive processes or neural activity interfering with retrieval (Levine & Schefner, 1991). Pastore et al. (2003) criticizes this description of noise, stating that referring to 'noise' as literal cognitive processes misses the original purpose of the concept of noise and negates the basic ideas behind signal detection theory. According to these theorists, noise refers solely to variability in statistical processes, and signal detection theory, rather than being concerned with sensory or cognitive processing, is a "general model of decision processing of evidence" (p. 560).

Regardless of the phenomenological bases behind the use of some of its concepts, the basic premise of signal detection theory as it applies to recognition is that two normal overlapping distributions exist along a continuum of familiarity. One distribution consists of new items and the other consists of old items, and the amount of overlap that exists between these two distributions determines how well a participant is able to distinguish between items in each distribution. The measure used to determine the difference between the two distributions is D′, and is calculated by subtracting the standardized mean of the distribution of hits from the standardized mean of the distribution of false alarms, as:

$$D' = Z_f - Z_h \qquad (4)$$

Two measures of bias have seen widespread use in signal detection analysis, the earliest, β, is computed as the height of the distribution of hits divided by the height of the distribution of false alarms, or:

$$\beta = f(Z_h) \,/\, f(Z_f) \qquad (5)$$

The use of β in measuring bias has been widely criticized for two key reasons. The first is that the very use of β in some situations, particularly those involving stimuli that are heterogeneously memorable, assumes that a participant is able to accurately classify the stimulus as belonging to the either the distribution of new or old items, which is exactly what most memory studies are trying to test (Snodgrass & Corwin, 1988). Another problem is that while measures of bias and sensitivity may show a statistical relationship in some data sets, due either to factors acting on both measures or changes in sensitivity affecting bias, they should be computationally independent, a condition β consistently fails to meet (Snodgrass & Corwin, 1988). For these reasons β has largely been replaced with C, another measure of bias. Rather than focusing on the heights of the two distributions, C is measured as the distance from the intersection of the two distributions. C can be computed as the average of the standard scores for hits and false alarms, or:

$$C = (Z_h + Z_f) \,/\, 2. \qquad (6)$$

According to signal detection theory, for each participant a point will exist where the two distributions overlap, marking the point where new and old items are equally familiar. If this point also marks the participant's criterion for responding, a neutral bias is said to exist, and C will be equal to 0.

In order for the preceding calculations to be valid, the primary assumption underlying signal detection theory that both distributions are normal must be met. Pollack and Norman (1964) were among the first to call this and other statistical assumptions of signal detection theory into question, as well as offer a distribution-free, or nonparametric, method of analyzing results of yes/no recognition tasks. Because of the difficulty in determining equal variances, particularly if receiver operating characteristic curves can not be calculated due to testing participants only a small number of times, the assumption of normality may in some cases be unwarranted. To best illustrate how nonparametric measures are calculated, data can be plotted in a unit square with hit rate on the x axis and false alarm rate on the y axis. Figure 2 uses this format to show the data point (E) of a subject with a hit rate of 0.7 and a false alarm rate of 0.1. Signal detection theory assumes that because both old and new items are normally distributed, a curve can be created (see Figure 3) on which data point P falls that describes performance based on this one point. Nonparametric analyses instead attempt to determine the average area under a calculated curve denoting performance in an initial trial. Figure 4 illustrates that a curve based on a data point for a subject with a hit rate of .75 and a false alarm rate of .25 could be expected to pass through areas A1 and A2.

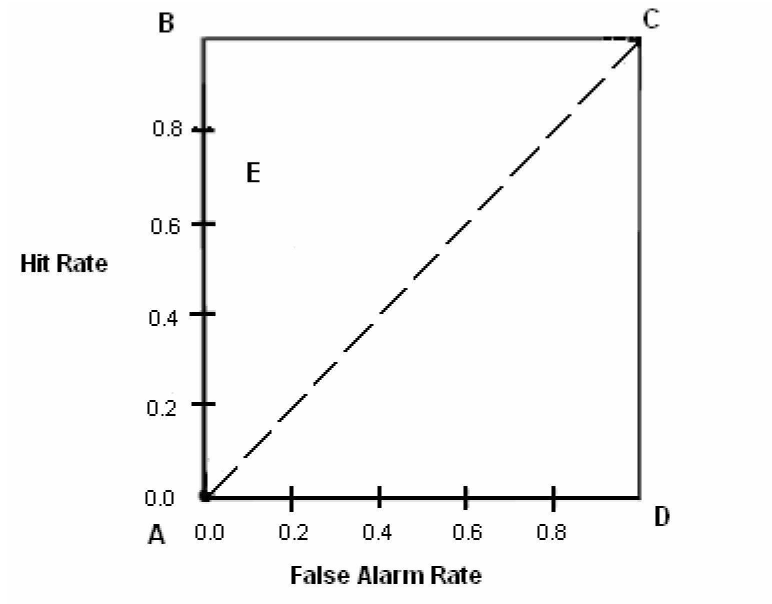


Figure 2. A hit rate of 0.5 on a unit square.

Note. Modified from I. Huibregtse, W. Admiraal, & P. Meara, 2002, *Language Testing*, 19, 227-245.

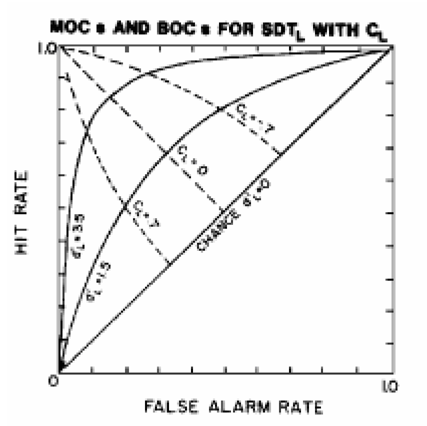


Figure 3. Using signal detection theory to describe data on a unit square

Note. From J. Snodgrass & J.Corwin, 1988, *Journal of Experimental Psychology*, 117, 34-50.

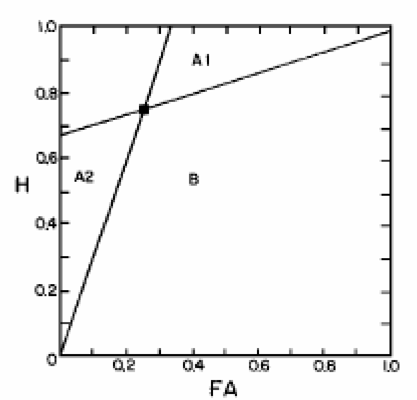


Figure 4. Nonparametric analyses using the unit square.

Note. Modified from I. Huibregtse, W. Admiraal, & P. Meara, 2002, *Language Testing*, 19, 227-245.

Several researchers have demonstrated that the area under the average curve created using areas A1 and A2 is a good indicator of memory performance (Pollack & Norman, 1964; Green & Moses, 1966). According to these researchers, such an index makes no assumption of normality or other statistical properties of the participants' distributions (Hodos, 1970). A′ is this sensitivity measure for nonparametric tests and can be calculated in terms of Figure 3 as:

$$A' = B + (A1 + A2) / 2. \qquad (7)$$

Two actual computational formulas for exist for A′ due to the fact that scores can possibly lie above or below the chance diagonal (Line AC in Figure 3). If the number of hits exceed the number of false alarms:

$$A' = .5 + [(P(h) - P(f)) * (1 + P(h) - P(f))] / [(4 * P(h)) * (1 - P(f))]. \qquad (8)$$

If the number of false alarms exceed the number of hits, the preceding formula can be modified by simply replacing each occurrence of hits with false alarms, and vice versa. If number of hits equal the number of false alarms, A′ = .5. Several computations exist for bias in a nonparametric model. Grier (1971) proposed the use of B′′, which can be seen in Figure 3 as B′′= A1-A2/A1+A2, and can be computed as:

$$B'' = [\ P(h) * (1-P(h)) - P(f) * (1 - P(f)\ ] / [\ P(h) * (1-P(h)) + P(f) * (1-P(f))\ ] \quad (9)$$

when the number of hits is greater than or equal to number of false alarms, and can be reversed when false alarm exceed hits by switching all occurrences of hits and false alarms in the formula. Hodos (1970) also proposed a bias index, referred to as B'_H, which can be seen in Figure 3 as B'_H = A1-A2/A1, and is calculated as:

$$B'_H = 1 - \{\ [\ P(f) * (1 - P(f))\ ]/[\ P(h) * (1-P(h))\ ]. \quad (10)$$

When hits exceed false alarms, B'_H can again be modified by reversing all occurrences of hits and false alarms and subtracting one from the total when false alarms exceed hits. Both equations for bias suggest neutral bias when the measure equals 0, liberal bias when positive, and conservative bias when negative (Snodgrass & Corwin, 1988).

For the past 35 years, a number of recognition memory researchers have espoused the use of nonparametric A′ due to its supposed lack of assumptions about underlying distributions (Hodos, 1970; Donaldson, 1992; Rhodes, Parkin, & Tremewan, 1993; Pastore et. al, 2003). Recently, Pastore et al (2003) called into question the rejection of signal detection based on its underlying assumptions, criticizing those who laud nonparametric measures as a distribution free alternative. Pastore first comments that the assumption that A′ measures the area under a theoretical average ROC curve falls apart at high levels of bias, underestimating sensitivity. Snodgrass and Corwin (1988) had previously made similar comments, and showed through several experiments that the fundamental assumption of independence between measures of bias and sensitivity does not hold true for nonparametric A′ and B′ measures. Pastore also demonstrates that A′ does indeed imply underlying distributions, suggesting that it is actually parametric.

Problems such as these have led a number of other researchers to reject the use of A′ and B′ as well as the use of β (Snodgrass & Corwin, 1988; Pastore et. al. 2003; Huibregtse, Admiraal & Meara, 2002), and have prompted others to suggest that all data be supported by several indexes, particularly lauding the independence of both P_r and D′ measures from their corresponding measures of bias, B_r and C , and suggest use of both sets of indexes in analyzing recognition data (Snodgrass and Corwin, 1988; Corwin, 1994; Feenan & Snodgrass, 1990). This is particularly true in light of Feenan and Snodgrass' (1990) study which showed significant effects of context on recognition of pictures and words that were observable through the use of some of the above measures, but not others. This is not a recent proposition, for as early as 1970 Lockhart and Murdock warned against the assumption that there was only one "correct" or "neutral" way to analyze recognition memory data.

The aforementioned paradigms have also served as the basis behind several new indexes. Meara (1992), developed an index which is a transformation of A′, estimating the hit rate that a participant would have scored had they not made any false alarms, calculated as:

$$\Delta m = [\ (P(h)-P(f)) * (1+P(h)-P(f))\ ] / [\ (P(h) * (1 - P(f))\ ] - 1, \qquad (11)$$

This formula is simply the transformation A′(4A′ - 3) and thus suffers from the same problems at high levels of bias. If a researcher does not wish to analyze bias separately, it may be factored out using equations such as I_{SDT}, which is presented as being based on a signal detection model, though it shares more similarity with nonparametric A′. I_{SDT}

was designed by Huibregtse et al (2002) to be used in analyzing tests of vocabulary, and can be computed by:

$$I_{SDT} = \frac{[4 * P(h) * (1 - P(f))] - [2 * (P(h) - P(f)) * (1 + P(h) - P(f)]}{[4 * P(h) * (1 - P(f))] - [(P(h) - P(f)) * (1 + P(h) - P(f)].} \qquad (12)$$

Huibregtse et al. (2002) attempts to correct for bias by basing his measure on the nonparametric calculations of A´ and determining the point at which the average ROC curve for a participant would intersect with the BD diagonal (see Figure 3). Any point on the BD diagonal is assumed to be free from bias, and Huibregtse cites Grier's (1971) bias measure as the basis for determining where the ROC curve intersects with this diagonal. How effectively Huibregtse's index incorporates bias correction into a nonparametric analysis has yet to be determined, though it initially appears that he has effectively eliminated the problems that A´ encountered at extremely high levels of bias.

Due to Snodgrass and Corwin's aforementioned recommendations (1988), analysis of APTT data will be conducted using several indices. Sensitivity analyses will be conducted using D´, P_r, and I_{SDT}, and bias will be assessed through the use of C and B_r. Conducting analyses without subscribing to a specific model is an attempt to obtain a well rounded picture of available data.

Aside from addressing the problem of scoring yes/no recognition tests, Beeckmans et al. (2001) outlined a number of other methodological concerns regarding their use in assessing student outcomes. Their analysis of second language yes/no recognition tests showed negative correlations between student performance on key terms and foils. They suggest that such an inverse relationship, which is likely a product of response bias clouding the results, calls into question the tests' discriminant validity.

They also suggest an analysis of any differences in distribution variance between key terms and foils to attempt to establish whether similar processes and distributions exist for the different types of items. While some of Beeckmans et al.'s criticisms are the basis of procedures for assessing the APTT, their concerns may not apply to the APTT test for several reasons. First of all, Beeckmans et al's examination was conducted using tests with unequal numbers of foils and key terms, a practice common in second language applications of the yes/no format. Differing numbers of key terms and foils necessitate several adjustments to resulting scores, and may confound some of the basic theoretical assumptions of key term and foil distributions inherent in some formulas. Beeckmans et al. also use a correction for guessing for part of their analysis that does not seem to meet the aforementioned requirements concerning independence of sensitivity and response bias.

In an effort to answer some preliminary questions about the format, and assess some basic psychometric properties of the APTT, eight hypotheses were tested, each corresponding to an addressed concern over yes/no recognition tests, validity and reliability of such tests, and test bias:

1. A significant relationship will exist between student scores on the APTT and other performance measures in introductory psychology courses.
2. The correlation between hits and total performance will be equal to the correlation between correct rejections and total performance.
3. Students who perform better on the APTT will show more conservative response biases.

4. The APTT will show adequate psychometric properties in each of the following analyses:
    a. Item and scale means and standard deviations
    b. Item total correlations and item scale correlations, using total performance as well as hit and correct rejection performance
    c. Item characteristic curve analysis to determine how well each item discriminates at all levels of performance
    d. Split half reliability between key terms and foils, as well as alpha
    e. An exploratory factor analysis to determine the dimensionality of the APTT
5. Some gender differences will exist in APTT performance.
6. Gender differences mentioned in hypothesis 7 will disappear once class performance is taken into account.
7. A significant relationship will exist between performance on the APTT and ability to recall information about key psychology terms.
8. Administration of an alternate form of the APTT, created using the same methodology will yield similar scores, and strong alternate form reliability.

Chapter II. EXPERIMENT 1

Method

Participants

Participants were 259 Auburn University students over the age of 19, enrolled in an introductory psychology course. The instruments were administered at the end of the semester, during the week of final exams.

Materials

Each student received one of two versions of the Auburn Psychology Term Test (APTT), each consisting of 50 key terms in psychology and 50 foils (see Appendices A and B). In part two of the task, students were given another form consisting of 20 randomly selected items from the alternate version. Between 12 and 15 of these items were key terms, the remainder were foils. Students were asked to determine which of these terms were correct and which were foils in the same manner as they did on the APTT. On the back of this form students chose 10 of the 20 terms that they have identified as key terms in psychology and were asked to "describe, define, or identify" the terms, giving as much information as they could recall in the space provided. Two versions of part two were created for each version of the APTT, totaling four distinct forms (see Appendix C).

Procedure

Introductory psychology students were given informed consent forms and administered the APTT, recording their responses on a scantron form. After the completion of the APTT, students were given part two. The relationship between scores on the APTT, using raw scores, I_{SDT} and D´, and course grades was assessed (hypothesis 1), as well as information concerning the relationship between performance on the APTT and hit and correct rejection performance (hypothesis 2), and bias (hypothesis 3). Several validity and reliability measures were assessed as discussed in hypothesis 4. Results were analyzed to assess any performance differences based on gender (hypothesis 5). After such differences were determined to exist, statistical analyses were conducted to determine if these differences could be accounted for by classroom performance (hypothesis 6).

The relationship between a student's APTT performance and his/her ability to recall information demonstrating a working knowledge of key psychology terms, as addressed in hypothesis 7, was also assessed. Responses in this section were graded on a five-point Likert-type scale, with scores (a) denoting that a student demonstrates an ability to recall a significant amount of correct information about the concept, or (b) demonstrates adequate recall ability of concept, consisting of correct statements or ideas that suggest a working knowledge of the item, or (c) demonstrates some knowledge of concept, recalling information that, while incomplete or only partially correct, suggests some knowledge of the core idea, or (d) does not demonstrate an adequate level of recall ability, but does seem to have some idea of the subject matter involved, or (e) demonstrates no recall ability of the term. Two raters independently assigned a score to each response. When the scores for an item were within one number value of each other, the response was scored as the mean of the two. When scores were two or more number values apart, the two raters discussed the item and agreed upon a score.

Results

Data were initially analyzed using raw scores, as well as indices D´, I_{SDT}, P_r, and A´. Because results for the following analyses were virtually identical using all of the above measures, only raw scores will be reported here.

A significant relationship was found between student introductory psychology course grade and performance on the APTT, as the correlation between course grade and APTT score was $r(257) = .63$, $p < .01$. This finding supports hypothesis 1, that the APTT would show significant relationships with other established measures of student performance. Figure 5 shows this relationship, in which APTT performance is represented as six levels, each representing approximately 20 percent of participants.

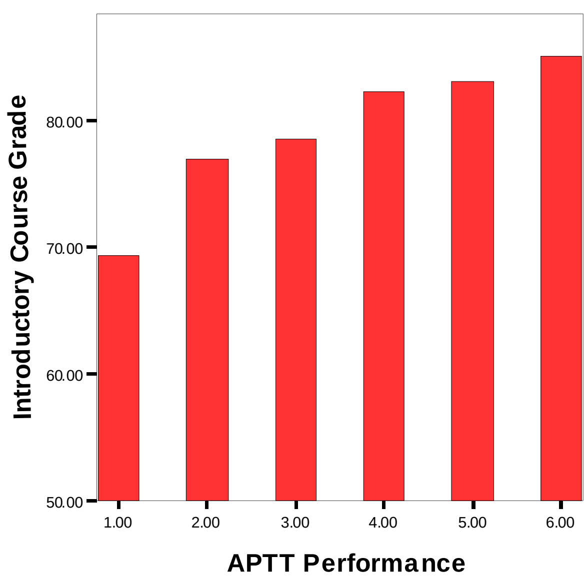


Figure 5. APTT performance and introductory psychology course grade.

The correlation between total APTT score and hits, $r(257) = .51$, $p < .01$, while significant, was significantly lower than the correlation between total APTT score and correct rejections $r(257) = .87$, $p < .01$. Figure 6 illustrates these relationships. A Fisher Z test of the differences between the correlations yielded $Z = 8.73$, $p < 0.01$. The second hypothesis predicted equality of the two correlations. This hypothesis was therefore rejected.

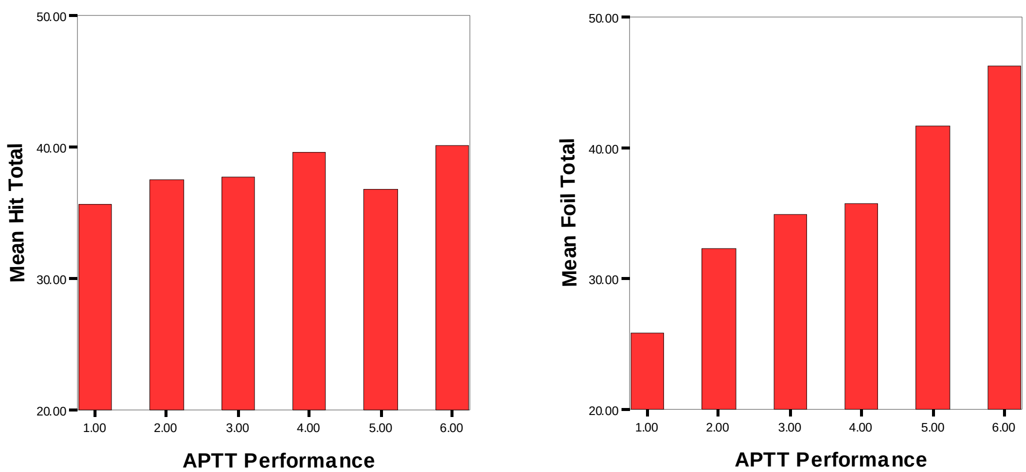


Figure 6. APTT performance as a function of key term and foil performance.

An analysis of the relationship between total score and response bias, measured using the bias index C, showed a strong correlation $r(257) = .49$, $p < .01$. This finding supports the assertions made in hypothesis 3, that a significant relationship would exist between participants' overall APTT performance and response bias. Figure 7 illustrates that as APTT score increases, C increases. An increase in C represents a more conservative responding strategy.

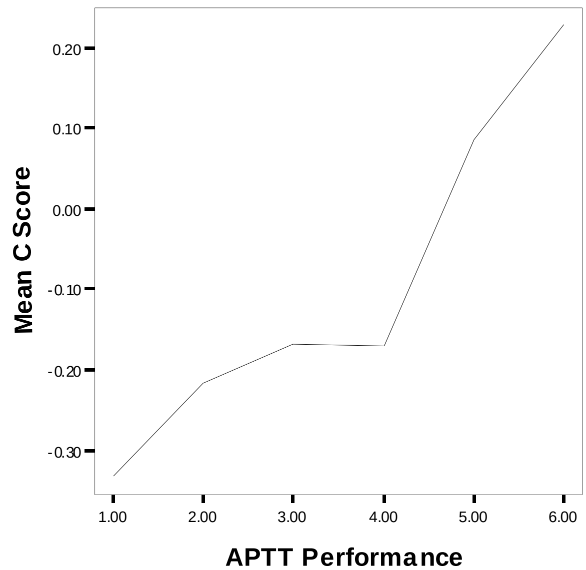


Figure 7. Response bias (C) and APTT performance.

The item analysis demonstrated several notable aspects of the test. The item total correlations of all items, shown in Appendix D, and the item characteristic curves, shown in Appendices E and F, suggest significant differences in the effectiveness of items in discriminating good and poor performers. Overall, as demonstrated in Figure 6, foils were far better discriminators of performance than key terms, with 48 of 50 item total correlations reaching significance at $p < .05$. Only 15 of 50 item total correlations for key terms reached significance at this level. Correlations reached significance more often when performance on items was compared with overall performance on the same class of items, as performance on 33 of 50 key terms significantly correlated with key term performance at $p < .05$, and all 50 foils significantly correlated with foil performance at $p < .05$. Scale correlations can be found in Appendices G and H. Cronbach's Alpha was

.81 for the test, and a split half reliability analysis between key terms and foils yielded a non-significant result at $r(257) = .02$, $p = .732$. Means and standard deviations for individual items can be found in Appendix I.

A principle components analysis was conducted to determine the number of factors among APTT items. Using parallel analysis criterion outlined by Lautenschlager (1989) one factor was determined to exist for foils. We then ran a one factor solution using maximum likelihood extraction with Obliman rotation. This produced a one factor model that accounted for 14.6 percent of the data. Items grouping into this factor (with a criterion of .35) are listed in Table 1.

Using the same parallel analysis criterion three factors were determined to exist for key terms. We then ran a one factor solution using maximum likelihood extraction with Obliman rotation which produced three factors for key terms. The three factor model for key terms accounted for 13.19 percent of the data. Items grouping into these three factors (using the same criterion of 0.35) are listed in Table 2.

Table 1.

Items grouping into the factor for foils.

Factor 1		
somatic transmission	unsystematic sensitization	retrograde amnesia
post-modern structuralism	interdependent variable	phobic malingering
conditional restriction	toddler directed speech	instinctual deprivation
intersubjective validity	proto-operational stage	functional flexibility
spontaneous salivation	neutral correlation	threshold of non-relativity
unconscious neuroticism	biological watch	multiple deviation
schema taking score	California-Binet test	

Table 2

Items grouping into the three factors for key terms.

Factor 1	Factor 2	Factor 3
bell-curve	bell-curve	
- inductive reasoning		
- unconditioned response	unconditioned response	
	fundamental attribution error	fundamental attribution error
		fixed action pattern
		cognitive dissonance
		just noticeable difference
		chunking
		episodic memory

Sixteen students did not report gender on their response sheet and were dropped from this analysis. The correlation was significant between gender and APTT score, $r(241) = .20$, $p < .01$. This finding supports hypothesis 5, which stated that gender differences would exist in APTT performance. Females performed significantly better than males. When controlling for introductory psychology course grade this correlation was reduced to $r(241) = .15$, p = .021, remaining significant at the .05 level. A Fisher Z test of differences between these two correlations was not significant at Z = .58, $p = .56$. Hypothesis 6, which stated that gender differences would be accounted for by differences in introductory psychology class performance, was therefore rejected.

A strong relationship was found between ability to recall information about psychology key terms and APTT performance, $r(229) = .60$, $p < .01$. Twenty-eight students did not complete the written section, and were subsequently dropped from the analysis. These results support hypothesis 6, which stated that a significant relationship would exist between APTT performance and recall ability. Figure 8 illustrates this relationship. Each of ten written items was worth between one and five points, bringing the total to 50 possible points.

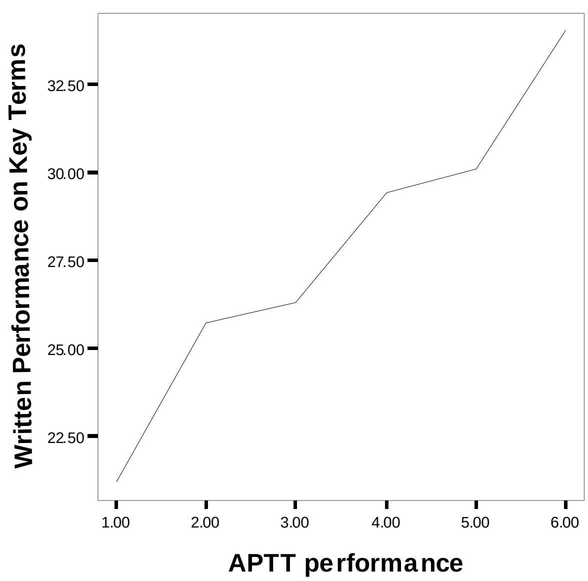


Figure 8. APTT performance and written recall performance on part two.

Chapter III. EXPERIMENT II

A second version of the APTT was created, consisting of 50 different key terms and 50 different foils, using the same procedure outlined in study one to determine alternate form reliability between the two instruments.

Method

Participants

Participants were students enrolled in a research methods course for credit at Auburn University. All participants had previously completed an introductory psychology course, though neither time elapsed from the completion of the course nor introductory course professor or content were controlled.

Measures

Both versions of the APTT (see Appendices A and B).

Procedure

Students (n = 40) enrolled in a research methods course in which no pre-testing had occurred were administered both versions of the APTT in random order. Data was analyzed to assess alternate form reliability in the two groups.

Results

Individual scores on the alternate form of the APTT correlated strongly with APTT performance $r(38) = .81$, $p < .01$, which was significant despite the smaller sample size. This supported the assertions of hypothesis 9, which stated that administration of an alternate form of the APTT would yield adequate alternate form reliability. Student scores on the two versions are graphically illustrated on Figure 9.

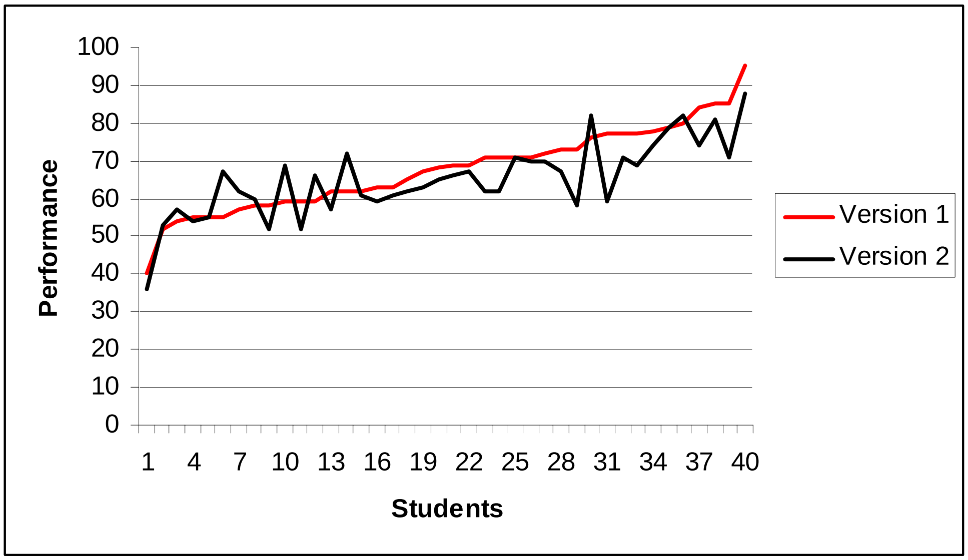


Figure 9. Performance comparison on versions one and two across students

Chapter IV. DISCUSSION

Preliminary analyses of the Auburn Psychology Term Test (APTT) suggest that it has strong potential for use in assessing psychology vocabulary knowledge. The significant relationship between classroom performance and APTT score suggests that the APTT is testing basic psychology knowledge. While classroom performance may not be a perfect indicator of student knowledge in the subject matter, the value placed on classroom performance in the educational system suggests that it must be considered to be among the indexes with the strongest potential for the assessment of the material covered. In Pilot studies, performance on the APTT prior to the start of an introductory psychology class has been shown to be at chance levels, suggesting that learning psychology in a classroom setting leads to better performance on the APTT. The results of this study suggest that performance on the APTT may be dependent on the amount of, and depth of understanding of, material learned

Beeckmans et al's (2001) criticism concerning the invariance of the contributions of foils and key terms to overall scores in yes/no recognition tests of learning may apply to the APTT. While the relationship between hit performance and total performance was strong, it was significantly lower than the relationship between foil performance and total performance, suggesting that performance on foils contributed to a student's total score more than hit performance. A ceiling effect may be observable for many of the key terms, as 22 of 50 key terms were correctly identified by 90% or more of the students taking the test, and the mean percentage correct for key terms was 77%. In comparison, only 5 of 50 foils were correctly identified by 90% or more students, and the mean percentage correct was 72%. In light of memory research on yes/no recognition tests, however, this result could be expected (i.e. Ruiz, Soler, & Dasi, 2004). Whether this invariance is an expected artifact of this testing methodology or should be cause for concern may be open for debate, though the analyses outlined in hypothesis four, which will be discussed shortly, do further our understanding of how each item contributes to overall performance.

The third hypothesis sought to confirm the findings in recognition memory literature (i.e. Ruiz, Soler, & Dasi, 2004) concerning the relationship between study time and response bias on yes/no recognition tasks, as well as provide additional evidence for the relationship between familiarity with psychology vocabulary and performance on the APTT. Because Ruiz et al. demonstrated that as study time increased, the propensity of a subject to reject terms that he/she was unsure about also increased, we expected to find a relationship between response bias and overall performance. Results indeed showed a strong relationship between the two measures ($r = .49$), hence, Ruiz et al's finding concerning the relationship between the amount of time spent with the material and performance on yes/no recognition tests may be generalizable to classroom settings, and thus the APTT. This also demonstrates that response bias on the APTT is not simply a random artifact of the test, but can be useful along with test performance in assessing student knowledge. Future research on this relationship may help researchers better understand student test taking strategies on yes/no recognition tests and how these strategies relate to actual vocabulary familiarity and knowledge.

An analysis of the psychometric properties of the APTT showed several notable points. Cronbach's Alpha for the test was .81, well above the .70 standard that Nunnaly (1978) deemed an acceptable reliability coefficient, which indicates high internal consistency. However, the most salient result of this analysis was the previously mentioned discrepancy between student performance on key terms and foils. While performance on most foils was significantly correlated with overall test performance (96%), performance on far fewer key terms (30%) showed significant correlations. This invariance can also be seen in split half reliability between key terms and foils, which was not significant at $r(257) = .02$, $p = .732$. In light of the previously mentioned finding demonstrating a significantly higher correlation between foils than key terms and overall test performance these findings are hardly surprising. Again, it is possible that we are observing a ceiling effect on key term performance, as 22 key terms were answered correctly by 90% of participants. Also of note was the variability across items in item total correlelations. Because of the nature of the learning environment, this effect on key terms could have been caused by either heterogeneously memorable items or differences in the amount of emphasis placed on concepts during the semester. However, as mentioned previously, all items used in this study were covered during the introductory psychology course, and all were contained in the required textbook.

Variability in foil performance is difficult to assess. Phonetic changes were far less common than semantic changes on the APTT administered to the participants, making analysis of differences along this dimension difficult. Difference in word length or number of syllables do not appear to be a factor (see foils in Appendices B and C). Since participants should have had no previous exposure to foils, their relationship to existing vocabulary items would be difficult to determine.

The exploratory factor analysis conducted seemed to suggest that the nature of the testing methodology did not lend itself to a salient grouping of items into identifiable factors. A low percentage of items grouped into factors during the analysis of hits and foils (see tables 1 and 2), and no discernable relationship could be established among those that did group into factors. While the dimensionality of the APTT could not be easily determined, the significance of this finding is unclear. Because participants are required to make yes/no decisions, as opposed to a Lickert or multiple choice testing format, and perhaps due in part to the presence of foils whose precise relationship to items in a participant's existing vocabulary cannot be determined, assessing the dimensionality of the test may not be possible at the present time through use of any available analyses.

Gender differences in performance were initially found on the APTT, with gender correlating with APTT performance at $r(241) = .20$, $p < .01$, and classroom performance $r(241) = .134$, $p < 0.05$. Gender differences were then analyzed by assessing the relationship between gender and APTT performance while controlling for classroom performance. When classroom performance was held constant the relationship between gender and APTT performance did decrease from $r = 0.20$ to $r = 0.15$, though this reduction was not significant. While the APTT may contain some gender differences, these could potentially be the result of other factors, such as differing study habits.

The nature of the relationship between recognition and recall may be particularly relevant in determining the effectiveness of recognition tests for assessing student knowledge. How effectively students were able to demonstrate general psychology vocabulary knowledge in an essay-type recall task was assessed by giving students random blocks of terms from the alternate version and asking them to provide "as much information as they know" about each term. While testing each student 's recall ability using the same terms he/she received on his/her version of the APTT would certainly have provided useful results, it was our intention to separate the assessment of students' overall psychology vocabulary recall ability from their knowledge of the particular terms in the version of the APTT that each student received. The inclusion of foils, which consisted of 5 to 8 out of the 20 items, also could have resulted in some artifacts of the APTT's testing methodology clouding the results. However, the strong correlations between written performance of psychology vocabulary items and APTT performance could not likely be seen wholly as the result of these methodological issues.

Several items were found to be poor predictors of student knowledge on both the essay and APTT portions of the study, and eliminating these items from the analysis resulted in correlations which were higher than those reported. The strong relationship between the recall and recognition portions of the test suggests similar processes or abilities at work in recognition on the APTT and recall of psychology vocabulary items, and may suggest a blurring of the distinction between the underlying processes. As in most educational tests, both consist of decontextualized vocabulary items, and may be testing the same basic ability. If so, the convenience of the yes/no format and sophistication of signal detection analyses provide further support for the usefulness of the test.

Administration of both an alternate form of the APTT and the original version to a group of student established a strong relationship between the two versions ($r = .81$). Aside from demonstrating strong alternate form reliability between the two versions, as well as establishing a viable second version of the test, the relationship between the two tests may speak to the stability of this testing methodology. Despite the fact that both versions contained entirely different key terms and foils, and (unlike in experiment one) that exposure to different key terms was not controlled for by participants having been enrolled in the same introductory psychology class at the time, performance was fairly reliable across versions. This may suggest that the testing methodology used is more important than the particular terms contained in the test, though, as found in experiment one, some terms were better than others at discriminating strong and weak performers.

Any analysis of the APTT as a test of psychology vocabulary knowledge should take into consideration certain theoretical differences inherent in educational and language testing discourse. Chapelle (1998) describes the division between trait and interactionalist approaches to second language acquisition research, a division that outlines some of the criticisms of yes/no recognition tests in that field. Trait theorists generally contend that test performance reflects relatively stable "underlying processes or structures" (Messick, 1989, p. 15). Such theorists view language performance along four dimensions of use, including vocabulary size, knowledge of word features and characteristics, organization in the mental lexicon, and use of fundamental semantic, phonological, and morphological vocabulary processes (Chapelle, 1998). Performance along these dimensions of general knowledge and cognitive processes are considered to represent a stable, measurable ability to use the target language.

Interactionalist theorists' most consistent criticism of trait theories, and thus, vocabulary tests as measures of language ability, stems from what they perceive as a disregard of context (Chapelle, 1998). Such theorists assert that tests of language should take the pragmatic and contextual features of the word into consideration. Several researchers suggest that subjects' ability to recognize words, or even their comprehension of these words, may not demonstrate an ability to use them in context (Laufer & Paribakht, 1998; Reed, 2000). However, other researchers have expressed concerns over the use of context, suggesting that some tests of language proficiency may measure inferencing skills as much as actual word knowledge (i.e. Laufer, 2004). While recognizing that not all terms that a participant recognizes may be fully understood, it is unlikely that participants will be capable of using in context, or recalling information about, terms that they are unable to recognize.

Some of the most frequent criticisms of the use of yes/no recognition tests in language testing do not necessarily apply to the proposed research. Many of these concerns involve factors such as phonotactic probability differences between languages (Beekmans et al., 2001). Cameron (2002) laments the possibility that students, having encountered unfamiliar words throughout the educational process, may become accustomed to such encounters and have more difficulty distinguishing words from non-words. Read (1997) expressed similar sentiments, arguing against the use of foils because low-level learners have more difficulty with the use of non-words.

However, these concerns may in fact demonstrate the strength of the format, rather than its weaknesses. Those who may be considered "low-level learners" should perform more poorly on this test, and those who are more familiar with psychology terms and concepts involved should have a better knowledge of what they do not know as well as what they know, leading to better performance on foils. This contention is supported by Ruiz et al.'s (2004) studies on study times and response bias discussed earlier, as well as the results of this study.

Another criticism of the yes/no format involves instructions given to test takers in language research (Beeckmans et al, 2001; Laufer & Paribakht, 1998). Typically, these tests ask participants to identify words for which they know the meaning, a standard that may have different implications for different test takers. By giving instructions in this manner, those in the language field are separating what many memory researchers contend are two types of recognition memory judgments (i.e. Atkinson & Juola, 1974; Mandler, 1980; Wixted & Stretch, 2004). These theorists suggest that one recognition process simply involves a sense of familiarity, and another involves a "conscious recollection" or identification of the information involved (Neath & Surprenant, 2003, p. 210). In an attempt to avoid this dichotomy, the instructions of the APTT simply ask the participant to discriminate actual psychology terms from foils. Either process, if such a distinction truly exists, could thus lead to the participant's response.

The APA task force on student assessment encouraged the use of locally developed tests to supplement in-class indices of student performance. The instrument designed and tested in this study, the Auburn Psychology Term Test (APTT), has been demonstrated to be reliable and valid as well as economical in terms of time and resources. This study looked at the relationship between this instrument and several indicators of student performance, including introductory course grade and ability to identify and define psychology vocabulary items, and found strong relationships between these variables. The internal properties of this test were also assessed through item analyses and an exploratory factor analysis, which demonstrated that some variance exists in the effectiveness of APTT items, and suggested that the dimensionality of the APTT may be difficult to determine. An alternate form was also created, and the two tests showed strong alternate form reliability, indicating the formats consistency. Other researchers using similar tests have found them to be good measures of a number of student characteristics, most notably; vocabulary knowledge. Such research has also found that students like the format in comparison with other testing formats. Additionally, the signal detection analysis encourages integration with an extensive literature on recognition memory. For these and other reasons, it is hopeful that other educators and researchers will find the APTT useful.

REFERENCES

Anderson,. R.C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), Advances in Reading/Language Research: A Research Annual. Greenwich, CT: JAI Press, 231-256.

Asthana, B., & Nagrani, S. (1984). Recall and recognition as a function of levels of processing. *Psycho-Lingua*, *14*, 85–94.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Luce, & P. Suppes (Eds.) *Contemporary developments in mathematical psychology* (Vol. 1). San Francisco: Freeman.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*, 81-99.

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 197-215.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the yes/no vocabulary test: some methodological issues in theory and practice. *Language Testing, 18,* 235-274.

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, 6, 145-173.

Challis, B. H., Velichkovsky, B. M., & Craik, F. (1996). Levels-of-Processing Effects on a Variety of Memory Tasks: New Findings and Theoretical Implications. *Consciousness and Cognition, 5,* 142-164.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman, A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research.* Cambridge: Cambridge University Press.

Corwin, J. (1994). On measuring discrimination and bias: Unequal numbers of targets and distractors and two classes of distractors. *Neuropsychology, 8,* 110-117.

Cunningham, A., & Stanovich, K. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33,* 934-945.

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General, 121(3),* 275-277.

Feenan, K., & Snodgrass, J. (1990). The effect of context on discrimination and bias in recognition memory for pictures and words. *Memory & Cognition, 18(5),* 515 -527.

Gillund, G., & Shiffrin, R. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short term memory. *Psychological Bulletin, 66,* 228-234.

Griggs, R., Bujak-Johnson, A., & Proctor, D., (2004) Using common core vocabulary in text selection and teaching the introductory course. Teaching of Psychology, 31, 265-269.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75(6),* 424-429.

Halonen et. al. (2002) The Assessment CyberGuide for Learning Goals and Outcomes in the Undergraduate Psychology Major. http://www.apa.org/ed/guide_outline.html

Hintzman, D. L. (1988). Judgments of frequency and recognition in a multiple-trace memory. *Psychological Review 95*, 528-551.

Hodos, W. (1970). Nonparametric index of response bias for use in detection and recognition experiments. *Psychological Bulletin, 74,* 351-354.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: correction for guessing and response style. *Language Testing*, 19, 227-245.

Kojic-Sabo, I., Lightbown, P. (1999). Students' approaches to vocabulary knowledge and their relationship to success. *The Modern Language Journal, 83,* 176-192.

Laufer, B. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing, 21,* 202-226.

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of learning context. *Language Learning, 48,* 365-391.

Lautenschlager, G. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Mulitvariate Behavioral Research, 24, 365- 396.*

Levine, M. W., & Shefner, J. M. (1991). Fundamentals of sensation and perception (2nd ed.). Pacific Grove, CA: Brooks & Cole.

Lockhart, R., & Murdock, B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100-109.

Loring, T. (1995) A Yes/No Vocabulary Test in a University Placement Setting. Unpublished masters thesis.

Luce, D. R. (1959). *Individual Choice Behavior*. New York: John Wiley & Sons.

Luce, D. R., Bush, R., & Galanter, E. (1963) *Handbook of mathematical psychology (Vol. 1*). New York: John Wiley & Sons.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252-271.

Meara, P. M., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. Language Testing, 4, 142-154.

Meara, P. M. (1992). *EFL Vocabulary Test.* Swansea, UK. Centre for Applied Language Studies.

Murdock, B. B. (1974). *Human memory: Theory and data*. Potomac, MD: Lawrence Erlbaum Associates.

Neath, I., Surprenant, A. (2003). *Human Memory: An introduction to research, data, and theory.* Belmont: Thompson & Wadsworth.

Nunnaly, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Pastore, R., Crawley, E., Berens, M., & Skelly, M. (2003). “Nonparametric” A′ and other modern misconceptions about signal detection theory. *Psychnomic Bulletin & Review*, 10, 556-569.

Pollack, I., Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science, 1,* 125-126.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83,* 190-214.

Reed, J. (2000). Assessing Vocabulary. Cambridge: Cambridge University Press.

Rhodes, G., Parkin, A. J., & Tremewan, T. (1993). Semantic priming and sensitivity in lexical decision. *Journal of Experimental Psychology: Human Perception and performance, 15*, 154-165.

Ruiz, J. C., Soler, M. J., & Dasi, C. (2004). Study time effects in recognition memory. *Perceptual and Motor Skills, 98,* 638-642.

Snodgrass, J., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology*, 117, 34-50.

Snodgrass, J., Levy-Berger, G., & Haydon, M. (1985*). Human experimental psychology*. New York: Oxford University Press.

Snodgrass, J., Volvovitz, R., & Walfish, E. R. (1972). Recognition memory for words, pictures, and words + pictures. *Psychonomic Science, 27,* 345-347.

Stanovich, K. E. (2000). *Progress in understanding reading*: Scientific foundations and new frontiers. New York: Guilford.

Stanovich, K. E. & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition*, 20, 51-68.

Stanovich, K. E. & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 24, 402-433.

Stanovich, K. E., West, R. F. & Harrison, M. (1995). Knowledge growth and maintenance across the life span: The role of print exposure. *Developmental Psychology*, 31, 811-826.

Tulving, E., & Thompson, D. M. (1971). Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology, 87,* 116 124.

West, R., & Stanovich, K. (1991). The incidental acquisition of information from reading. *Psychological Science, 2,* 325-330.

West, R. F., Stanovich, K. E. & Mitchell, H. R. (1993). Reading in the real world and its correlates. *Reading Research Quarterly*, 28, 34-50.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*(4), 616-41

APPENDICES

Appendix A

Auburn Psychology Term Test Version 1 *(*Bold items are key terms)*

Below, 100 terms are listed. Some of them are key psychological terms that you encountered in lectures and reading the textbook. Others will be unfamiliar to you, because they are bogus, fabricated terms that sound like psychological terms, but are not "real" psychology terms.
Your task is to identify which of the terms are real and which are fabricated. For example, terms such "memory" and "Ivan Pavlov" are both associated with psychology, so you would mark "A" on the scantron. Likewise, "intestinal myopia" and "terminal distress" are not part of psychology, so for these terms you would mark "B." Please look at each item, then bubble in "A" if you recognize it as a real term, and "B" if you think the term is bogus.

1 adolescent amnesia
2 **transduction**
3 **action potential**
4 **comfort touch**
5 schema taking score (STS)
6 **sexual identity**
7 **secondary reinforcer**
8 James Farber
9 **cognitive dissonance**
10 **critical period**
11 **token economy**
12 **chunking**
13 alpha-wave effect
14 ghost limb
15 **empiricism**
16 gestation psychology
17 **standard deviation**
18 **Jean Piaget**
19 **language acquisition device**
20 dendritic hypo-potential
21 **longitudinal study**
22 **negative feedback**
23 **libido**
24 superstitious relaxation
25 **bell curve**
26 antisocial facilitation
27 animalism
28 functional flexibility
29 neurostasis
30 **fixation**
31 **dendrite**
32 motivational intelligence
33 **attachment**
34 **big 5 personality factors**
35 hapless motivation
36 sleep activation
37 multiple deviation
38 **Shaping**
39 **general intelligence (g)**
40 proto-operational stage
41 **James-Lange theory**
42 neutral correlation
43 retrograde memory
44 species-typical behavior
45 **Wernicke's area**
46 latitudinal study
47 somatic transmission
48 **Synapse**
49 Psychotransference
50 biological watch
51 **inductive reasoning**
52 instinctual deprivation
53 indifferent schizophrenia
54 unconscious neuroticism
55 **null hypothesis**
56 successful approximation
57 psychogenic amnesia
58 **reaction range**
59 toddler-directed speech (TDS)
60 **obsessive compulsive disorder (OCD)**
61 **proactive interference**
62 terminal stasis
63 distance IQ
64 Bronski's area
65 **test-retest reliability**
66 **Temperament**
67 objective well-being
68 **law of effect**
69 **unconditioned response**
70 **dark adaptation**
71 unsystematic sensitization
72 **operational definition**
73 threshold of non-relativity
74 bystander apathy effect (BAE)
75 insensitive period
76 **circadian rhythm**
77 **paradoxical sleep**
78 spontaneous salivation
79 **fundamental attribution error**
80 **unipolar disorder**
81 Festinger-Maslow effect
82 **just noticeable difference(JND)**
83 **William James**
84 California-Binet test
85 interdependent variable
86 **sensorimotor stage**
87 **introspection**
88 duozygotic twins
89 phobic malingering
90 ego complex
91 **episodic memory**
92 **cognitive-behavioral therapy**
93 conditional restriction
94 **activation-synthesis hypothesis**
95 intersubjective validity
96 operant encoding
97 **systematic desensitization**
98 post-modern structuralism
99 latent gratification
100 **fixed action pattern (FAP)**

Appendix B

Auburn Psychology Term Test Version 2 *(*Bold items are key terms)*

Below, 100 terms are listed. Some of them are key psychological terms that you encountered in lectures and reading the textbook. Others will be unfamiliar to you, because they are bogus, fabricated terms that sound like psychological terms, but are not "real" psychology terms.
Your task is to identify which of the terms are real and which are fabricated. For example, terms such "memory" and "Ivan Pavlov" are both associated with psychology, so you would mark "A" on the scantron. Likewise, "intestinal myopia" and "terminal distress" are not part of psychology, so for these terms you would mark "B." Please look at each item, then bubble in "A" if you recognize it as a real term, and "B" if you think the term is bogus.

1. **blindsight**
2. id therapy
3. **anterograde** amnesia
4. **aphagia**
5. **homeostasis**
6. tri-delta waves
7. physiological clock
8. discontinuous reinforcement
9. **dissociation**
10. "Big Ten" Personality Factors
11. **adaptation**
12. **phenotype**
13. work memory
14. **convergence**
15. indiscriminate learning
16. replicated repetition
17. **retinal disparity**
18. **conservation of volume**
19. observational validity
20. Intellectual Quotient (IQ)
21. **neurosis**
22. involutional study
23. liquid intelligence
24. **psychosexual stages**
25. **linguistic relativity hypothesis**
26. semantic loop
27. **kin selection**
28. inheritability
29. need-for-improvement theory
30. Edward Dubranski
31. **hedonism**
32. **Cannon-Bard theory**
33. **experimenter bias**
34. **ecological validity**
35. Genomotypic
36. **Thalamus**
37. **Structuralism**
38. **stimulus generalization**
39. group-actualization theory
40. unnatural selection
41. Fractionalism
42. **confounding variable**
43. **activation-synthesis hypothesis**
44. invalidation therapy
45. polar cells
46. **Assimilation**
47. Maslow's Hierarchy of Emotion
48. split-cell research (SCR)
49. RPM Sleep
50. **myelin sheath**
51. narcissistic schizophrenia
52. **set point theory**
53. **somatosensory cortex**
54. variable ration schedule
55. **serial position effect**
56. sensitization cycle
57. unconditional negative regard
58. **Schema**
59. bottom-down processes
60. general activation syndrome (GAS)
61. **Biofeedback**
62. **attribution theory**
63. type C Personality
64. **monozygotic twin**
65. **learned helplessness**
66. collective conscience
67. delay theory
68. Stanford-WAIS
69. **bystander effect**
70. **Wilhelm Wundt**
71. **zeitgeiber**
72. **transference**
73. language imprinting device (LID)
74. transdifferentation
75. **social loafing**
76. arm-in-the-door technique
77. frustration-repression hypothesis
78. **self-actualization**
79. **psychosomatic disorder**
80. synaptic contusion
81. parallel amnesia
82. person esteem
83. **factor analysis**
84. **DSM-IV**
85. **crystalized intelligence**
86. Flynn defect
87. telegram speech
88. phenome
89. **inprinting**
90. mental set
91. group mind
92. **retroactive interference**
93. somalization
94. **hypochondriasis**
95. **free association**
96. tetrogen
97. algorerhythm
98. **conversion disorder**
99. Stroop defect
100. **spontaneous recovery**

Appendix C

Materials for the written recall portion of Experiment I

On this concluding portion of the study, pick any of the ten terms that you marked “real” on the reverse side, and briefly identify them. Write the term in the space provided, and describe/define/identify that term in one or two sentences in the space provided.

Version 1 terms

Form A	Form B
replicated repetition	blindsight
retinal disparity	id therapy
conservation of volume	anterograde amnesia
intellectual quotient (IQ)	aphagia
neurosis	homeostasis
involutional study	tri-delta waves
liquid intelligence	structuralism
psychosexual stages	physiological clock
linguistic relativity hypothesis	learned helplessness
kin selection	dissociation
semantic loop	attribution theory
myelin sheath	adaptation
bottom down process	phenotype
factor analysis	indiscriminate learning
conversion disorder	work memory
RPM sleep	convergence
Cannon-Bard theory	Thalamus
sensitization cycle	schema
serial position effect	spontaneous recovery
natural selection	DSM-IV

Version 2 terms

Form A

transduction
adolescent amnesia
action potential
sexual identity
cognitive dissonance
comfort touch
critical period
token economy
chunking
alpha-wave effect
empiricism
standard deviation
Jean Piaget
longitudinal study
bell curve
superstitious relaxation
dendrite
attachment
shaping
hapless motivation

Form B

Big 5 personality factors
sleep activation
general intelligence (g)
James-Lange theory
Wernicke's area
attitudinal study
synapse
biological watch
inductive reasoning
proactive interference
reaction range
terminal stasis
test-retest reliability
objective well-being
dark adaptation
operational definition
bystander-apathy effect (BAE)
circadian rhythm
bipolar disorder
independent variable

Appendix D

Item Total Correlations

Correlations

		TOTAL
Q2	Pearson Correlation	.008
	Sig. (2-tailed)	.923
	N	133
Q3	Pearson Correlation	.282(**)
	Sig. (2-tailed)	.001
	N	133
Q4	Pearson Correlation	-.157
	Sig. (2-tailed)	.071
	N	133
Q6	Pearson Correlation	-.140
	Sig. (2-tailed)	.109
	N	133
Q9	Pearson Correlation	.127
	Sig. (2-tailed)	.144
	N	133
Q10	Pearson Correlation	.163
	Sig. (2-tailed)	.061
	N	133
Q11	Pearson Correlation	.063
	Sig. (2-tailed)	.474
	N	133
Q12	Pearson Correlation	.256(**)
	Sig. (2-tailed)	.003
	N	133
Q15	Pearson Correlation	.135
	Sig. (2-tailed)	.122
	N	133
Q17	Pearson Correlation	.150
	Sig. (2-tailed)	.084
	N	133
Q18	Pearson Correlation	.296(**)
	Sig. (2-tailed)	.001
	N	133
Q19	Pearson Correlation	.159
	Sig. (2-tailed)	.068
	N	133
Q21	Pearson Correlation	-.045
	Sig. (2-tailed)	.610
	N	133
Q22	Pearson Correlation	-.008
	Sig. (2-tailed)	.930
	N	133
Q23	Pearson Correlation	-.165
	Sig. (2-tailed)	.057
	N	133
Q25	Pearson Correlation	.157
	Sig. (2-tailed)	.071
	N	133
Q30	Pearson Correlation	-.164
	Sig. (2-tailed)	.059
	N	133
Q31	Pearson Correlation	-.015
	Sig. (2-tailed)	.867
	N	133
Q33	Pearson Correlation	.267(**)
	Sig. (2-tailed)	.002
	N	133
Q34	Pearson Correlation	.103
	Sig. (2-tailed)	.239
	N	133
Q38	Pearson Correlation	.005
	Sig. (2-tailed)	.952
	N	133
Q39	Pearson Correlation	.114
	Sig. (2-tailed)	.191
	N	133
Q41	Pearson Correlation	.351(**)
	Sig. (2-tailed)	.000

	N	133		N	133
Q45	Pearson Correlation	.132	Q76	Pearson Correlation	.258(**)
	Sig. (2-tailed)	.130		Sig. (2-tailed)	.003
	N	133		N	133
Q48	Pearson Correlation	.030	Q77	Pearson Correlation	-.056
	Sig. (2-tailed)	.728		Sig. (2-tailed)	.524
	N	133		N	133
Q51	Pearson Correlation	.078	Q79	Pearson Correlation	.245(**)
	Sig. (2-tailed)	.373		Sig. (2-tailed)	.005
	N	133		N	133
Q55	Pearson Correlation	.211(*)	Q80	Pearson Correlation	.134
	Sig. (2-tailed)	.015		Sig. (2-tailed)	.124
	N	133		N	133
Q58	Pearson Correlation	.136	Q82	Pearson Correlation	.521(**)
	Sig. (2-tailed)	.119		Sig. (2-tailed)	.000
	N	133		N	133
Q60	Pearson Correlation	.112	Q83	Pearson Correlation	.204(*)
	Sig. (2-tailed)	.200		Sig. (2-tailed)	.018
	N	133		N	133
Q61	Pearson Correlation	-.015	Q86	Pearson Correlation	.375(**)
	Sig. (2-tailed)	.866		Sig. (2-tailed)	.000
	N	133		N	133
Q65	Pearson Correlation	.219(*)	Q87	Pearson Correlation	.133
	Sig. (2-tailed)	.011		Sig. (2-tailed)	.127
	N	133		N	133
Q66	Pearson Correlation	-.017	Q91	Pearson Correlation	.248(**)
	Sig. (2-tailed)	.848		Sig. (2-tailed)	.004
	N	133		N	133
Q68	Pearson Correlation	.270(**)	Q92	Pearson Correlation	.004
	Sig. (2-tailed)	.002		Sig. (2-tailed)	.962
	N	133		N	133
Q69	Pearson Correlation	.121	Q94	Pearson Correlation	-.053
	Sig. (2-tailed)	.166		Sig. (2-tailed)	.541
	N	133		N	133
Q70	Pearson Correlation	.070	Q97	Pearson Correlation	-.089
	Sig. (2-tailed)	.424		Sig. (2-tailed)	.310
	N	133		N	133
Q72	Pearson Correlation	.116	Q100	Pearson Correlation	.174(*)
	Sig. (2-tailed)	.182		Sig. (2-tailed)	.045

	N	133		N	133
Q1	Pearson Correlation	.294(**)	Q32	Pearson Correlation	.299(**)
	Sig. (2-tailed)	.001		Sig. (2-tailed)	.000
	N	133		N	133
Q5	Pearson Correlation	.344(**)	Q35	Pearson Correlation	.260(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.003
	N	133		N	133
Q7	Pearson Correlation	.017	Q36	Pearson Correlation	.318(**)
	Sig. (2-tailed)	.842		Sig. (2-tailed)	.000
	N	133		N	133
Q8	Pearson Correlation	.199(*)	Q37	Pearson Correlation	.356(**)
	Sig. (2-tailed)	.022		Sig. (2-tailed)	.000
	N	133		N	133
Q13	Pearson Correlation	.190(*)	Q40	Pearson Correlation	.384(**)
	Sig. (2-tailed)	.028		Sig. (2-tailed)	.000
	N	133		N	133
Q14	Pearson Correlation	.207(*)	Q42	Pearson Correlation	.393(**)
	Sig. (2-tailed)	.017		Sig. (2-tailed)	.000
	N	133		N	133
Q16	Pearson Correlation	.175(*)	Q43	Pearson Correlation	.379(**)
	Sig. (2-tailed)	.044		Sig. (2-tailed)	.000
	N	133		N	133
Q20	Pearson Correlation	.160	Q44	Pearson Correlation	.473(**)
	Sig. (2-tailed)	.066		Sig. (2-tailed)	.000
	N	133		N	133
Q24	Pearson Correlation	.257(**)	Q46	Pearson Correlation	.271(**)
	Sig. (2-tailed)	.003		Sig. (2-tailed)	.002
	N	133		N	133
Q26	Pearson Correlation	.337(**)	Q47	Pearson Correlation	.438(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q27	Pearson Correlation	.364(**)	Q49	Pearson Correlation	.241(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.005
	N	133		N	133
Q28	Pearson Correlation	.330(**)	Q50	Pearson Correlation	.349(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q29	Pearson Correlation	.476(**)	Q52	Pearson Correlation	.292(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.001

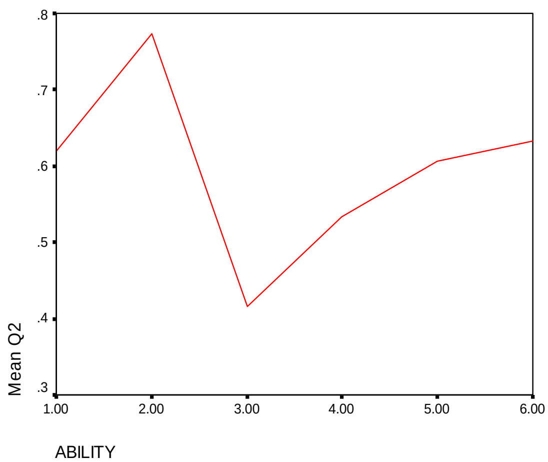
	N	133
Q53	Pearson Correlation	.411(**)
	Sig. (2-tailed)	.000
	N	133
Q54	Pearson Correlation	.401(**)
	Sig. (2-tailed)	.000
	N	133
Q56	Pearson Correlation	.256(**)
	Sig. (2-tailed)	.003
	N	133
Q57	Pearson Correlation	.512(**)
	Sig. (2-tailed)	.000
	N	133
Q59	Pearson Correlation	.287(**)
	Sig. (2-tailed)	.001
	N	133
Q62	Pearson Correlation	.468(**)
	Sig. (2-tailed)	.000
	N	133
Q63	Pearson Correlation	.231(**)
	Sig. (2-tailed)	.007
	N	133
Q64	Pearson Correlation	.441(**)
	Sig. (2-tailed)	.000
	N	133
Q67	Pearson Correlation	.477(**)
	Sig. (2-tailed)	.000
	N	133
Q71	Pearson Correlation	.369(**)
	Sig. (2-tailed)	.000
	N	133
Q73	Pearson Correlation	.272(**)
	Sig. (2-tailed)	.002
	N	133
Q74	Pearson Correlation	.457(**)
	Sig. (2-tailed)	.000
	N	133
Q75	Pearson Correlation	.312(**)
	Sig. (2-tailed)	.000

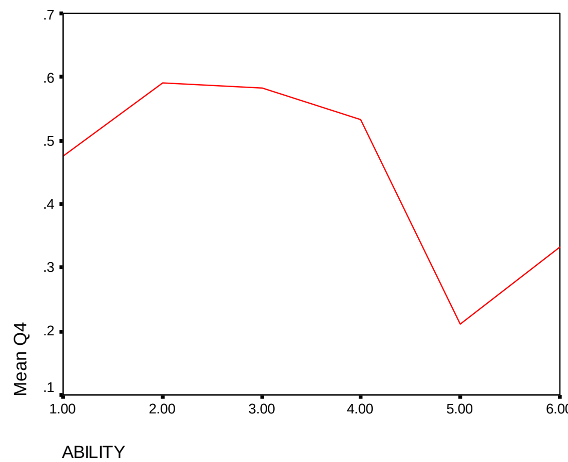
	N	133
Q78	Pearson Correlation	.396(**)
	Sig. (2-tailed)	.000
	N	133
Q81	Pearson Correlation	.216(*)
	Sig. (2-tailed)	.012
	N	133
Q84	Pearson Correlation	.387(**)
	Sig. (2-tailed)	.000
	N	133
Q85	Pearson Correlation	.420(**)
	Sig. (2-tailed)	.000
	N	133
Q88	Pearson Correlation	.542(**)
	Sig. (2-tailed)	.000
	N	133
Q89	Pearson Correlation	.394(**)
	Sig. (2-tailed)	.000
	N	133
Q90	Pearson Correlation	.166
	Sig. (2-tailed)	.056
	N	133
Q93	Pearson Correlation	.411(**)
	Sig. (2-tailed)	.000
	N	133
Q95	Pearson Correlation	.410(**)
	Sig. (2-tailed)	.000
	N	133
Q96	Pearson Correlation	.483(**)
	Sig. (2-tailed)	.000
	N	133
Q98	Pearson Correlation	.432(**)
	Sig. (2-tailed)	.000
	N	133
Q99	Pearson Correlation	.219(*)
	Sig. (2-tailed)	.011
	N	133

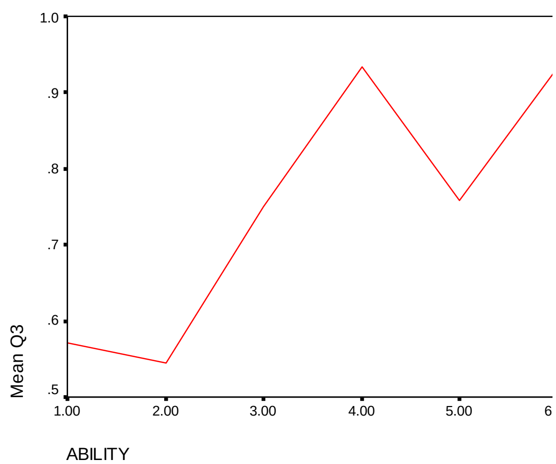
** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

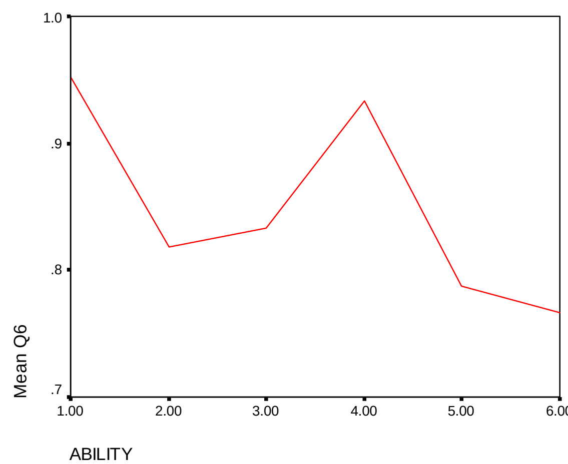
Appendix E

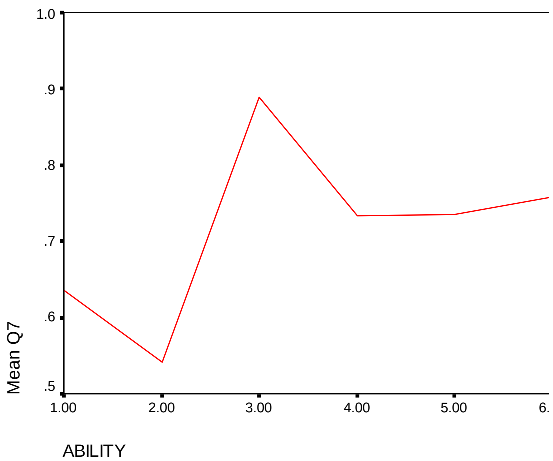
Item Characteristic Curves for Key Terms

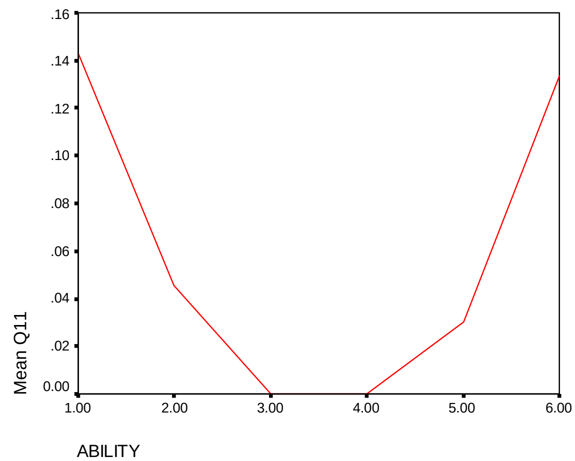


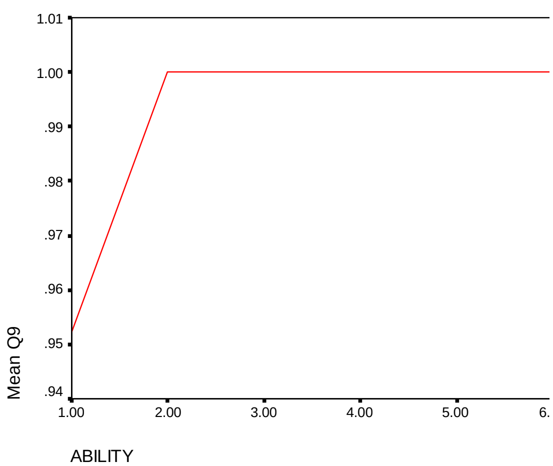


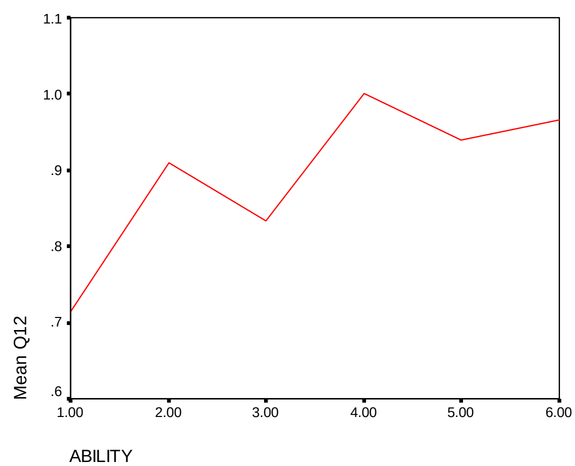


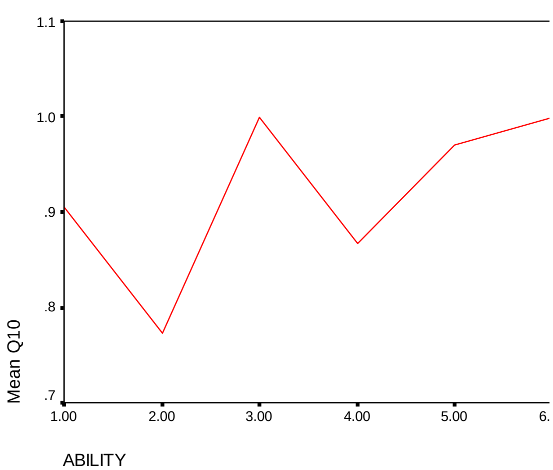


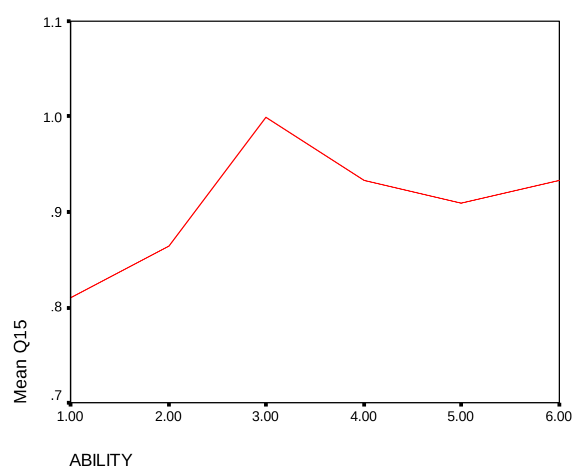

Mean Q7
1.0
.9
.8
.7
.6
.5
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

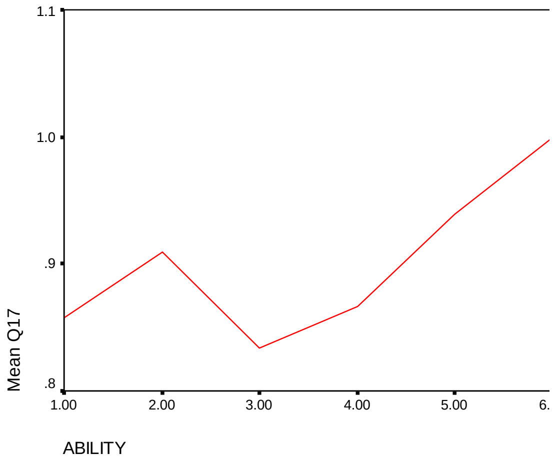

Mean Q11
.16
.14
.12
.10
.08
.06
.04
.02
0.00
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

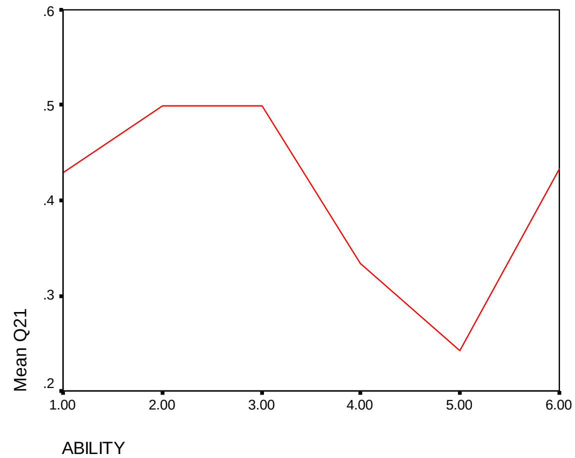

Mean Q9
1.01
1.00
.99
.98
.97
.96
.95
.94
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

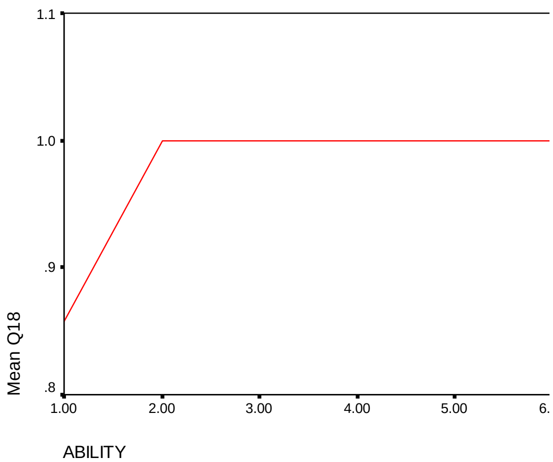

Mean Q12
1.1
1.0
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

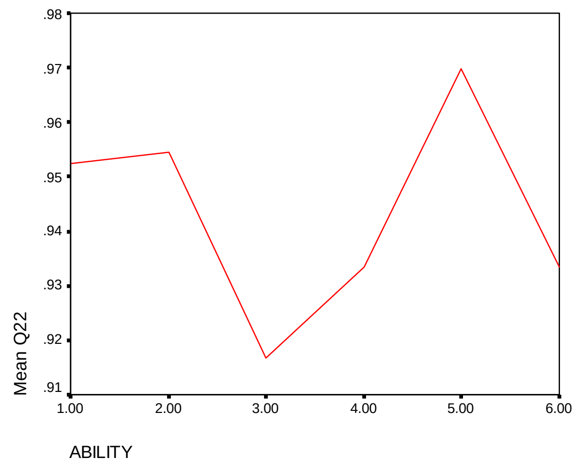

Mean Q10
1.1
1.0
.9
.8
.7
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

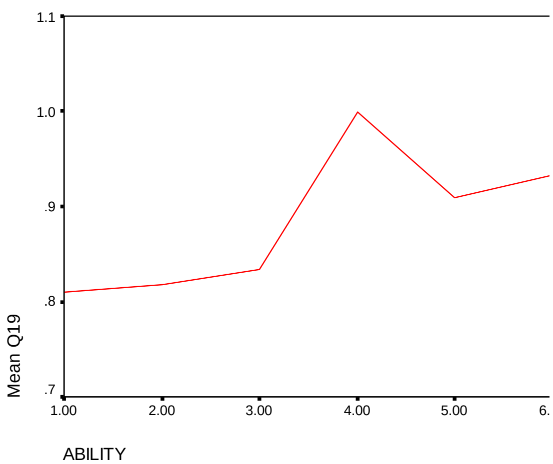

Mean Q15
1.1
1.0
.9
.8
.7
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

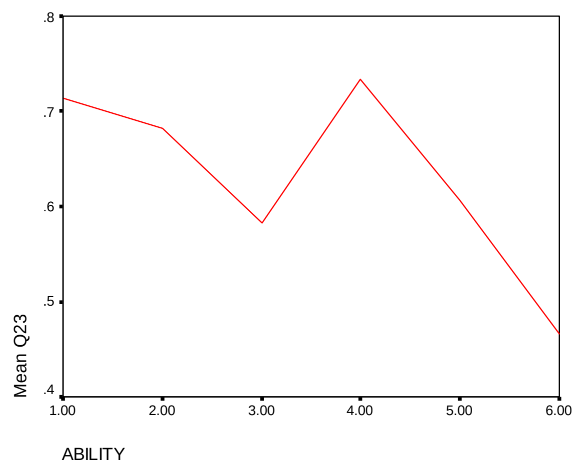

Mean Q17
1.1
1.0
.9
.8
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

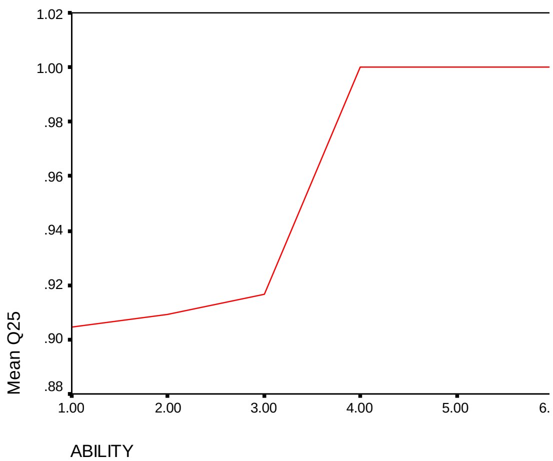

Mean Q21
.6
.5
.4
.3
.2
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

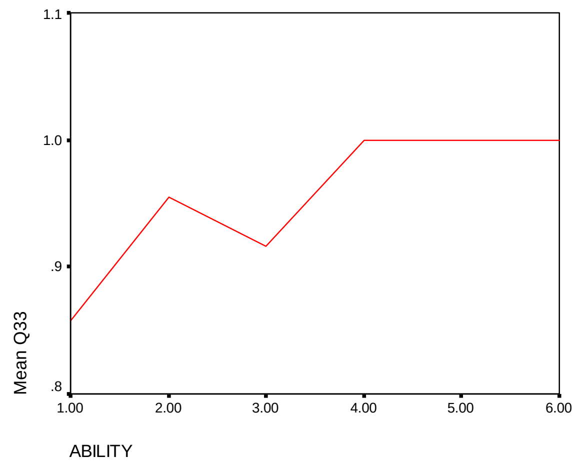

Mean Q18
1.1
1.0
.9
.8
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

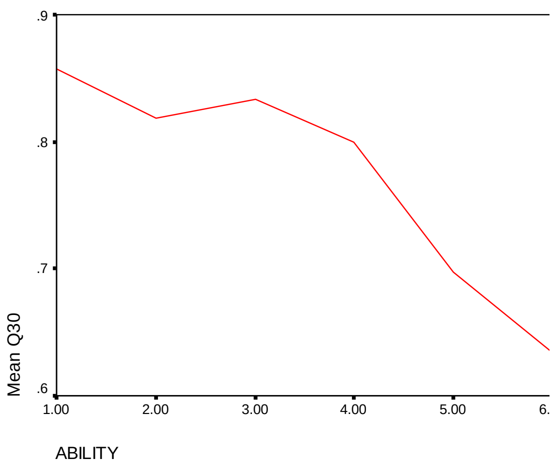

Mean Q22
.98
.97
.96
.95
.94
.93
.92
.91
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

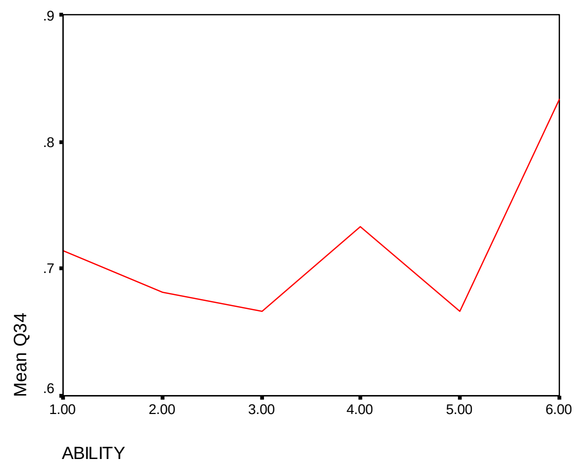

Mean Q19
1.1
1.0
.9
.8
.7
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

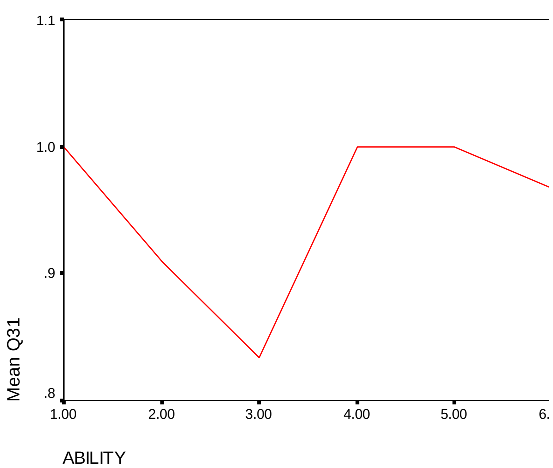

Mean Q23
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

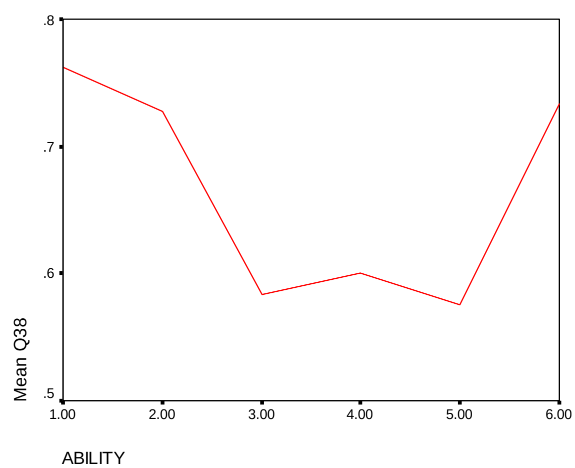

Mean Q25
1.02
1.00
.98
.96
.94
.92
.90
.88
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

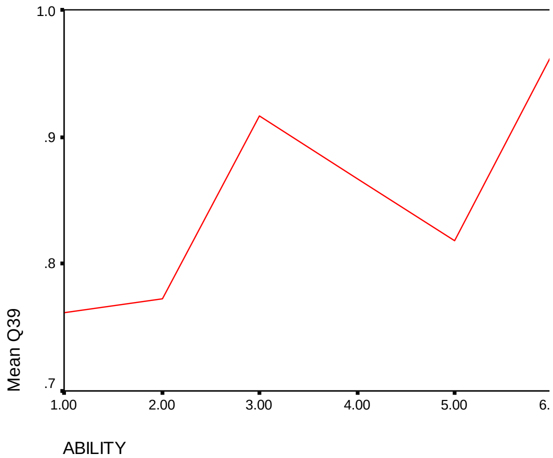

Mean Q33
1.1
1.0
.9
.8
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY


Mean Q30
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

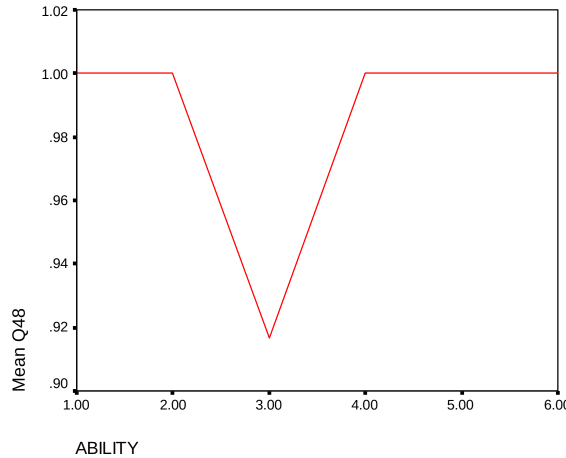

Mean Q34
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY


Mean Q31
1.1
1.0
.9
.8
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

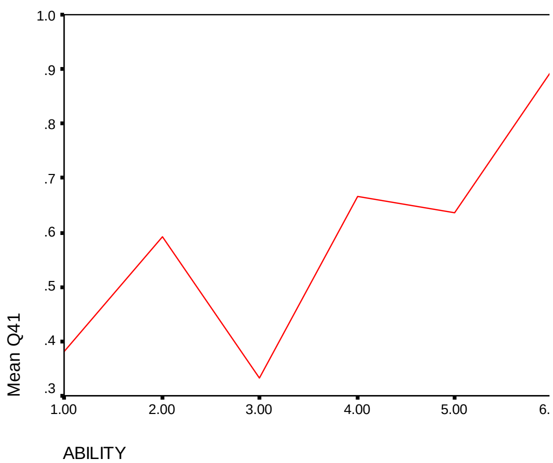

Mean Q38
.8
.7
.6
.5
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY



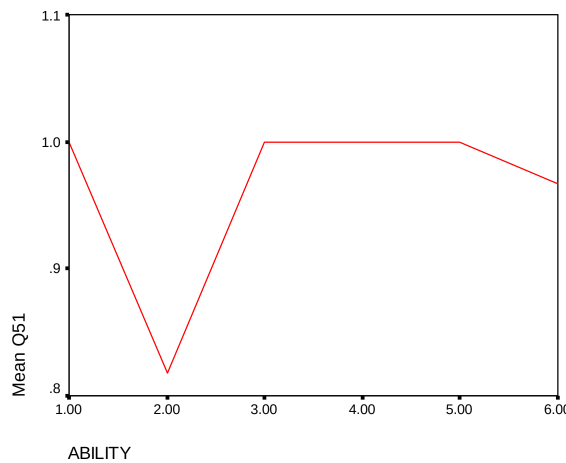
1.0
.9
.8
.7
Mean Q39
1.00
2.00
3.00
4.00
5.00
6.
ABILITY



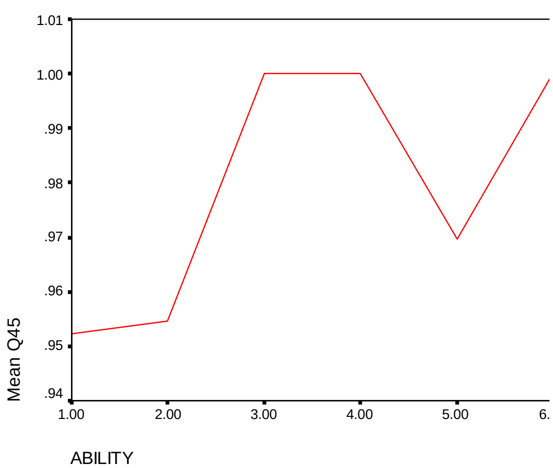
1.02
1.00
.98
.96
.94
.92
.90
Mean Q48
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY



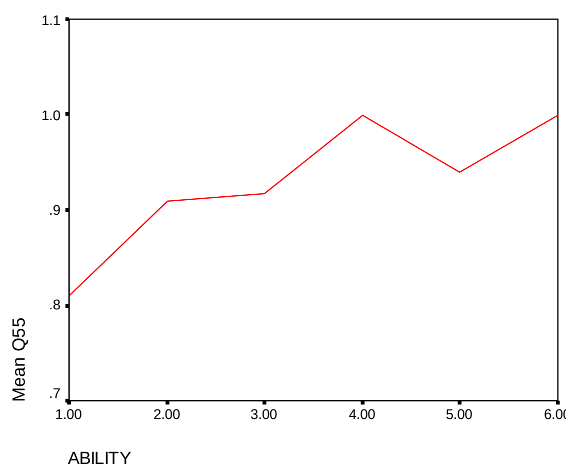
1.0
.9
.8
.7
.6
.5
.4
.3
Mean Q41
1.00
2.00
3.00
4.00
5.00
6.
ABILITY



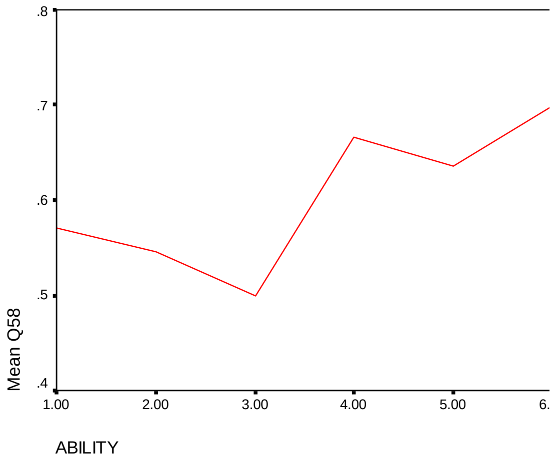
1.1
1.0
.9
.8
Mean Q51
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY



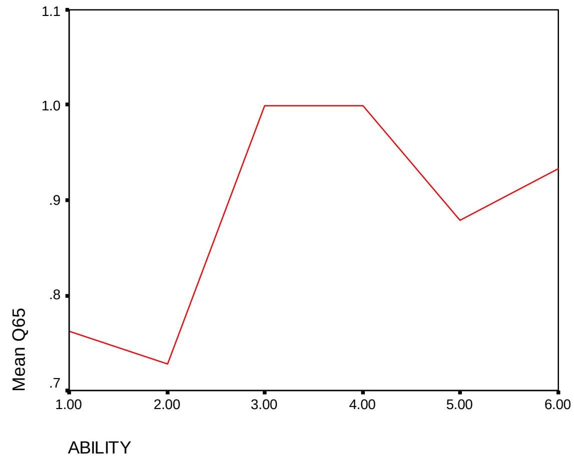
1.01
1.00
.99
.98
.97
.96
.95
.94
Mean Q45
1.00
2.00
3.00
4.00
5.00
6.
ABILITY



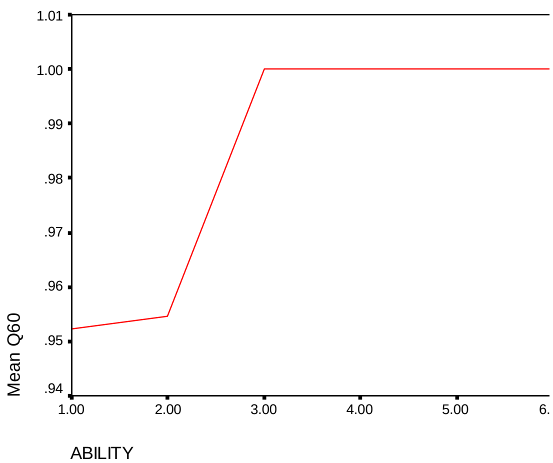
1.1
1.0
.9
.8
.7
Mean Q55
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY



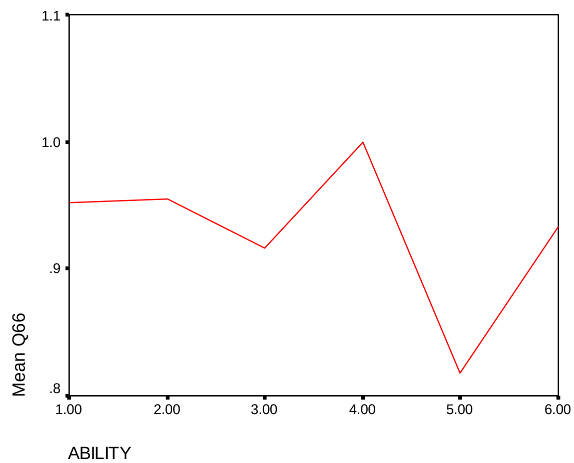
.8
.7
.6
.5
.4
Mean Q58
1.00
2.00
3.00
4.00
5.00
6.
ABILITY



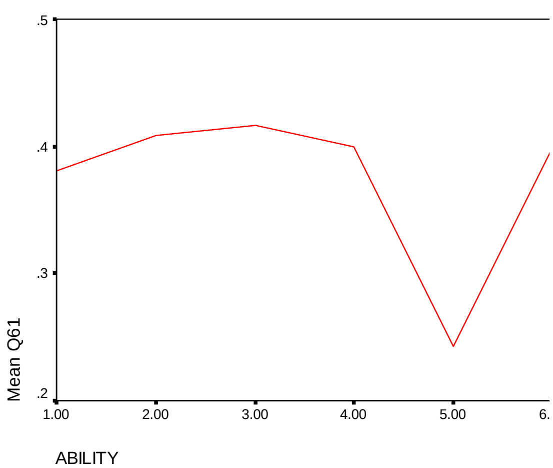
1.1
1.0
.9
.8
.7
Mean Q65
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY



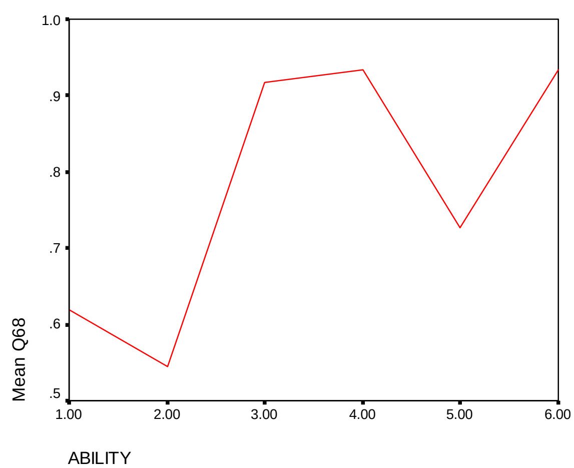
1.01
1.00
.99
.98
.97
.96
.95
.94
Mean Q60
1.00
2.00
3.00
4.00
5.00
6.
ABILITY



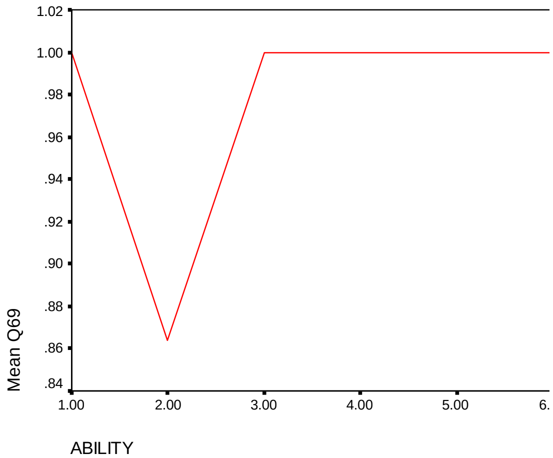
1.1
1.0
.9
.8
Mean Q66
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

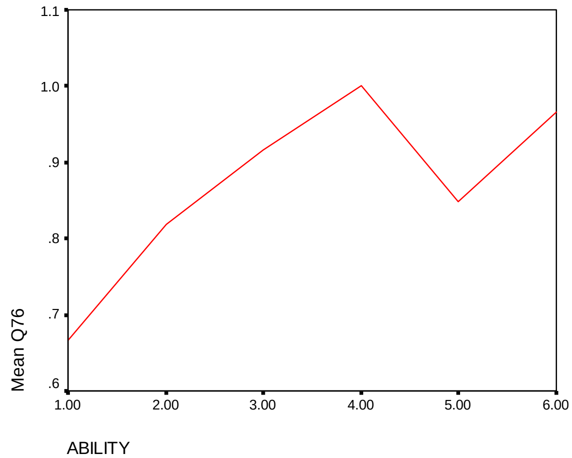


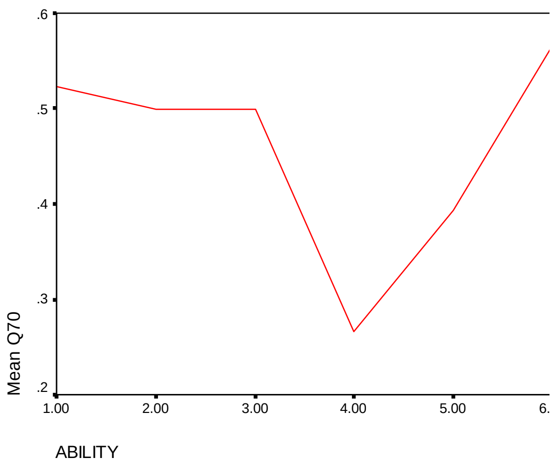
.5
.4
.3
.2
Mean Q61
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

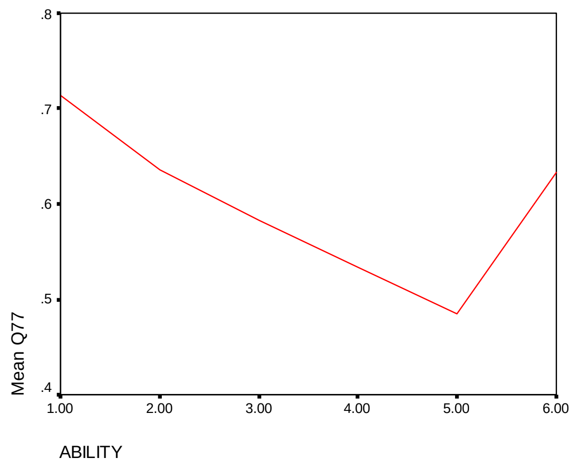


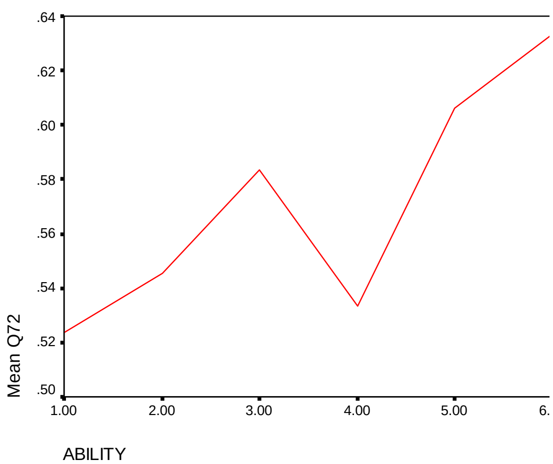
1.0
.9
.8
.7
.6
.5
Mean Q68
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

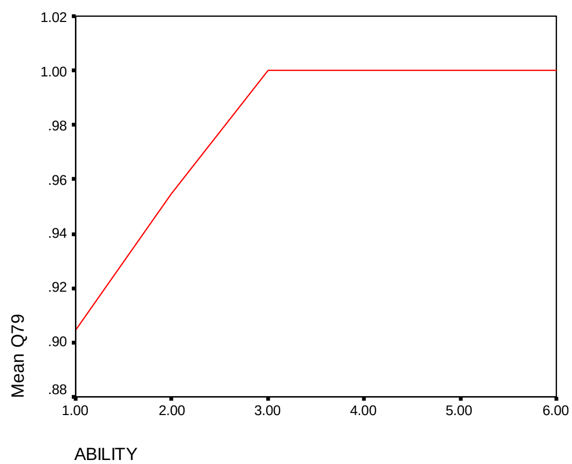

Mean Q69
1.02
1.00
.98
.96
.94
.92
.90
.88
.86
.84
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

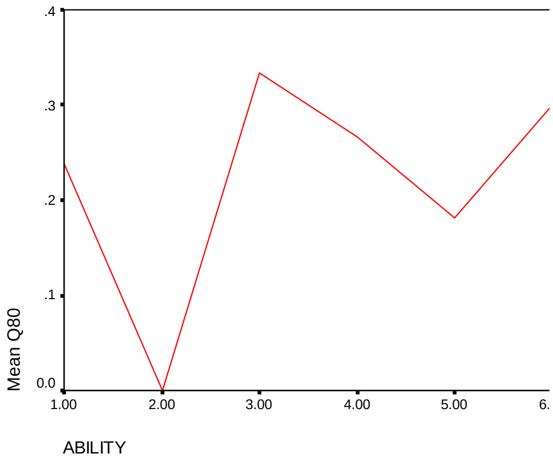

Mean Q76
1.1
1.0
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

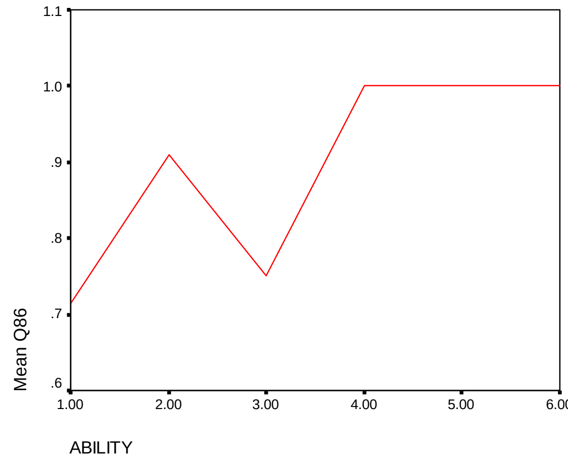

Mean Q70
.6
.5
.4
.3
.2
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

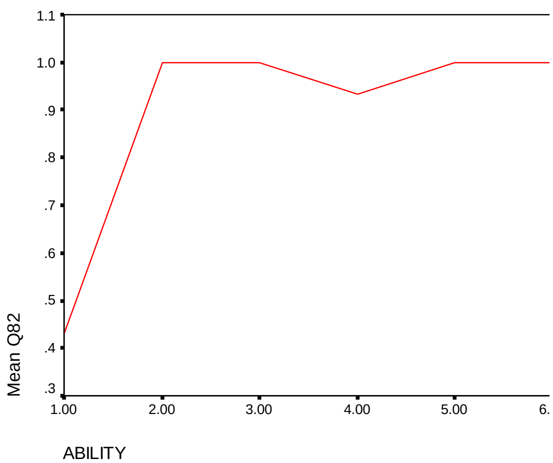

Mean Q77
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

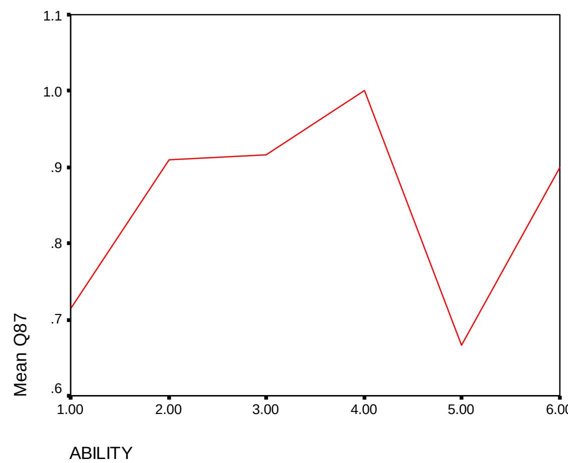

Mean Q72
.64
.62
.60
.58
.56
.54
.52
.50
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

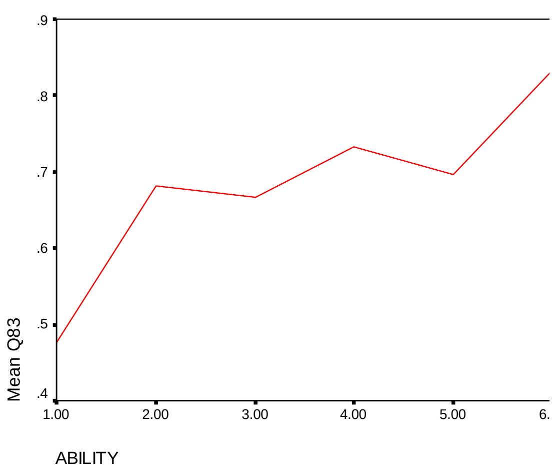

Mean Q79
1.02
1.00
.98
.96
.94
.92
.90
.88
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

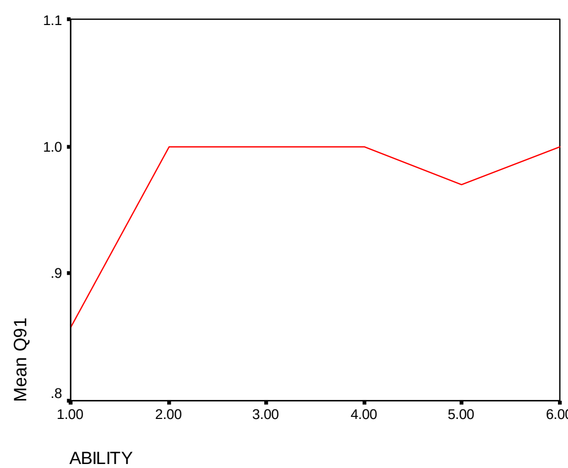

Mean Q80
.4
.3
.2
.1
0.0
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

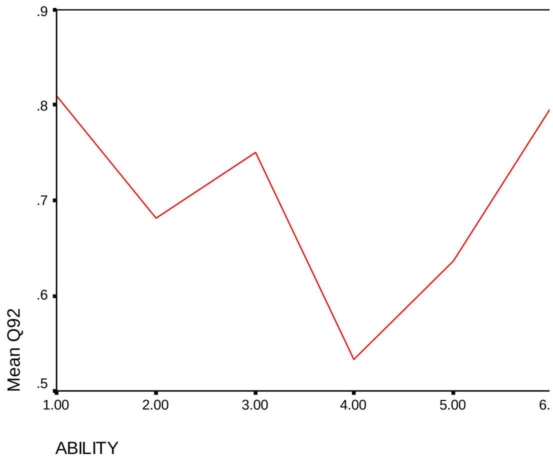

Mean Q86
1.1
1.0
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

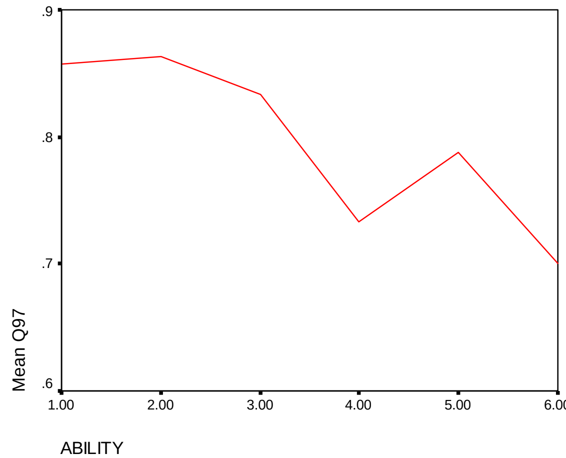

Mean Q82
1.1
1.0
.9
.8
.7
.6
.5
.4
.3
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

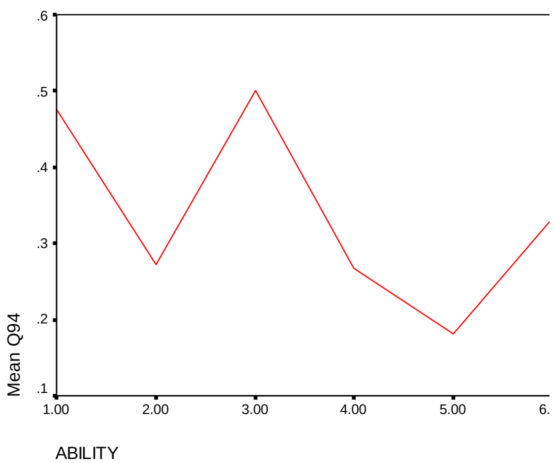

Mean Q87
1.1
1.0
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

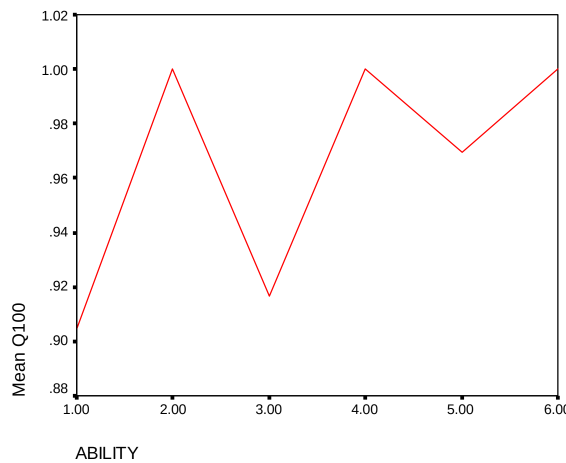

Mean Q83
.9
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.
ABILITY


Mean Q91
1.1
1.0
.9
.8
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY


.9
.8
.7
.6
.5
Mean Q92
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

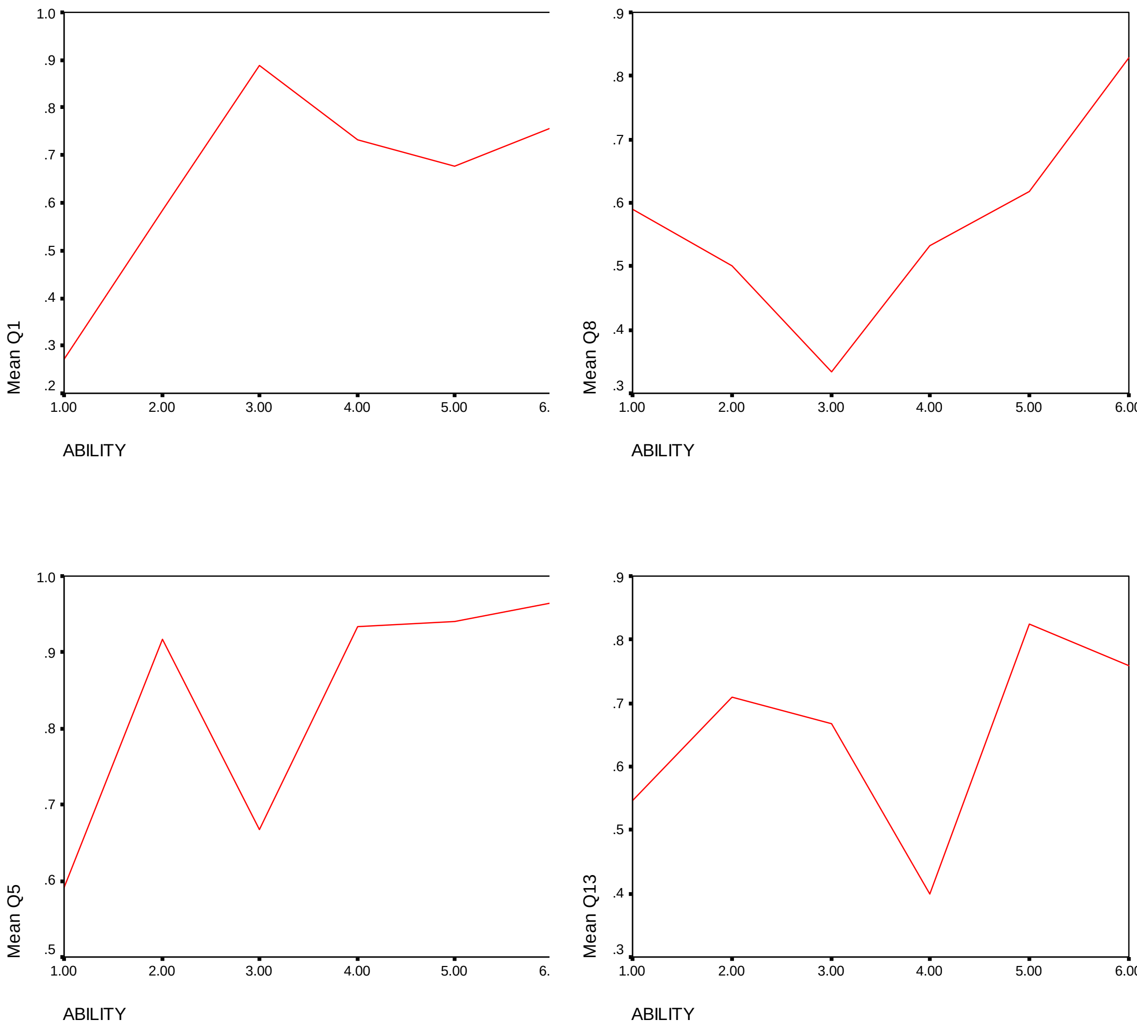

.9
.8
.7
.6
Mean Q97
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

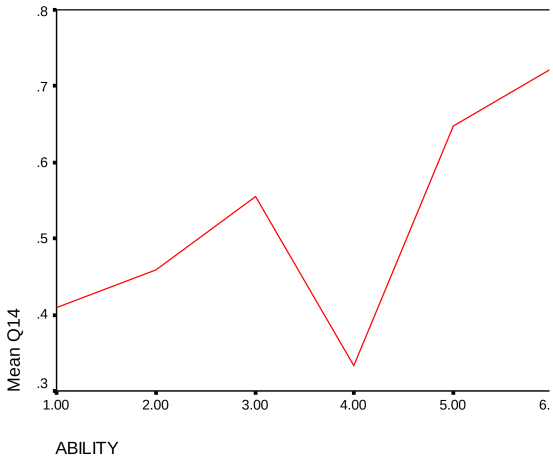

.6
.5
.4
.3
.2
.1
Mean Q94
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

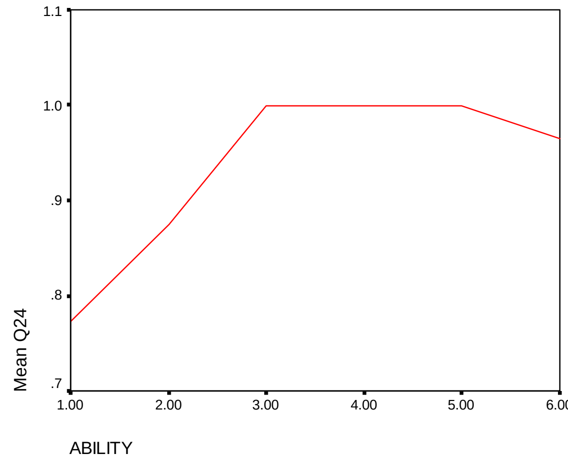

1.02
1.00
.98
.96
.94
.92
.90
.88
Mean Q100
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

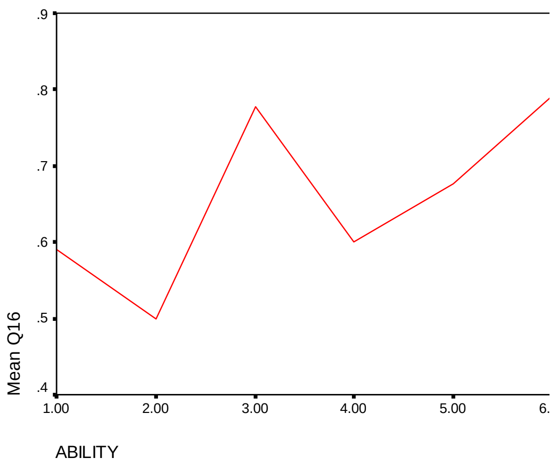
Appendix F

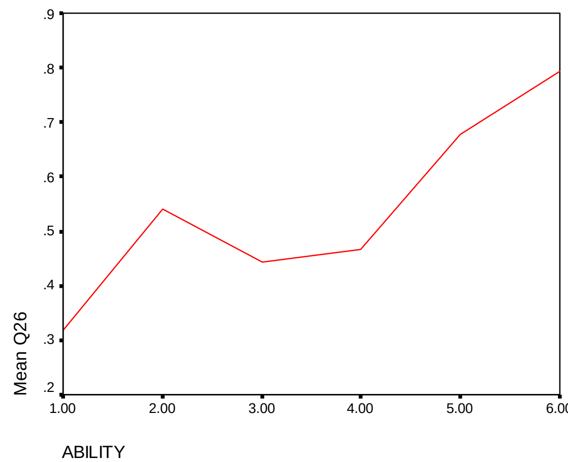
Item Characteristic Curves for Foils

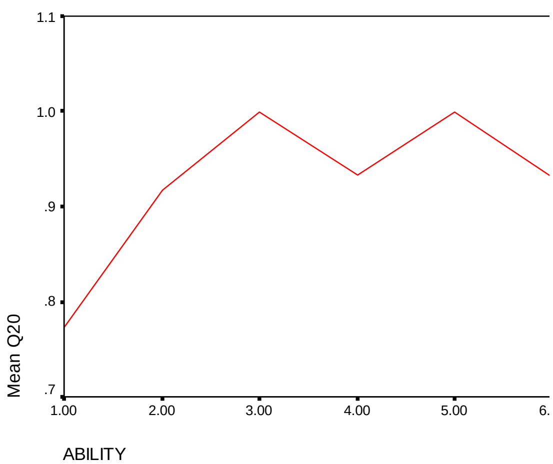


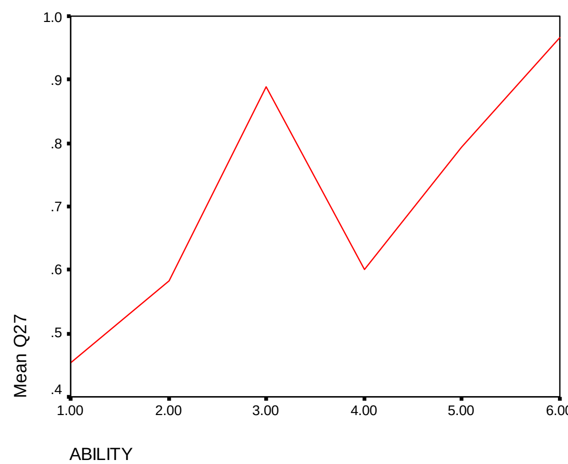

Mean Q14
.8
.7
.6
.5
.4
.3
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

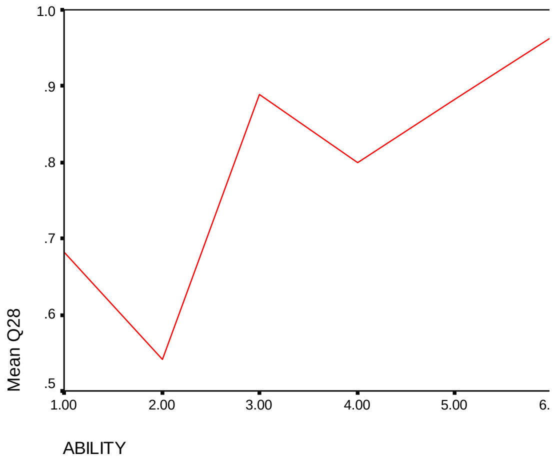

Mean Q24
1.1
1.0
.9
.8
.7
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

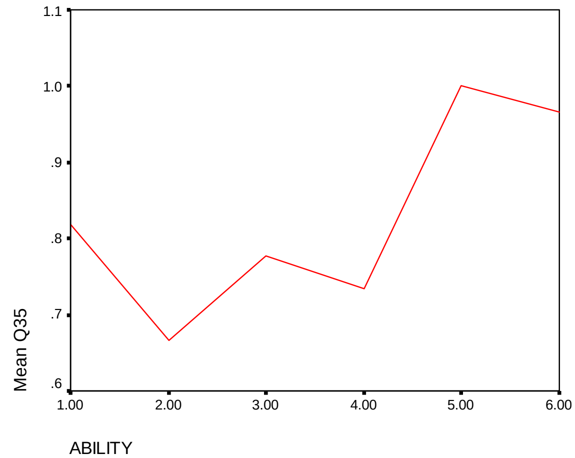

Mean Q16
.9
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

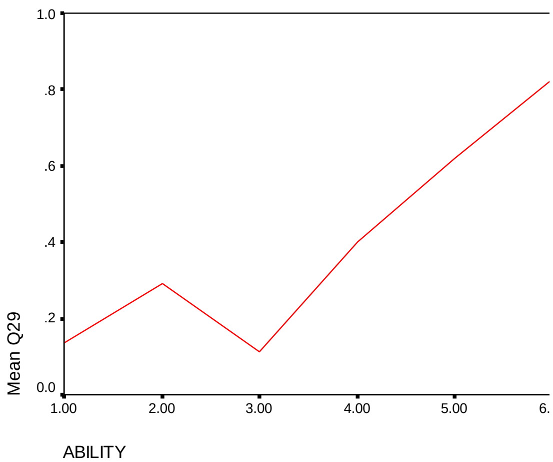

Mean Q26
.9
.8
.7
.6
.5
.4
.3
.2
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

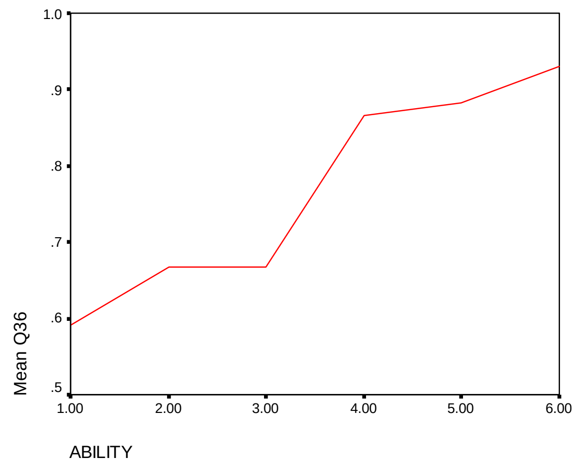

Mean Q20
1.1
1.0
.9
.8
.7
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

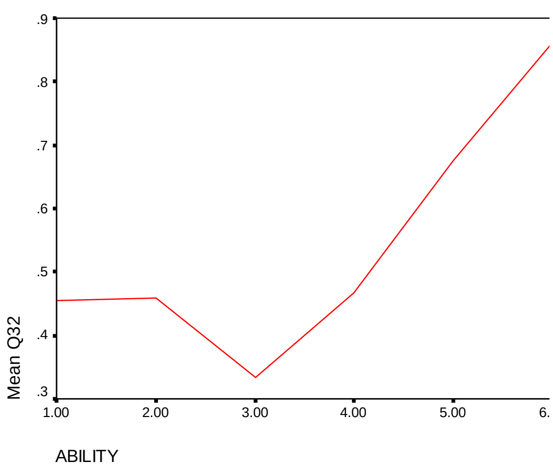

Mean Q27
1.0
.9
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

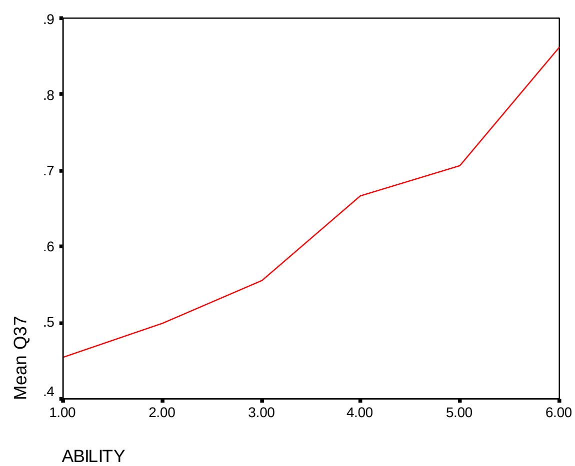

1.0
.9
.8
.7
.6
.5
Mean Q28
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

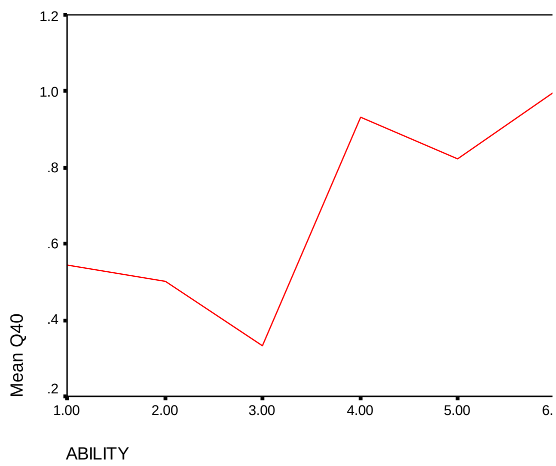

1.1
1.0
.9
.8
.7
.6
Mean Q35
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

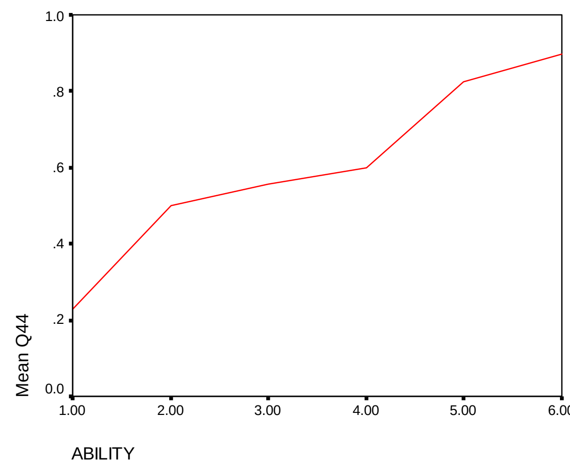

1.0
.8
.6
.4
.2
0.0
Mean Q29
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

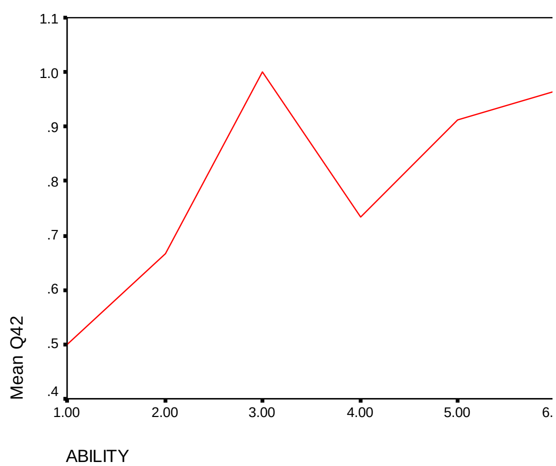

1.0
.9
.8
.7
.6
.5
Mean Q36
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

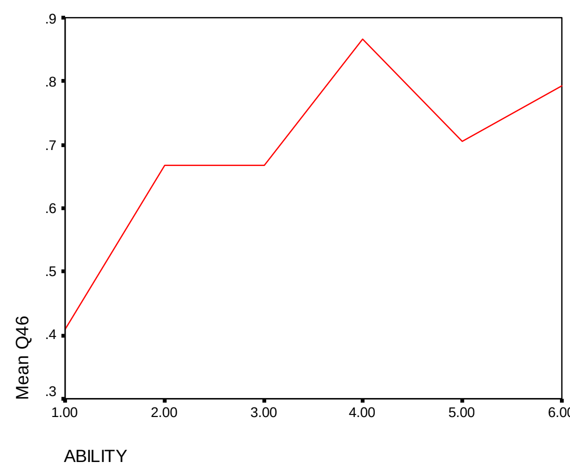

.9
.8
.7
.6
.5
.4
.3
Mean Q32
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

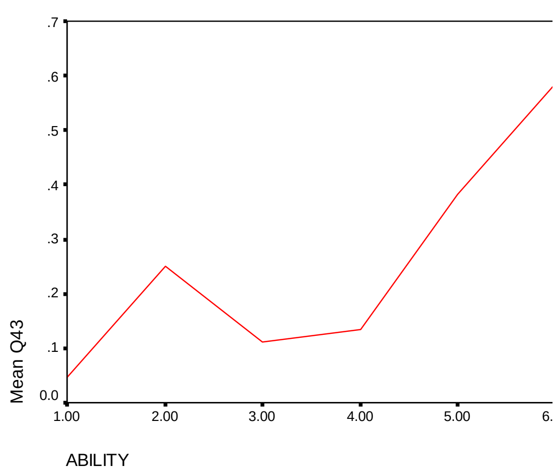

.9
.8
.7
.6
.5
.4
Mean Q37
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

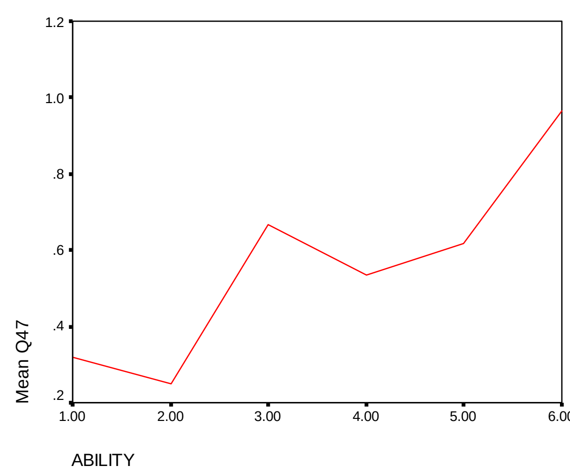

Mean Q40
1.2
1.0
.8
.6
.4
.2
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

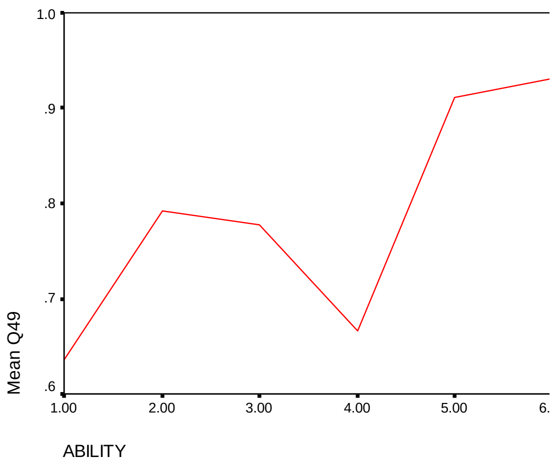

Mean Q44
1.0
.8
.6
.4
.2
0.0
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

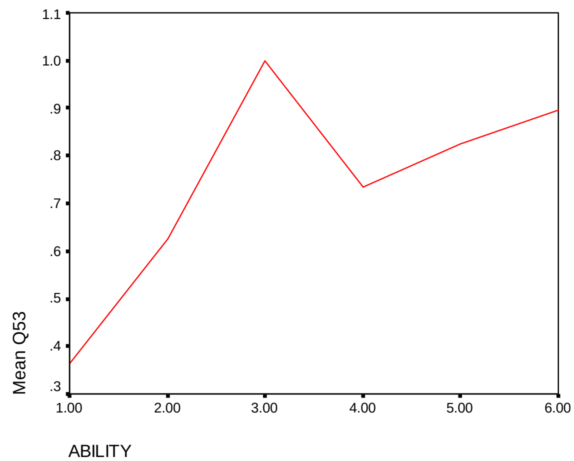

Mean Q42
1.1
1.0
.9
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

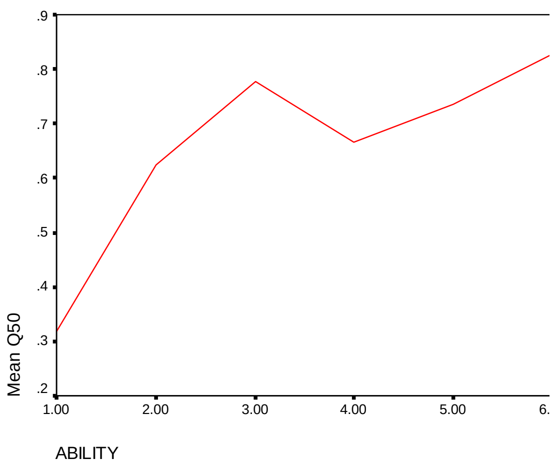

Mean Q46
.9
.8
.7
.6
.5
.4
.3
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

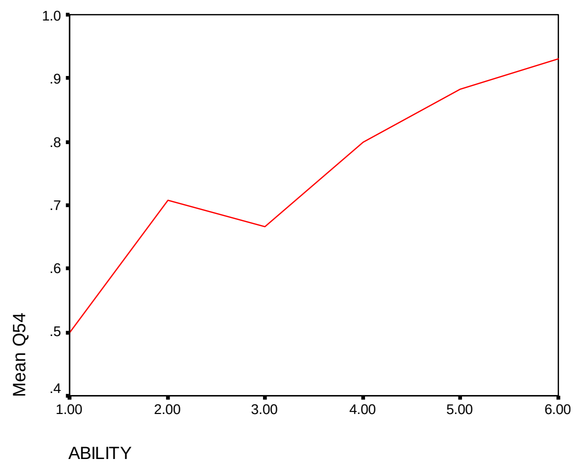

Mean Q43
.7
.6
.5
.4
.3
.2
.1
0.0
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

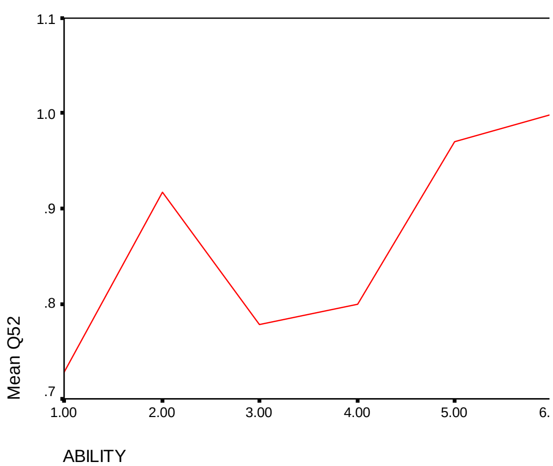

Mean Q47
1.2
1.0
.8
.6
.4
.2
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

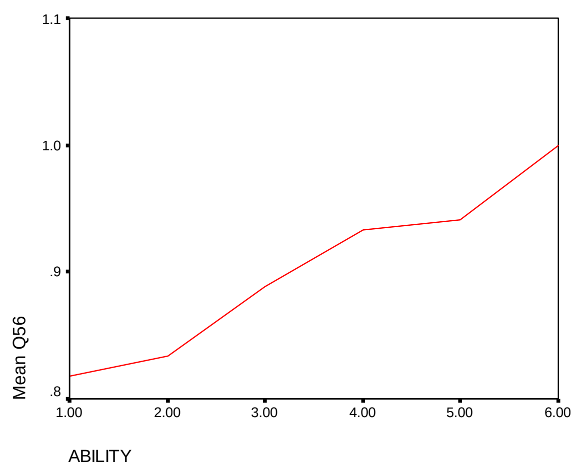

Mean Q49
1.0
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

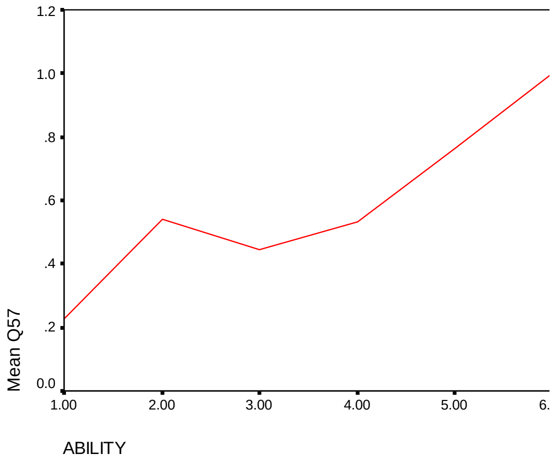

Mean Q53
1.1
1.0
.9
.8
.7
.6
.5
.4
.3
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

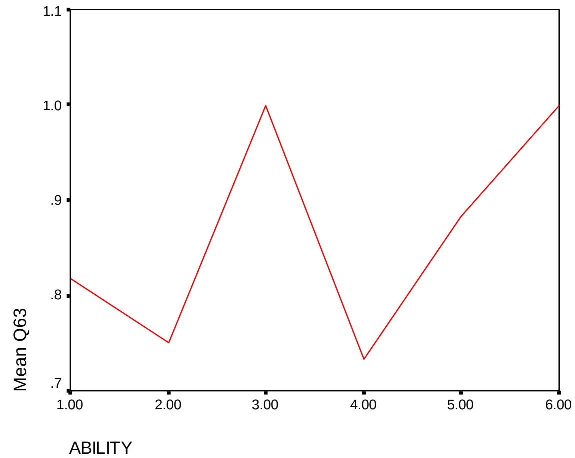

Mean Q50
.9
.8
.7
.6
.5
.4
.3
.2
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

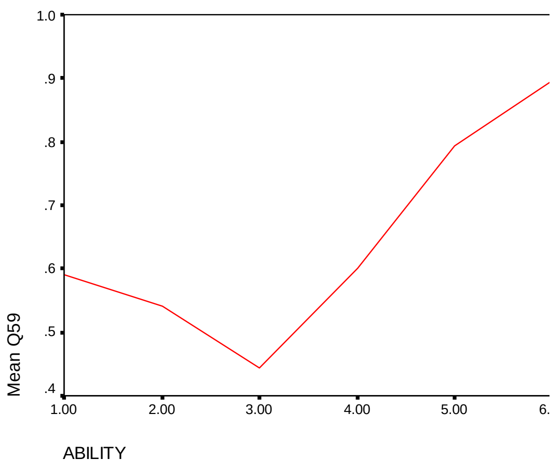

Mean Q54
1.0
.9
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

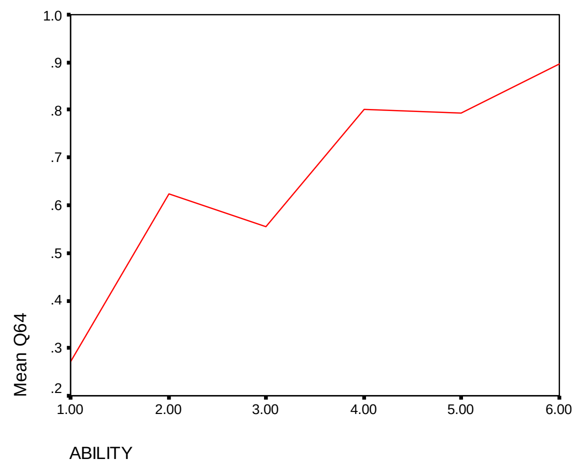

Mean Q52
1.1
1.0
.9
.8
.7
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

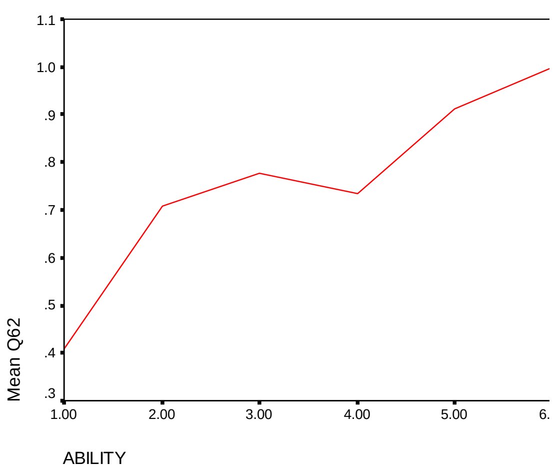

Mean Q56
1.1
1.0
.9
.8
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

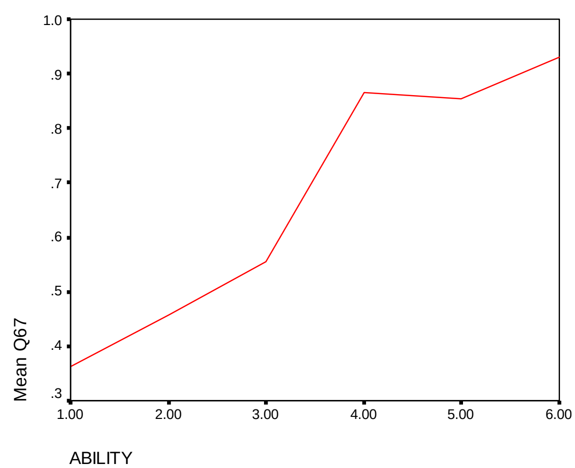

1.2
1.0
.8
.6
.4
.2
0.0
Mean Q57
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

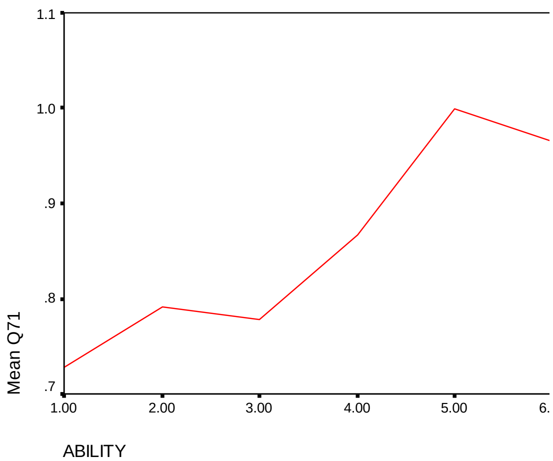

1.1
1.0
.9
.8
.7
Mean Q63
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

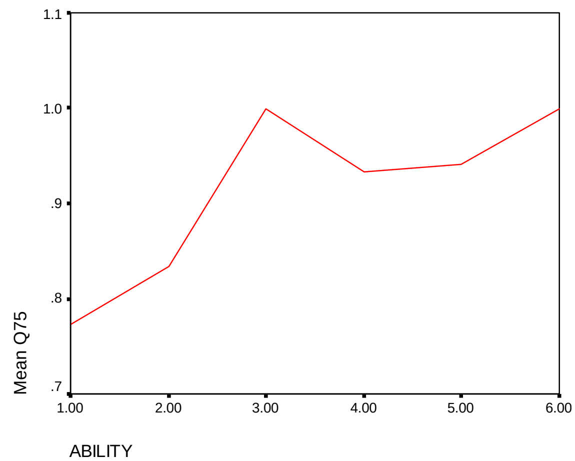

1.0
.9
.8
.7
.6
.5
.4
Mean Q59
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

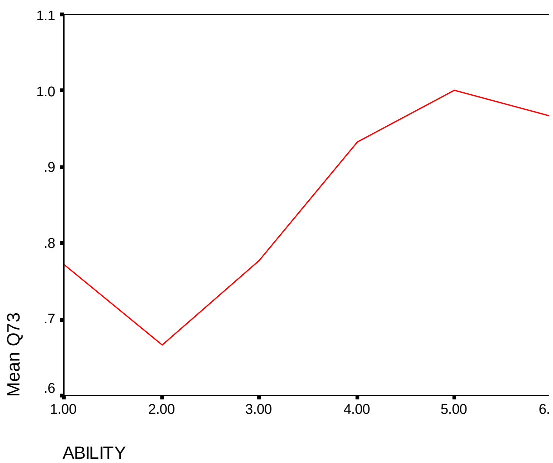

1.0
.9
.8
.7
.6
.5
.4
.3
.2
Mean Q64
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

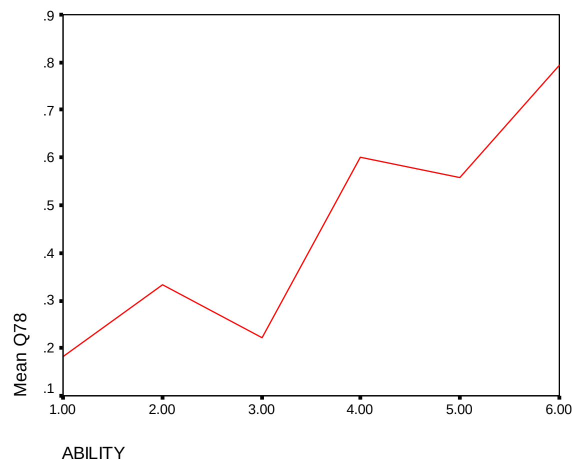

1.1
1.0
.9
.8
.7
.6
.5
.4
.3
Mean Q62
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

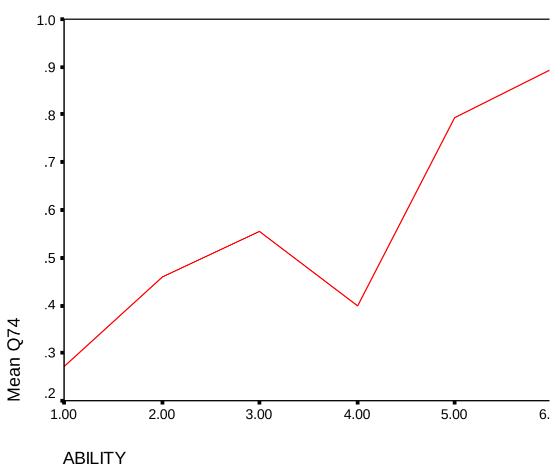

1.0
.9
.8
.7
.6
.5
.4
.3
Mean Q67
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

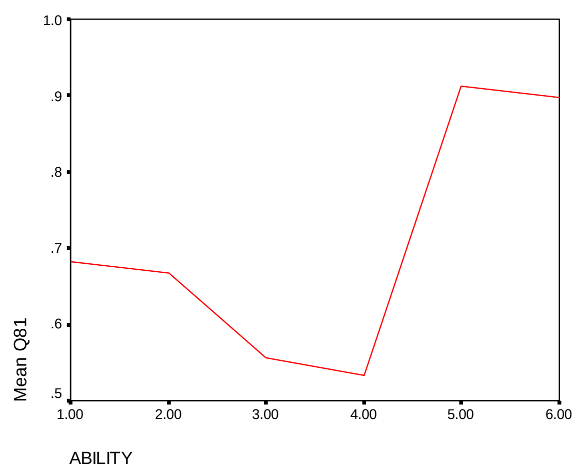

Mean Q71
1.1
1.0
.9
.8
.7
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

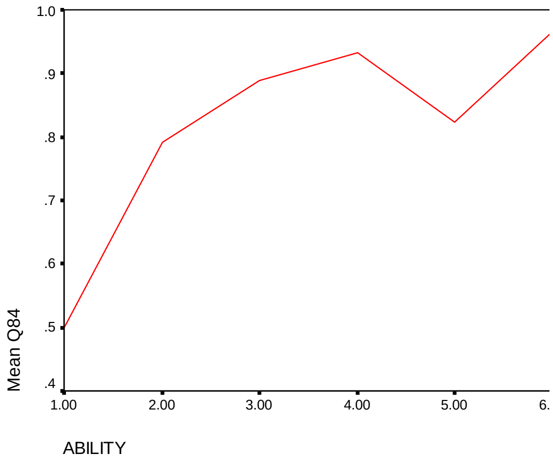

Mean Q75
1.1
1.0
.9
.8
.7
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY


Mean Q73
1.1
1.0
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

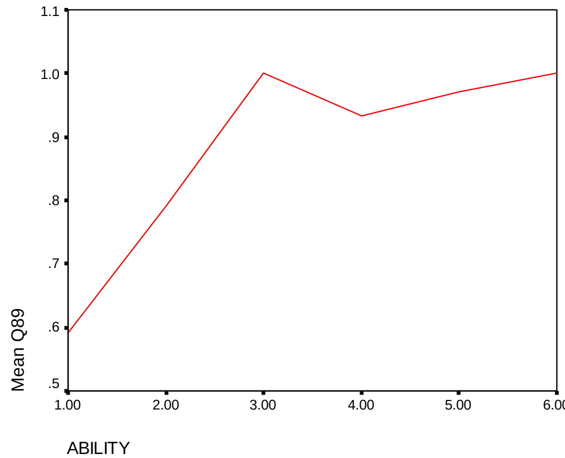

Mean Q78
.9
.8
.7
.6
.5
.4
.3
.2
.1
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY


Mean Q74
1.0
.9
.8
.7
.6
.5
.4
.3
.2
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

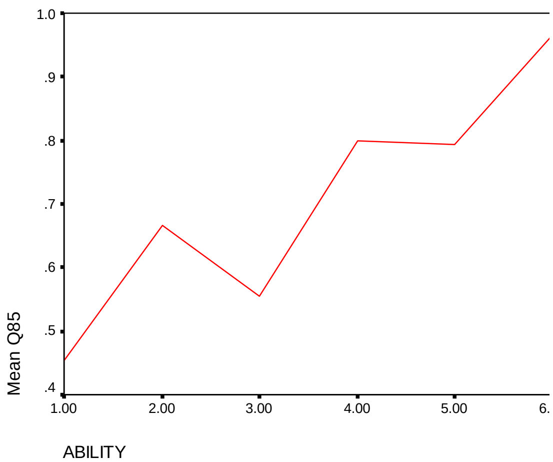

Mean Q81
1.0
.9
.8
.7
.6
.5
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY



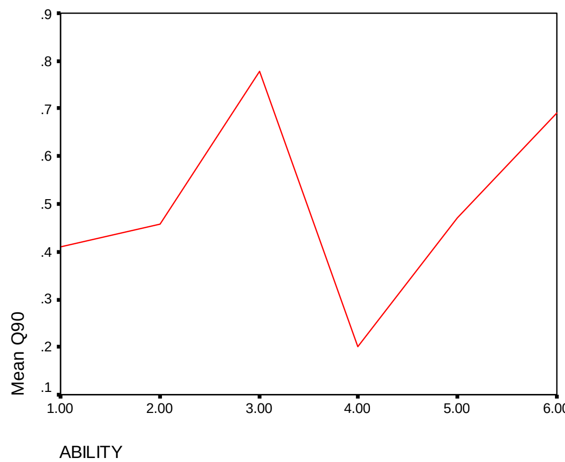
Mean Q84
1.0
.9
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.
ABILITY



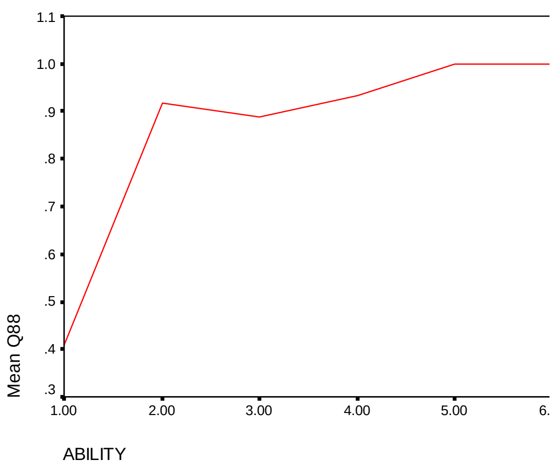
Mean Q89
1.1
1.0
.9
.8
.7
.6
.5
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY



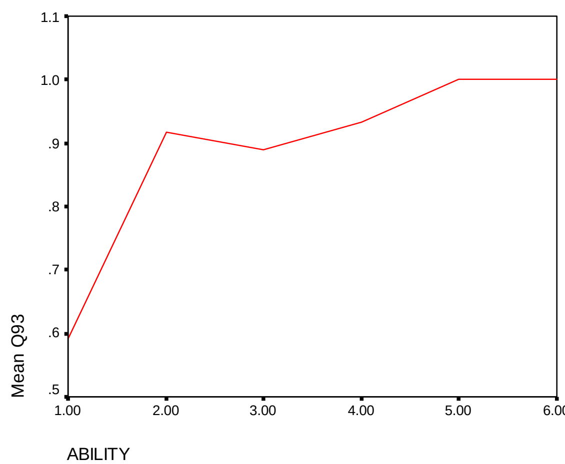
Mean Q85
1.0
.9
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.
ABILITY



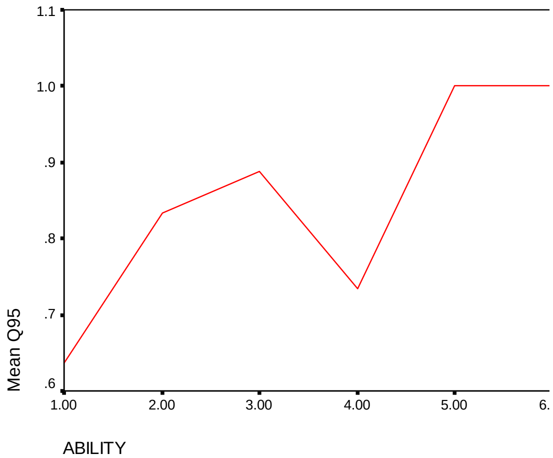
Mean Q90
.9
.8
.7
.6
.5
.4
.3
.2
.1
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

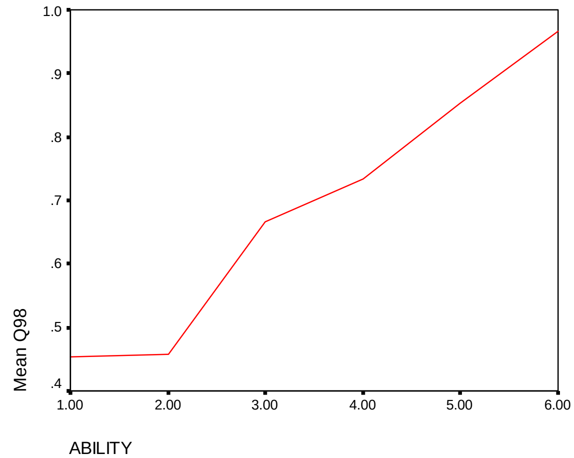


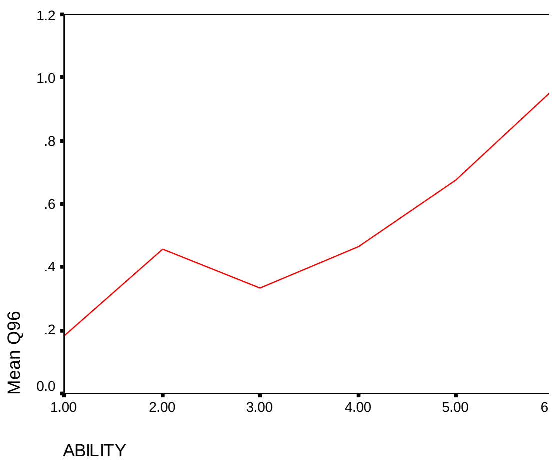
Mean Q88
1.1
1.0
.9
.8
.7
.6
.5
.4
.3
1.00
2.00
3.00
4.00
5.00
6.
ABILITY

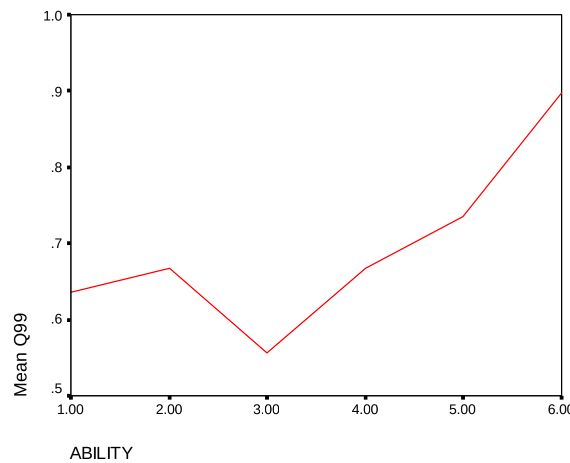


Mean Q93
1.1
1.0
.9
.8
.7
.6
.5
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY


Mean Q95
1.1
1.0
.9
.8
.7
.6
1.00
2.00
3.00
4.00
5.00
6.
ABILITY


Mean Q98
1.0
.9
.8
.7
.6
.5
.4
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY


Mean Q96
1.2
1.0
.8
.6
.4
.2
0.0
1.00
2.00
3.00
4.00
5.00
6
ABILITY


Mean Q99
1.0
.9
.8
.7
.6
.5
1.00
2.00
3.00
4.00
5.00
6.00
ABILITY

Appendix G

Scale Correlations for Key Terms

Correlations

		HITTOTAL
Q2	Pearson Correlation	.135
	Sig. (2-tailed)	.123
	N	133
Q3	Pearson Correlation	.233(**)
	Sig. (2-tailed)	.007
	N	133
Q4	Pearson Correlation	.168
	Sig. (2-tailed)	.053
	N	133
Q6	Pearson Correlation	.154
	Sig. (2-tailed)	.077
	N	133
Q7	Pearson Correlation	.039
	Sig. (2-tailed)	.655
	N	133
Q9	Pearson Correlation	.213(*)
	Sig. (2-tailed)	.014
	N	133
Q10	Pearson Correlation	.110
	Sig. (2-tailed)	.207
	N	133
Q11	Pearson Correlation	.141
	Sig. (2-tailed)	.106
	N	133
Q12	Pearson Correlation	.199(*)
	Sig. (2-tailed)	.022
	N	133
Q15	Pearson Correlation	.326(**)
	Sig. (2-tailed)	.000
	N	133
Q17	Pearson Correlation	.309(**)
	Sig. (2-tailed)	.000
	N	133
Q18	Pearson Correlation	.104
	Sig. (2-tailed)	.234
	N	133
Q19	Pearson Correlation	.230(**)
	Sig. (2-tailed)	.008
	N	133
Q21	Pearson Correlation	.291(**)
	Sig. (2-tailed)	.001
	N	133
Q22	Pearson Correlation	.202(*)
	Sig. (2-tailed)	.020
	N	133
Q23	Pearson Correlation	.053
	Sig. (2-tailed)	.546
	N	133
Q25	Pearson Correlation	.208(*)
	Sig. (2-tailed)	.016
	N	133
Q30	Pearson Correlation	.184(*)
	Sig. (2-tailed)	.034
	N	133
Q31	Pearson Correlation	.109
	Sig. (2-tailed)	.210
	N	133
Q33	Pearson Correlation	.186(*)
	Sig. (2-tailed)	.032
	N	133
Q34	Pearson Correlation	.365(**)
	Sig. (2-tailed)	.000
	N	133
Q38	Pearson Correlation	.301(**)
	Sig. (2-tailed)	.000
	N	133
Q39	Pearson Correlation	.192(*)
	Sig. (2-tailed)	.027
	N	133

Q41	Pearson Correlation	.386(**)	Q72	Pearson Correlation	.331(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q45	Pearson Correlation	.005	Q76	Pearson Correlation	.160
	Sig. (2-tailed)	.953		Sig. (2-tailed)	.065
	N	133		N	133
Q48	Pearson Correlation	-.102	Q77	Pearson Correlation	.139
	Sig. (2-tailed)	.243		Sig. (2-tailed)	.111
	N	133		N	133
Q51	Pearson Correlation	.120	Q79	Pearson Correlation	.245(**)
	Sig. (2-tailed)	.167		Sig. (2-tailed)	.005
	N	133		N	133
Q55	Pearson Correlation	.342(**)	Q80	Pearson Correlation	.205(*)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.018
	N	133		N	133
Q58	Pearson Correlation	.332(**)	Q82	Pearson Correlation	.185(*)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.033
	N	133		N	133
Q60	Pearson Correlation	.027	Q83	Pearson Correlation	.219(*)
	Sig. (2-tailed)	.756		Sig. (2-tailed)	.011
	N	133		N	133
Q61	Pearson Correlation	.279(**)	Q86	Pearson Correlation	.187(*)
	Sig. (2-tailed)	.001		Sig. (2-tailed)	.031
	N	133		N	133
Q65	Pearson Correlation	.222(*)	Q87	Pearson Correlation	.194(*)
	Sig. (2-tailed)	.010		Sig. (2-tailed)	.025
	N	133		N	133
Q66	Pearson Correlation	.233(**)	Q91	Pearson Correlation	.198(*)
	Sig. (2-tailed)	.007		Sig. (2-tailed)	.022
	N	133		N	133
Q68	Pearson Correlation	.339(**)	Q92	Pearson Correlation	.279(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.001
	N	133		N	133
Q69	Pearson Correlation	.062	Q94	Pearson Correlation	.323(**)
	Sig. (2-tailed)	.482		Sig. (2-tailed)	.000
	N	133		N	133
Q70	Pearson Correlation	.277(**)	Q97	Pearson Correlation	.170
	Sig. (2-tailed)	.001		Sig. (2-tailed)	.051
	N	133		N	133

Q100	Pearson Correlation	.173(*)
	Sig. (2-tailed)	.046
	N	133

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Appendix H

Scale Correlations for Foils

Correlations

		FOILTOTA
Q1	Pearson Correlation	.356(**)
	Sig. (2-tailed)	.000
	N	133
Q5	Pearson Correlation	.404(**)
	Sig. (2-tailed)	.000
	N	133
Q8	Pearson Correlation	.298(**)
	Sig. (2-tailed)	.000
	N	133
Q13	Pearson Correlation	.274(**)
	Sig. (2-tailed)	.001
	N	133
Q14	Pearson Correlation	.227(**)
	Sig. (2-tailed)	.009
	N	133
Q16	Pearson Correlation	.182(*)
	Sig. (2-tailed)	.036
	N	133
Q20	Pearson Correlation	.212(*)
	Sig. (2-tailed)	.014
	N	133
Q24	Pearson Correlation	.256(**)
	Sig. (2-tailed)	.003
	N	133
Q26	Pearson Correlation	.402(**)
	Sig. (2-tailed)	.000
	N	133
Q27	Pearson Correlation	.376(**)
	Sig. (2-tailed)	.000
	N	133
Q28	Pearson Correlation	.392(**)
	Sig. (2-tailed)	.000
	N	133
Q29	Pearson Correlation	.493(**)
	Sig. (2-tailed)	.000
	N	133
Q32	Pearson Correlation	.351(**)
	Sig. (2-tailed)	.000
	N	133
Q35	Pearson Correlation	.371(**)
	Sig. (2-tailed)	.000
	N	133
Q36	Pearson Correlation	.276(**)
	Sig. (2-tailed)	.001
	N	133
Q37	Pearson Correlation	.380(**)
	Sig. (2-tailed)	.000
	N	133
Q40	Pearson Correlation	.392(**)
	Sig. (2-tailed)	.000
	N	133
Q42	Pearson Correlation	.378(**)
	Sig. (2-tailed)	.000
	N	133
Q43	Pearson Correlation	.422(**)
	Sig. (2-tailed)	.000
	N	133
Q44	Pearson Correlation	.479(**)
	Sig. (2-tailed)	.000
	N	133
Q46	Pearson Correlation	.321(**)
	Sig. (2-tailed)	.000

	N	133		N	133
Q47	Pearson Correlation	.471(**)	Q71	Pearson Correlation	.397(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q49	Pearson Correlation	.236(**)	Q73	Pearson Correlation	.376(**)
	Sig. (2-tailed)	.006		Sig. (2-tailed)	.000
	N	133		N	133
Q50	Pearson Correlation	.383(**)	Q74	Pearson Correlation	.467(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q52	Pearson Correlation	.349(**)	Q75	Pearson Correlation	.346(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q53	Pearson Correlation	.445(**)	Q78	Pearson Correlation	.426(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q54	Pearson Correlation	.434(**)	Q81	Pearson Correlation	.334(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q56	Pearson Correlation	.270(**)	Q84	Pearson Correlation	.355(**)
	Sig. (2-tailed)	.002		Sig. (2-tailed)	.000
	N	133		N	133
Q57	Pearson Correlation	.525(**)	Q85	Pearson Correlation	.371(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q59	Pearson Correlation	.407(**)	Q88	Pearson Correlation	.532(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q62	Pearson Correlation	.489(**)	Q89	Pearson Correlation	.397(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q63	Pearson Correlation	.254(**)	Q90	Pearson Correlation	.227(**)
	Sig. (2-tailed)	.003		Sig. (2-tailed)	.009
	N	133		N	133
Q64	Pearson Correlation	.464(**)	Q93	Pearson Correlation	.420(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
	N	133		N	133
Q67	Pearson Correlation	.503(**)	Q95	Pearson Correlation	.419(**)
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000

	N	133
Q96	Pearson Correlation	.517(**)
	Sig. (2-tailed)	.000
	N	133
Q98	Pearson Correlation	.465(**)
	Sig. (2-tailed)	.000
	N	133
Q99	Pearson Correlation	.249(**)
	Sig. (2-tailed)	.004
	N	133

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Appendix I

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Q2	133	.00	1.00	.6165	.48807
Q3	133	.00	1.00	.7519	.43355
Q4	133	.00	1.00	.4135	.49433
Q6	133	.00	1.00	.8346	.37296
Q9	133	.00	1.00	.9925	.08671
Q10	133	.00	1.00	.9248	.26469
Q11	133	.00	1.00	.0677	.25213
Q12	133	.00	1.00	.9023	.29809
Q15	133	.00	1.00	.9023	.29809
Q17	133	.00	1.00	.9173	.27648
Q18	133	.00	1.00	.9774	.14905
Q19	133	.00	1.00	.8872	.31752
Q21	133	.00	1.00	.3910	.48981
Q22	133	.00	1.00	.9474	.22414
Q23	133	.00	1.00	.6165	.48807
Q25	133	.00	1.00	.9624	.19093
Q30	133	.00	1.00	.7519	.43355
Q31	133	.00	1.00	.9624	.19093
Q33	133	.00	1.00	.9624	.19093
Q34	133	.00	1.00	.7218	.44980
Q38	133	.00	1.00	.6692	.47229
Q39	133	.00	1.00	.8496	.35879
Q41	133	.00	1.00	.6241	.48620
Q45	133	.00	1.00	.9774	.14905
Q48	133	.00	1.00	.9925	.08671
Q51	133	.00	1.00	.9624	.19093
Q55	133	.00	1.00	.9323	.25213
Q58	133	.00	1.00	.6165	.48807
Q60	133	.00	1.00	.9850	.12216
Q61	133	.00	1.00	.3609	.48208
Q65	133	.00	1.00	.8722	.33515
Q66	133	.00	1.00	.9173	.27648
Q68	133	.00	1.00	.7669	.42439
Q69	133	.00	1.00	.9774	.14905
Q70	133	.00	1.00	.4662	.50074
Q72	133	.00	1.00	.5789	.49559
Q76	133	.00	1.00	.8647	.34338
Q77	133	.00	1.00	.5940	.49294
Q79	133	.00	1.00	.9774	.14905
Q80	133	.00	1.00	.2105	.40922
Q82	133	.00	1.00	.9023	.29809

Q83	133	.00	1.00	.6917	.46352
Q86	133	.00	1.00	.9173	.27648
Q87	133	.00	1.00	.8271	.37962
Q91	133	.00	1.00	.9699	.17144
Q92	133	.00	1.00	.7068	.45697
Q94	133	.00	1.00	.3158	.46659
Q97	133	.00	1.00	.7895	.40922
Q100	133	.00	1.00	.9699	.17144
Q1	133	.00	1.00	.6316	.48420
Q5	133	.00	1.00	.8647	.34338
Q7	133	.00	1.00	.6992	.46032
Q8	133	.00	1.00	.6090	.48981
Q13	133	.00	1.00	.6842	.46659
Q14	133	.00	1.00	.5489	.49949
Q16	133	.00	1.00	.6541	.47745
Q20	133	.00	1.00	.9248	.26469
Q24	133	.00	1.00	.9323	.25213
Q26	133	.00	1.00	.5789	.49559
Q27	133	.00	1.00	.7218	.44980
Q28	133	.00	1.00	.7970	.40376
Q29	133	.00	1.00	.4662	.50074
Q32	133	.00	1.00	.5940	.49294
Q35	133	.00	1.00	.8571	.35125
Q36	133	.00	1.00	.7895	.40922
Q37	133	.00	1.00	.6466	.47983
Q40	133	.00	1.00	.7368	.44201
Q42	133	.00	1.00	.7970	.40376
Q43	133	.00	1.00	.3008	.46032
Q44	133	.00	1.00	.6391	.48208
Q46	133	.00	1.00	.6842	.46659
Q47	133	.00	1.00	.5714	.49674
Q49	133	.00	1.00	.8120	.39217
Q50	133	.00	1.00	.6617	.47494
Q52	133	.00	1.00	.8947	.30805
Q53	133	.00	1.00	.7293	.44599
Q54	133	.00	1.00	.7744	.41953
Q56	133	.00	1.00	.9098	.28759
Q57	133	.00	1.00	.6391	.48208
Q59	133	.00	1.00	.6917	.46352
Q62	133	.00	1.00	.7820	.41448
Q63	133	.00	1.00	.8647	.34338
Q64	133	.00	1.00	.6842	.46659
Q67	133	.00	1.00	.6992	.46032
Q71	133	.00	1.00	.8797	.32654
Q73	133	.00	1.00	.8722	.33515
Q74	133	.00	1.00	.6090	.48981

Q75	133	.00	1.00	.9098	.28759
Q78	133	.00	1.00	.4887	.50176
Q81	133	.00	1.00	.7594	.42906
Q84	133	.00	1.00	.8120	.39217
Q85	133	.00	1.00	.7368	.44201
Q88	133	.00	1.00	.8722	.33515
Q89	133	.00	1.00	.8797	.32654
Q90	133	.00	1.00	.4962	.50188
Q93	133	.00	1.00	.9023	.29809
Q95	133	.00	1.00	.8722	.33515
Q96	133	.00	1.00	.5714	.49674
Q98	133	.00	1.00	.7143	.45346
Q99	133	.00	1.00	.7218	.44980
Valid N (listwise)	133				