DEVELOPMENT, RELIABILITY, AND VALIDITY OF A COMPUTERIZED ADAPTIVE VERSION OF THE SCHEDULE FOR NONADAPTIVE AND ADAPTIVE PERSONALITY

by

Leonard Jay Simms

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Psychology (Clinical Psychology) in the Graduate College of The University o f Iowa

August 2002

Thesis Supervisor: Professor Lee Anna Clark

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Leonard Jay Simms

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Psychology (Clinical Psychology) at the August 2002 graduation.

Thesis Committee: _____
Lee Anna Clark, Thesis Supervisor

_____
James Marchman

_____
Gregg Oden

_____
Walter Vispoel

_____
David Watson

ACKNOWLEDGEMENTS

I wish to thank my research mentor, Lee Anna Clark, for her unwavering support of this research. I also thank the other current and past members of my dissertation committee—Eva Klohnen, James Marchman, Gregg Oden, Walter Vispoel, and David Watson—for their helpful comments throughout the study. Finally, I thank the undergraduate research participants who kindly volunteered to participate in this study.

This work was supported through a research grant from the University of Minnesota Press.

ABSTRACT

Computerized adaptive testing (CAT) and Item Response Theory (IRT) techniques were applied to the Schedule for Nonadaptive and Adaptive Personality (SNAP) to create a more efficient measure with little or no cost to test reliability or validity. The SNAP includes 15 factor analytically derived and relatively unidimensional traits relevant to personality disorder. IRT item parameters were calibrated on item responses from a sample of 3,995 participants who completed the traditional paper-and-pencil (P&P) SNAP in a variety of university, community, and patient settings. Computerized simulations were conducted to test various adaptive testing algorithms, and the results informed the construction of the CAT version of the SNAP (SNAP-CAT). Live testing of the SNAP-CAT was conducted on a sample of 413 undergraduates who completed the SNAP twice, separated by one week. Participants were randomly assigned to one of four groups: (1) completed a modified P&P version of the SNAP (SNAP-PP) twice (n = 106), (2) completed the SNAP-PP first and the SNAP-CAT second (n = 105), (3) completed the SNAP-CAT first and the SNAP-PP second (n = 102), and (4) completed the SNAP-CAT twice (n = 100). Results indicated that the SNAP-CAT was 57.8% and 60.1% faster than the traditional P&P version, at Times 1 and 2, respectively, and mean item savings across scales were 36.3% and 36.7%, respectively. These savings came with little cost to reliability or validity, and the two test forms were largely equivalent. Descriptive statistics, rank-ordering of scores, internal factor structure, and convergent/discriminant validity were highly comparable across testing modes and methods of scoring, and very few differences between forms replicated across testing sessions. In addition, participants overwhelmingly preferred the computerized version to

the P&P version. However, several specific problems were identified for the Self-harm and Propriety scales of the SNAP-CAT that appeared to be broadly related to IRT calibration difficulties. Reasons for these anomalous findings are discussed, and follow-up studies are suggested. Despite these specific problems, the SNAP-CAT appears to be a viable alternative to the traditional P&P SNAP.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure

CHAPTER 1
STUDY RATIONALE AND LITERATURE REVIEW

Introduction

With the evolution of the computer over the past 40 years has come the recognition that computers have the power to improve our lives in innumerable ways. Without the massive computing power afforded by today's computers, major human advances such as life-saving brain imaging techniques (e.g., magnetic resonance imaging) as well as mapping of the human genome would not be possible. In the world of psychology, computers have given researchers the ability to "crunch" massive amounts of data more quickly than ever before; and now a small piece of desktop equipment can do what would have required a machine that filled an entire room not so long ago. Factor analysis, for example, once required researchers to spend months, even years, completing the calculations necessary to transform a modest covariance matrix into a meaningful factor structure. Today, computers can complete these calculations in a matter of milliseconds.

Computers have been helpful and influential in applied psychological settings as well. In the area of psychological assessment, for example, computerized tests increasingly are being used for the measurement of personality (e.g., Butcher, 1987; Karson & O'Dell, 1987; Lachar, 1987; Waller & Reise, 1989), intellectual abilities (e.g., Weiss & Vale, 1987), neuropsychological status (e.g., Golden, 1987), and vocational interests (e.g., Hansen, 1987). The earliest computerized applications simply administered and provided scores for traditional paper-and-pencil tests (e.g., Lushene, O'Neil, & Dunn, 1974); others, however, went further by providing computerized

interpretive reports for each examinee (Butcher, Perry, & Atlis, 2000; Fowler, 1985). The proper and ethical use of computerized interpretive reports has been a topic of great debate, and although this issue is not a focus of the current review, interested readers are referred to several recent articles (Butcher et al., 2000; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Snyder, 2000) for discussion of the issue.

In a review of various issues related to computerized assessment, Butcher (1987) described the following advantages: (a) *objectivity*--computers are unbiased when they administer, score, or interpret a psychological test, (b) *speed*--computers can score and interpret a psychological test more quickly than a clinician, and computerized administration of tests is often less time-consuming than traditional paper-and-pencil methods, (c) *efficiency*--due to its increased speed and accuracy, the computer can be more cost-effective than a technician or clinician, and (d) *reliability*--the computer's reliability in scoring and interpreting tests is always 1.00. That is, given the same item responses, the computer will always assign the same score and generate the same interpretive report.

A number of personality and measurement researchers (e.g., Forbey, Handel, & Ben-Porath, 2000; Roper, Ben-Porath, & Butcher, 1991, 1995; Reise & Henson, 2000; Waller & Reise, 1989; Zickar, 2001) have discussed or attempted to use the flexibility afforded by personal computers to increase assessment efficiency. Borrowing from the ability and achievement testing community, these researchers have begun to explore adaptive administration of personality items by computer. In the most basic sense, computerized adaptive tests (CATs) select and administer items that are individually tailored to the trait level of the examinee, with the potential of substantial item and time

savings (Weiss, 1985). The primary purpose of the present project was to use CAT methodology to increase the efficiency with which a particular multi-scale personality inventory, the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993), is administered and scored.

Equivalence of Computerized Assessments

The equivalence of computer-administered tests to their paper-and-pencil counterparts cannot be assumed (Honaker, 1988; Wilson, Genco, & Yager, 1985). Indeed, the literature concerning the equivalence between paper-and-pencil tests and their conventional computer counterparts has yielded wildly discrepant results across studies. Before examining these conflicting results in detail, it is necessary to discuss the concept of equivalence. In this domain, the most common way to define equivalence operationally is through the concept of *psychometric equivalence*, or parallel forms. As outlined by a number of authors (e.g., Ghiselli, Campbell, & Zedeck, 1981, pp.192-227; Hofer & Green, 1985), two tests can be considered psychometrically equivalent, or parallel, if they yield equal descriptive statistics (i.e., mean-level stability) and ranking of scores (i.e., rank-order stability), and if they correlate to the same degree with scores on other variables. If two forms of the same test meet these criteria, both may share the same normative and validity data for the purposes of score interpretation (Honaker, 1988). If the two forms yield only different descriptive statistics, simple algebraic score transformations may be able to equate the two tests for the purposes of score interpretation. If, however, the two forms do not yield similar rankings or do not correlate equally well with other important extra-test variables, then it is likely that the

two forms of the test are actually measuring two different constructs, and separate normative and validity data may be necessary for each (Honaker, 1988).

The equivalence of test forms also can be evaluated in terms of *experiential equivalence* (Honaker, 1988). Experiential equivalence refers to the ways in which the test forms are experienced by the examinees (e.g., emotional, perceptual, and attitudinal reactions to the two test forms). Test forms can be psychometrically equivalent yet different experientially in ways which may make the forms nonequivalent in the overall context of the assessment process (Honaker, 1988). For example, consider a test for which two parallel forms exist; despite being psychometrically equivalent, one form is consistently rated as more aversive than the other by examinees. The greater aversiveness associated with that form may influence the client's attitude toward further assessments as well as overall rapport between the clinician and the client. Thus, determinations of test equivalence should take into account both psychometric and non-psychometric considerations.

A number of common psychological tests have computerized versions, but none has been in use as long or has been as widely studied as the those of the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1951) and its second edition (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). As far back as the 1960s, researchers at the Mayo Clinic and the University of Alabama independently created computer systems for automated MMPI scoring and interpretation (Fowler, 1985), and over 25 years ago, Lushene, O'Neil, and Dunn (1974) published the first equivalency study. They evaluated a completely computerized version of the booklet form of the MMPI and found that conventional computerized (CC)

administration of the original MMPI resulted in significant scale score differences on all scales, except *F*, *Mf*, and *Ma*, with differences generally in the direction of higher scores on paper-and-pencil (P&P) administration. However, these differences tended to be small (about two raw score points) and not clinically significant.

A number of studies that followed (Biskin & Kolotkin, 1977; Hart & Goldstein, 1985; Schuldberg, 1988, 1990) also found significant differences on various scales, but these differences were highly inconsistent across studies and far fewer in number. Biskin and Kolotkin (1977) found small but significant differences for the *Pa* scale (lower on CC administration). They also found that the CC version yielded significantly more "cannot say" responses than did the P&P version. Hart and Goldstein (1985), using a between-subjects design, found the CC and P&P versions to be largely comparable with the exception of the *K* scale (where CC participants scored somewhat lower).

More recently, Schuldberg (1988, 1990) investigated whether participants differed in their responses to individual items across the CC and P&P formats. He found that participants responded differently on 40 items, tending to acknowledge less pathology on the CC version. Correspondingly, he found *F* scale scores to be lower with CC testing (Schuldberg, 1988) and that participants used the "cannot say" option more often on the CC version (Schuldberg, 1990). Based on these modest significant differences, Schuldberg concluded that repeated testing was a more salient variable concerning response changes and that factors such as carelessness and maladjustment were more important than administration format.

In contrast, other studies (Honaker, Harrell, & Buffaloe, 1988; Pinsoneault, 1996; Rozensky, Honor, Rasinski, Tovian, & Hertz, 1986; Russell, Peace, & Mellsop, 1986)

have found no significant mean scale score differences across P&P and CC formats. Still others (Lambert, Andrews, Rylee, & Skinner, 1987; White, Clements, & Fowler, 1985) found no significant differences on the clinical and validity scales but yielded significantly different "cannot say" scores. The most recent study to find a lack of significant scale score differences (Pinsoneault, 1996) found the CC and P&P formats to be quite comparable. Neither the validity nor clinical scales differed in terms of descriptive statistics, and test-retest reliabilities between the two formats were significant and compared favorably with those previously reported for repeated P&P testing.

Given the discrepant findings outlined above, a number of researchers have hypothesized about the interpretation of the findings. In the first critical review of the computerized testing literature, Honaker (1988) suggested that many of these studies have various methodological problems. First, many of the studies failed to specify exactly *how* the MMPI was administered by computer. Second, he commented that many of the computer programs used to administer the MMPI varied somewhat with respect to how responses were made on the computer and differed considerably with respect to the response options available to subjects. Given this variation, Honaker suggested that the research up to that point had addressed the potential equivalency of several alternative forms of computer administration. He concluded that these studies, given their methodological problems, have not adequately demonstrated the comparability of these computer-administered formats.

In a somewhat more quantitative review of the literature, Watson et al. (1990) reported a mean scale score reduction of 1.2 *T*-score points across nine studies comparing standard with computerized MMPI administration, a difference that is generally less than

a test-taker's response to a single item. They noted that these differences were well within the standard error of measurement of MMPI scales and were not associated with a change in scale interpretation, suggesting high comparability between computerized and booklet formats. Two years later, Watson and his colleagues (Watson, Thomas, & Anderson, 1992) published a more systematic meta-analysis of the same nine equivalency studies, providing a total of 967 subjects for analysis. They found significant differences on 8 of the 13 basic and validity scales, and differences approaching significance on three of the remaining five scales. However, they noted that the differences were small and averaged only about 1.5 *T*-score points. Interestingly, though, the authors concluded from these data that the CC version of the MMPI might be improved by developing separate norms.

Most recently, in the most complete and sophisticated analysis to date, Finger and Ones (1999) applied psychometric meta-analysis (Hunter & Schmidt, 1990) to pool results across all available studies to examine the equivalence of the P&P and CC forms of the MMPI. In addition to mean effect sizes, the authors also calculated and compared standard deviations and cross-form correlations across studies. Once the authors statistically accounted for sampling error across individual studies, differences in means and standard deviations across studies were near zero, and the cross-form correlations were near perfect. The authors concluded that the CC and P&P forms of the MMPI are psychometrically equivalent, and moreover, suggested that the disparate findings found in the literature to this point can be explained in terms of sampling error across individual studies.

Computerized Adaptive Testing

Computerized adaptive testing (CAT; Weiss, 1985; Weiss & Vale, 1987) is a subtype of computerized testing that combines the speed and flexibility of computerized assessment with the power and efficiency afforded by item response theory (IRT). The critical difference between CAT and other forms of standard and computerized assessment is its ability to administer individually tailored tests. A typical CAT selects and administers only those items that provide the most psychometric information (i.e., have the lowest standard errors of measurement) at a given trait level, eliminating the need to administer items that have very low or very high endorsement probabilities given a particular examinee's trait level. For example, consider a hypothetical scale composed of 40 items designed to assess neuroticism. In contrast to traditional testing (in which all 40 items must be administered in order to obtain a score), CATs may be able to administer a subset of 20 highly informative items, tailored to given examinee's trait level, without any loss of measurement precision. This greater efficiency, represented by item and time savings of 50% in this hypothetical example, is the primary advantage of CAT over other forms of traditional and computerized assessment.

A typical IRT-based CAT includes three interrelated elements: (a) a procedure for estimating a given examinee's trait or ability level, (b) a procedure for selecting items from the item pool, and (c) a termination rule to determine when testing should be discontinued (Waller & Reise, 1989). In practice, CAT begins with the administration of an item representative of the median trait level. The computer then scores that item, calculates an estimate of the trait level, and then determines whether the termination rule has been satisfied. If not, the computer administers a new item that will provide

maximum information at the newly calculated trait level, scores the item, re-estimates the trait level, and determines whether the termination rule has been satisfied. This cycle of item selection and administration, item scoring, and trait level estimation continues until the termination rule has been satisfied.

Non-IRT CAT

One potential problem with IRT involves its assumption of scale unidimensionality (to be discussed in greater detail later). In short, multidimensionality precludes the use of most IRT-based measurement models. Thus, researchers who wish to adapt existing measures into IRT-based CATs must assess this assumption before proceeding with an IRT-based CAT. Popular instruments like the MMPI and the California Personality Inventory (CPI; Gough, 1975) were constructed using empirical criterion-keying, which generally results in highly heterogeneous and multidimensional scales (Ben-Porath & Butcher, 1986). Within the ability testing domain, Thomas (1990) tested several methods for utilizing a multidimensional item pool with an IRT-based CAT. He found that presenting the examinee with several shorter "mini-CATs" can provide a reasonably accurate estimation of ability while ameliorating the problems that usually result from multidimensional item pools. Such methods have not yet been extended to multidimensional personality scales such as those from the MMPI and CPI.

Given the heterogeneity of MMPI scales, an alternative method—known as the *countdown method*—has been developed to administer the MMPI adaptively (Butcher, Keller, & Bacon, 1985). The countdown method can generate substantial item savings if the clinical assessment question involves only the classification of individuals into two groups: those with either clinically elevated or non-elevated scores. In one permutation

of this method, the *classification procedure* (Butcher et al., 1985), items are administered on a given scale only until a clinically meaningful elevation (e.g., *T*-score = 65) on that scale is either certain or impossible. For example, suppose a given scale contains 20 items, and 10 items endorsed in the keyed direction is the cutoff for a clinically meaningful elevation. If an individual endorses 11 items in the non-keyed direction, clinically significant elevation becomes impossible, and scale administration is terminated. Conversely, if the individual endorses 10 items in the keyed direction, clinical elevation is certain, and scale administration is then terminated. An alternative countdown method, termed the *full scores on elevated scales procedure* (Ben-Porath, Slutske, & Butcher, 1989), is identical to the classification procedure except that all items are administered for scales on which scale elevation is certain.

Roper, Ben-Porath, and Butcher (1991) utilized the countdown method to administer a CAT and P&P versions of the MMPI-2 to 155 college students over a one-week interval. In order to provide CAT means that could be directly compared to the P&P means, they administered all skipped items once the adaptive test was completed. This strategy resulted in significantly different scores on scales *F*, *K*, *Si*, as well as several content scales across the two forms. Further, the authors reported significantly lower CAT-P&P retest correlations for scales *Ma*, *Hs*, *Mf*, *Pt*, and *Sc* than have been reported in other published P&P retest studies of the MMPI-2. Despite these differences, Roper et al. (1991) noted that none of the mean score differences reached clinical significance and that the clinical profiles generated across the two modalities would yield identical clinical interpretations. Thus, they concluded that for practical purposes (i.e., clinical interpretation), the CAT and P&P versions of the MMPI-2 were equivalent. The CAT

version yielded item savings ranging from 26.7% to 28.7%, depending on the specific adaptive procedure employed.

In a follow-up study, Roper, Ben-Porath, and Butcher (1995) improved upon the first study by including direct comparison groups and assessing convergent validity with other established measures. They randomly assigned a sample of 571 undergraduates to one of three experimental groups: (a) the *P&P retest* group completed the standard P&P version of the MMPI-2 twice; (b) the *P&P-CAT* group completed the MMPI-2 once by PP and once by CAT; and (c) the *CC-CAT* group completed the MMPI-2 once by conventional computerized (CC) testing (i.e., computerized administration of the whole instrument) and once by CAT. All administrations were separated by one week, and order of administration was counterbalanced in each group. The mean scale differences that resulted from the CAT administration were no greater than those found in repeated administration of the P&P form. Further, they found no significant mean scale score differences in the CC-CAT condition. Finally, to examine convergent validity, the authors computed the correlations between the MMPI-2 scales and several other common psychological measures (e.g., the Beck Depression Inventory), finding no significantly different correlations across the three administration modes. Once again, the CAT version yielded significant item savings, ranging from 30% to 34%.

Handel, Ben-Porath, and Watt (1999) extended this line of research to a sample of 140 inpatients and outpatients being treated for substance abuse and other addictive behaviors. Participants completed the MMPI-2 twice, in one of two conditions: (a) CAT vs. CC (CAT-CC), and (b) CC test-retest (CC-CC). The results revealed that the two forms were largely comparable, with two notable exceptions: compared to the CC-CC

group, the CAT-CC group produced significantly lower *F* scale scores during CA testing as well as a significantly lower CAT-CC correlation for the *Mf* scale. Again, the most impressive results were the item (31.5%) and time (31.6%) savings achieved in the CAT-CC condition. Recently, Ben-Porath's group (Forbey, Handel, & Ben-Porath, 2000) further extended these techniques to the adolescent version of the MMPI (MMPI-A; Butcher et al., 1992) by conducting real data CAT simulations on existing MMPI-A datasets. Looking at various administration formats and termination criteria, Forbey et al. achieved mean simulated item savings ranging from 10.7% to 26.4%, which is somewhat lower than that found with the MMPI-2 using similar procedures. The authors attributed this drop to the MMPI-A's smaller item pool.

The preceding studies largely support the equivalence of the computerized adaptive version of the MMPI-2 and MMPI-A when compared to both conventional computer and standard paper-and-pencil administration. However, given the heterogeneous nature of MMPI-2 and MMPI-A scales, these studies utilized adaptive algorithms that are not state-of-the-art. Thus, while these studies provide promising evidence that adaptive personality testing can yield substantial item savings without a significant loss of reliability or validity, they have done so without the efficiency and precision afforded by IRT-based adaptive testing algorithms (Weiss, 1985).

Item Response Theory

The term IRT applies to a group of psychometric models that characterize test items by one or more item parameters (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). These item parameters define an item characteristic curve (ICC; see Figure 1.1 for a prototypic ICC). An ICC describes the

regression of the probability of a particular item response on an underlying trait. In IRT models, this trait is referred to as *theta* (θ). The ICC can be defined by one, two, or three of the following item parameters (Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Lord, 1980): *item discrimination* (*a*; also referred to as "slope"), *item difficulty* (*b*; sometimes referred to as "item threshold"), and the *pseudoguessing parameter* (*c*). Item difficulty refers to the point along the theta (trait or ability level) continuum that is associated with a 50% probability that a given examinee, at that theta level, will respond to the item in the keyed direction (after adjusting the curve for guessing when the *c* parameter is included in the item response model). High values of item difficulty are associated with items that have low endorsement probabilities (i.e., items that are more difficult in ability or achievement testing, or that reflect higher levels of the trait in question). Item discrimination reflects the slope of the ICC (i.e., the probability of a keyed response as a function of trait level) at the difficulty level for the item (which is the point of maximum item discrimination). Steeper slopes reflect greater discriminatory power. Finally, the pseudoguessing parameter is used to adjust the ICC to account for the fact that individuals who are very low on a trait sometimes answer items in the keyed direction solely on the basis of chance. These three parameters are represented in Figure 1.1.

Depending upon the particular application, one, two, or three of these parameters can be modeled using IRT (Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Lord, 1980). The one-parameter model, also known as the Rasch model (Rasch, 1960; Wright, 1977), assumes constant item discrimination and allows only item difficulty to vary across items. The two-parameter model (Birnbaum, 1968) allows items to vary both

in difficulty and discrimination. The three-parameter model includes difficulty and discrimination, as well as the pseudoguessing parameter described above. While the Rasch model certainly has its vocal proponents (e.g., Wright, 1977), those who value closer model fit over parsimony have criticized it for being overly simplistic (e.g., Embretson & Hershberger, 1999). The three-parameter model, with its pseudoguessing parameter, is ideally suited to applications such as ability or achievement testing, where accounting for guessing or chance responding is important for appropriate model fit. In applications where guessing is not as viable a concept (e.g., personality testing), the two-parameter model has been shown to be both reasonable and powerful (e.g., Kamakura & Balasubramanian, 1989; Reise, 1999; Reise & Waller, 1990; Waller, 1999; Waller & Reise, 1989). Interestingly, however, one recent study (Rouse, Finger, & Butcher, 1999) applied the three-parameter model to the Personality Psychopathology (PSY-5) scales of the MMPI-2 and obtained *c* values for some items that were significantly correlated with independent ratings of item social desirability. These relations were not entirely consistent, however, and further research is necessary to determine whether the *c* parameter is meaningful for personality scales.

The formula for the two-parameter logistic model (Birnbaum, 1968; Reise & Waller, 1990; Waller & Reise, 1989) is represented as follows:

$$P_i(\theta) = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

where $P_i(\theta)$ is the probability of a keyed response to item *i* at a given level of the underlying trait, theta (θ); θ is the continuous latent trait underlying test performance; *a* is the item discrimination parameter for item *i*; *b* is the item difficulty parameter for item

i, and *D* is a scaling constant that is often set to 1.7 to approximate the model to the normal ogive function.

The parameters of an ICC can be combined into a single index that describes how precisely an item measures the trait at various points along the trait continuum (Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Lord, 1980). This index is referred to as *item information* and can be represented graphically in an *item information curve* (IIC). On an IIC, item information is plotted as a function of trait level (see Figure 1.2 for several prototypic IICs). An IIC has its peak at the difficulty of the item, and the height of its peak is positively related to the item discrimination. In addition, the spread of the IIC is influenced by item discrimination. The IIC of an item of relatively high discrimination, for example, will have a high peak and relatively narrow spread. In contrast, an item of low discrimination will have a comparatively low peak and will spread over a wider range of the trait continuum.

An important property of IICs is that they can be summed to produce a test information curve (TIC), a function that describes the range on the trait continuum where measurement is most and least precise for a given test. In contrast to classical test theory (cf. Gulliksen, 1950), IRT allows for the calculation of differential standard errors of measurement (i.e., measurement precision) across the trait continuum (Embretson, 1996; Embretson & Hershberger, 1999). In IRT, the standard error of measurement is equal to the inverse square root of test information. Thus, knowing the TIC for a given test permits one to compute differential standard errors of measurement for any possible trait level. Given these features, CAT applications can utilize item and test information to

select only those items for administration that provide maximum psychometric information throughout a given test administration (Weiss, 1985).

Assumptions of IRT

Most popular IRT models rest on two basic assumptions. The first of these, the assumption of *scale unidimensionality*, requires that items from scales cohere and measure a common latent trait. In other words, all items in a given item pool must load on a single factor. As discussed above, this requirement has precluded the use of IRT for the analysis of relatively heterogeneous scales such as those from the MMPI. Multidimensional IRT models have recently been developed and utilized in the ability testing domain (e.g., Embretson, 1991; Knol & Berger, 1991), but these methods have not yet been extended to multidimensional personality scales. Despite these recent advances, most available IRT software is based on models that assume scale unidimensionality; thus, personality measures that have been designed to measure relatively homogeneous latent traits (i.e., those developed using factor analytic methods) are perfect candidates for IRT calibration.

The second basic assumption of IRT, referred to as the assumption of *local independence*, requires that a given examinee's item responses are statistically independent once his or her trait level has been taken into account. Stated less formally, this assumption states that an individual's performance on one item of a scale does not influence his or her response to any other item on the scale. For example, if one item of a hypothetical reading test contained information that would aid an examinee on a later test item, the assumption of local independence would be violated, and IRT models should not be used. Personality researchers wishing to use IRT models have not been very

concerned about this assumption, believing that responses to items on personality tests are largely independent of one another. This may be an error, however, as small item context and serial position effects have been reported in the personality literature (Knowles, 1988; Steinberg, 1994). Further research is needed to clarify whether these effects represent a significant violation of IRT's assumption of local independence.

IRT in the Personality Literature

The use of IRT as a tool to aid personality researchers in the investigation of practical personality measurement questions has become increasingly popular. Some researchers, for example, have utilized IRT models to help clarify the interpretation of the structural properties a given measure. In one such analysis of Hare's Psychopathy Checklist—Revised (PCL-R; Hare, 1991), Cooke and Michie (1997) utilized Samejima's graded item response model (Samejima, 1969) to analyze both the test and item functioning of the PCL-R in a large sample of 2,067 participants. Previous work has found that two correlated but distinct factors underlie scores on the PCL-R: (a) *selfish, callous, and remorseless use of others*, and (b) *chronically unstable and antisocial life.* Cooke and Michie found that items related to the first factor were generally more discriminating (i.e., higher *a* values) and provided more information about the psychopathy trait than items relating to the second factor.

In a later study, Cooke and colleagues (Cooke, Michie, Hart, & Hare, 1999) again used the graded IRT model to determine whether the screening version of the Psychopathy Checklist—Revised (PCL:SV; Hart, Cox, & Hare, 1995) could be regarded as a short form of the longer PCL-R. In doing so, the authors reported that 8 of 12 PCL:SV items were strongly parallel (i.e., nearly identical *a* and *b* values) to their

equivalent PCL-R items, and that the 4 remaining items had equivalent or superior discrimination properties (i.e., *a* values) compared to their equivalent PCL-R items. Based on these IRT analyses, the authors were able to conclude that the PCL:SV has structural properties very similar to those of the PCL-R and that the PCL:SV can be considered a short or parallel form of the PCL-R.

IRT has also been used to create, modify, or shorten personality scales. To that end, Waller, Tellegen, McDonald, and Lykken (1996) recently utilized nonlinear factor analysis and IRT to construct a 30-item Negative Emotionality scale that reflects individual differences on Tellegen's (1982) higher order Negative Emotionality factor. Waller et al. drew their initial item pool from five lower order scales (Stress Reaction, Alienation, Aggression, Absorption, and Social Closeness) of the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982) that have loaded consistently on this factor in previous investigations. This initial pool of 122 items was first reduced to 62 by an examination of each item's convergent and discriminant relations with each of the MPQ's three higher order factors. The two-parameter logistic IRT model (Birnbaum, 1968) was then used to reduce this number to 30 by identifying highly discriminating items (i.e., high *a* values) that evenly represented each of the five content areas and that spanned a range of item thresholds (i.e., *b* values). The resulting 30-item scale correlated equivalently with the scales of the CPI (Gough, 1975) compared to factor scores derived from four times as many items.

In another example, Kim & Pilkonis (1999) used Samejima's (1969) graded item response model to construct shorter versions of five scales (Pilkonis, Kim, Proietti, & Barkham, 1996) relevant to the measurement of personality disorder derived from the

Inventory of Interpersonal Problems (IIP; Horowitz, Rosenberg, Baer, Ureño, & Villaseñor, 1988). The original scales totaled 47 items; Kim and Pilkonis sought to make them more attractive for screening purposes by shortening them to five items each. To do this, the authors identified the five most informative items from each scale, based on the magnitude of the item discrimination values as well as the elevation and range of individual item information functions. The shortened scales demonstrated satisfactory levels of internal consistency (all alphas greater than .80) as well as convergent and predictive validity, supporting the ability of IRT to help create shorter scales with little or no loss of reliability or validity.

A final and increasingly popular application of IRT involves the assessment of differential item functioning (DIF; Holland & Wainer, 1993) across samples or groups of individuals. DIF occurs when the probability of endorsing a given questionnaire item depends on both an individual's trait level as well as other factors such as the individual's group membership (e.g., gender, racial, or age cohort groups). Largely because of the attention focused on cultural bias on achievement and ability tests, numerous IRT-based methods have been developed to assess DIF. In IRT terms, DIF is defined as occurring when "individuals having the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, et al., 1991, p. 110). While DIF analyses were originally developed for and utilized in ability testing applications, these methods have begun to appear more often in the personality testing literature to identify DIF related to gender (e.g., Smith & Reise, 1998), age cohort (e.g., Mackinnon, et al., 1995), and culture (e.g., Huang, Church, & Katigbak, 1997).

Smith & Reise (1998) studied DIF related to gender on the Stress Reaction Scale of the MPQ (Tellegen, 1982) and found that, after accounting for mean differences on the scale, women were more likely to endorse items describing emotional vulnerability, and men were more likely to endorse items describing tension, irritability, and being easily upset. Mackinnon et al. (1995) looked at DIF associated with age cohort (elderly vs. younger samples) on the Eysenck Personality Questionnaire (EPQ; Eysenck & Eysenck, 1975). While finding some evidence of DIF, they concluded that the measurement properties of the EPQ were largely equivalent in older and younger samples. Finally, Huang et al. (1997) examined DIF related to cultural differences (American vs. Filipino college students) on the NEO Personality Inventory (NEO PI; Costa & McCrae, 1992). The authors reported that nearly 40% of the items showed DIF and that mean scale score differences between groups disappeared when purified scales were constructed that removed these items.

IRT-based CAT

Clearly, as the examples in the preceding section illustrate, IRT has been used increasingly within the personality domain to examine important testing questions. As described briefly above, another domain to which IRT has been applied is computerized adaptive testing. For tests that meet its assumption of scale unidimensionality, IRT can afford CATs greater precision and efficiency (Weiss, 1985) than the non-IRT CAT methods discussed previously (Forbey et al., 2000; Ben-Porath et al., 1989; Handel et al., 1999; Roper et al., 1991, 1995). As discussed above, items calibrated using IRT can be characterized in terms of the amount of information they provide along the trait continuum; in CAT, then, items can be selected for administration that are maximally

informative at the examinee's current trait level estimate. This precision leads to overall test efficiency. By selecting items that are maximally informative for the examinee at each point of the test, the computer typically can obtain a reliable estimate of the trait with far fewer items than traditional paper-and-pencil tests (Weiss, 1985).

Within the ability testing domain, IRT-based CATs have been developed across a range of abilities and settings. One of the earliest examples is the CAT version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB), which has was developed through more than 20 years of research and has been used to assess military recruits since 1990 (Sands, Waters, & McBride, 1997; Segall & Moreno, 1999). Both simulation and live testing data support the reliability and validity of the CAT-ASVAB (e.g., McBride & Martin, 1983; Moreno & Segall, 1997), and these studies have indicated that adaptive tests can reduce test lengths by more than 50% when compared to conventional tests, without any loss of reliability or validity.

Another prominent example of CAT in the ability testing community is the Graduate Record Examination (GRE-CAT; Educational Testing Service, 2002). Development began on the GRE-CAT in 1988, some examinees were first assessed with it in 1992, and today all GRE examinees taking the GRE adaptively by computer (Mills, 1999; Schaeffer et al., 1998). The GRE-CAT uses a fixed-length adaptive algorithm, yielding item savings of 61%, 53%, and 30% on the verbal, quantitative, and analytical subtests, respectively (Schaeffer, Reese, Steffen, McKinley, & Mills, 1993). In addition, cross-mode correlations between the GRE-CAT and a computer-based full-scale version approach the scale internal consistency reliabilities, suggesting good recovery of score rank-ordering (Schaeffer, Steffen, Golub-Smith, Mills, & Durso (1995). However, a

recent report (Schaeffer et al., 1998) revealed that scores on all three subtests were significantly higher when adaptively administered, but the authors attributed these differences to the manner of scoring for examinees who failed to respond to all items and suggested a new proportional scoring algorithm that successfully eliminated these differences.

IRT, CAT, and Personality

Relatively few applications of IRT-based CAT appear in the personality literature, and these few examples are all computerized simulations. In the first such study, Waller and Reise (1989) reported data supporting the feasibility of IRT-based computerized adaptive personality testing. Applying the two-parameter logistic model (Birnbaum, 1968) to the Absorption scale of the MPQ, Waller and Reise conducted computerized real-data simulations. They examined two adaptive testing strategies. In the first of these, *fixed-test-length adaptive testing*, the computer administers a fixed number of items to each examinee. Alternatively, when the goal of assessment is only classification, *clinical-decision adaptive testing* can be used to reduce testing time without sacrificing diagnostic accuracy. In clinical-decision adaptive testing, items are administered until the confidence interval surrounding the current point estimate of the trait score no longer includes the cutoff value used to classify the subjects. Real-data simulations, based on responses from 1000 subjects who had previously completed the Absorption scale in the traditional paper-and-pencil format, were used to illustrate these strategies. The results suggested that computerized adaptive personality assessment can work quite well. With the fixed-test-length strategy, Waller and Reise achieved a 50% savings in administered items with little loss of measurement precision. Using the clinical-decision testing

strategy, individuals who were extreme on the Absorption trait were identified with perfect accuracy using, on average, only 25% of the available items.

Shortly after Waller and Reise (1989), in a similar demonstration, Kamakura and Balasubramanian (1989) applied IRT to the 54-item Socialization scale of the CPI. Based on these analyses, they eliminated 10 items due to poor discrimination values, yielding a final item pool of 44 items, and then constructed a computerized simulation of three adaptive administration strategies using real data. In the first strategy (MAXIT), the test proceeded until a maximum of 15 items were asked of the examinees. In the second strategy (MINSTD), testing continued until an arbitrarily specified minimum standard error (in this case, 0.40) was achieved for each examinee's trait estimate. The last adaptive strategy (MIXED) combined the first two: A minimum of 10 items were administered to each examinee, and testing continued until a minimum standard error of 0.40 was attained. These strategies resulted in significant item savings. The computer algorithm administered only 34.1%, 38.6%, and 39.8% of the items for the MAXIT, MINSTD, and MIXED adaptive strategies, respectively. However, in this simulation, the items savings were obtained at the expense of measurement error. The authors reported small but significant losses in measurement precision for scores estimated from adaptive testing, but they concluded that these small increases in error were overshadowed by the substantial item and time savings that were realized.

In a more recent study, Reise and Henson (2000) conducted similar real-data simulations on the facet scales of the NEO Personality Inventory-Revised (NEO PI-R; Costa & McCrae, 1992). Using a maximum information item selection adaptive testing algorithm, they were able to achieve item savings of 50%, on average, across the 30 facet

scales, with little loss in measurement precision. Interestingly, however, their results also suggested that the CAT design was unnecessary to obtain these result. Instead, they found that using the four most informative items from most facet scales yielded equally impressive results.

Finally, Waller (1999) completed a similar analysis of one scale of the MMPI. As described above, the multidimensionality of the clinical scales of the MMPI has precluded the successful application of IRT to its scales in the past, and CAT versions of the MMPI have been based on methods that do not require this assumption. To circumvent this problem, Waller (1999) factor analyzed the MMPI item responses of a large sample of 28,390 medical and psychiatric patients who were treated at the University of Minnesota Hospital between 1940 and 1976. He identified 16 unidimensional factors within the MMPI item pool and chose one such factor, "Denial of Somatic Complaints," for an IRT-based CAT simulation similar to the one described above (Waller & Reise, 1989). The simulation involved a cycle of theta estimation, maximum-information item selection, and item scoring that continued until two termination criteria were satisfied. First, each subject was administered at least 20 items. Second, items were administered until the conditional item information in the next item fell below a threshold of .10. The results were impressive. Over half of the sample finished the test after only 20 items (from a total of 51 items), effectively shortening the scale by 61%, and no person answered more than 44 items. However, Waller did not report any extra-test correlates, making convergent and discriminant validity comparisons impossible. Further, while this simulation showed that IRT-based CATs are possible with the MMPI, it did so using a factor analytically derived scale that is not scored in

traditional settings. Thus, unless it is revised and restructured along factor analytic lines, the MMPI cannot be considered a practical candidate for IRT-based CAT in the foreseeable future.

Measure Selection

The first step in the development of a multi-scale personality CAT is the selection of an instrument from which to draw the items and scales. The primary requirement is that each scale of the measure must contain items that represent a single latent factor. To that end, measures that have been developed using factor analytic techniques have a distinct advantage over measures that were constructed via empirical criterion-keying (e.g., the MMPI) or by rational methods (e.g., the Tennessee Self-Concept Scale; Fitts, 1965). One such instrument is the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993). The SNAP is a self-report questionnaire that measures 15 relatively distinct personality trait dimensions relevant to personality disorder, which is a type of psychopathology involving pervasive dysfunction in interpersonal behavior, emotional dysregulation, and behavioral dyscontrol (American Psychiatric Association [APA], 1994). Personality pathology is highly prevalent (e.g., Ekselius, Tillfors, Furmark, & Fredrikson, 2001). Moreover, the presence of personality pathology has been associated with a variety of negative outcomes, including poorer psychosocial functioning in community samples (e.g., Soldz & Vaillant, 1999) and poorer prognosis for individuals with various other conditions (e.g., Mennin, & Heimberg, 2000). Thus, assessment of personality pathology is an important clinical activity.

The 375-item SNAP questionnaire, which was developed using a combination of content analytic and factor analytic methods, yields scores on the following core traits of

personality disorder: Negative Temperament, Mistrust, Manipulativeness, Aggression, Self-harm, Eccentric Perceptions, Dependency, Positive Temperament, Exhibitionism, Entitlement, Detachment, Disinhibition, Impulsivity, Propriety, and Workaholism. The scales generally group into three broad higher order factors: Negative Affectivity, Positive Affectivity, and Disinhibition vs. Contraint. In addition, the SNAP includes five validity scales—Rare Virtues, Deviance, Variable Response Inconsistency (VRIN), True Response Inconsistency (TRIN), and Desirable Response Inconsistency (DRIN)—and an overall Invalidity Index, that were designed to detect several common response patterns (e.g., random, defensive, or exaggerated responding). These scales, given their design and psychometric properties, are not likely amenable to IRT or CAT. However, because of their importance in both clinical and research settings, they will be included in the computerized adaptive version of the SNAP (SNAP-CAT). Luckily, for the sake of simplicity, the items for all of the validity scales, except Rare Virtues, are shared with the standard trait and temperament scales. Finally, the SNAP includes a set of diagnostic scales that provide indices of personality pathology linked to the official Diagnostic and Statistical Manual of the American Psychiatric Association (DSM-IV; APA, 1987). Because of complexities associated with inconsistent item overlap, the diagnostic scales were not included in the SNAP-CAT at this stage of the project.

A major limitation of the SNAP is that it is long and generally takes over one hour to complete. Scoring the SNAP by hand is time consuming as well, and the available computerized scoring program, while convenient, requires some time for a clinician or technician to set up, enter the data, and score the protocol. In clinical settings, managed care companies have begun to control tightly the types of assessments for which they will

reimburse psychologists, with an eye toward shorter, more efficient measures. Shorter measures are also desirable in research settings, as participant and researcher time is valuable and often in short supply. Thus, given its factor analytic foundation and a desire to create shorter, more time-efficient measures, the SNAP was an excellent candidate for the sort of IRT-based CAT that was the primary objective of this project.

Summary

The CAT simulations described above have built an impressive foundation upon which to begin exploring the feasibility of developing truly adaptive personality measures. However, no studies exist in the personality literature in which live participants complete IRT-based CATs. Such studies are sorely needed. The equivalence of simulated and live testing conditions cannot be assumed, as experiential aspects of CAT may lead to reliability and validity results that are different from those obtained during CAT simulation studies. Moreover, the equivalency of CATs to their traditional paper-and-pencil counterparts must also be established empirically. Computerized adaptive tests may administer items from the same item pool as conventional paper-and-pencil tests, but the mode of stimulus presentation is quite different. The order of item presentation varies considerably across the two modes, and instead of a booklet, an answer sheet, and a pencil, the examinee is presented with words on a computer screen, a keyboard, and a mouse. These differences may lead to differential responding, and thus presentation mode equivalence cannot be assumed.

Thus, the primary objectives of this dissertation project were to (a) develop an IRT-based CAT version of the SNAP, an existing multi-scale personality inventory that appears in the literature and is commercially available, (b) use live participants to provide

equivalence, reliability, and validity data that are more ecologically valid than simulation data, and (c) attempt to replicate the item and time savings that have been obtained previously using non-IRT methodology (e.g., Ben-Porath et al., 1989; Handel et al., 1999; Roper et al., 1991, 1995) as well as IRT-based CAT simulations (Kamakura & Balasubramanian, 1989; Reise & Henson, 2000; Waller, 1999; Waller & Reise, 1989).

The development and validation of the computerized adaptive version of the SNAP-CAT was conducted in five phases: (a) IRT-based item calibration of all SNAP trait scales, (b) computerized CAT simulations of the SNAP to assess various termination rules, (c) design of the SNAP-CAT program based on the results from the simulation study, (d) validation of the SNAP-CAT using live research participants, and (e) conducting analyses to assess the stability, validity, and equivalence of the SNAP-CAT compared to the paper-and-pencil version of the SNAP.

Figure 1.1: Parameters of a Typical Item Characteristic Curve (ICC).

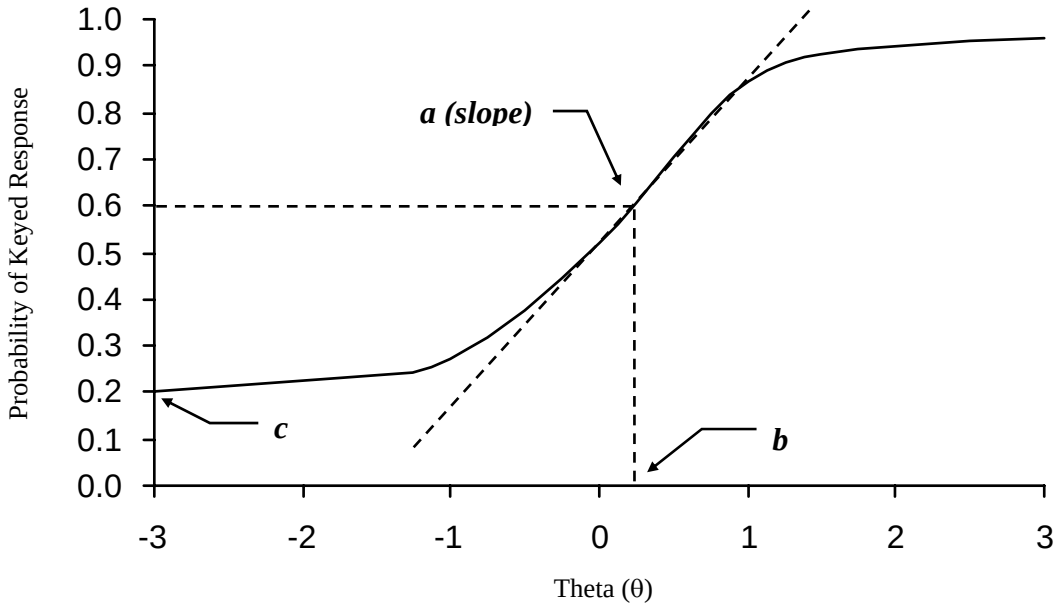
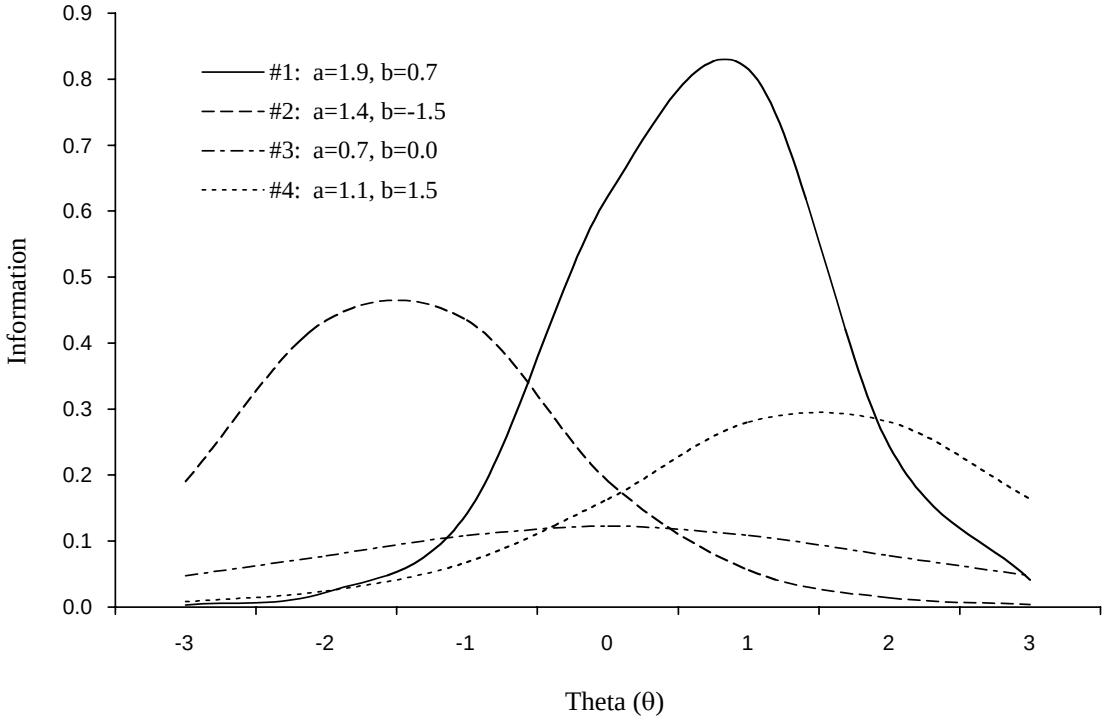


Figure 1.2: Several Typical Item Information Curves (IIC).



CHAPTER 2
DEVELOPMENT OF SNAP-CAT

Item Calibration

Item parameters were calibrated on a large sample of 3,995 individuals (40.8% male and 59.2% female; 86.3% white, 6.1% African American, and 7.6% other ethnic origins) who completed the standard paper-and-pencil version of the SNAP in a variety of university, community, and patient settings over the past decade. Demographic characteristics of this large calibration sample are presented in Table 2.1, and the sources that formed this sample are described more fully in Appendix A. To summarize, the calibration sample was comprised of 809 community-dwelling adults (43.0% male and 57.0% female), 1,886 college undergraduates (37.0% male and 63.0% female), and 1300 psychiatric patients (44.9% male and 55.2% female). The size of this calibration sample was more than adequate for an IRT calibration of this sort (Reise, 1999; Zickar, 2001).

IRT item parameters were estimated for each of the SNAP's 15 trait and temperament scales. Complicating this process was item overlap across several SNAP scales. Most notably, 19 of the Disinhibition's 35 items are also scored on other scales (e.g., Impulsivity, Manipulativeness, and Propriety). Based on feedback received from the prospectus committee, I decided to ignore item overlap associated with Disinhibition and to include full items for each overlapping scale. This decision maintained the integrity of these scales and simplified matters considerably during the calibration phase, but resulted in item presentation duplicates in the live testing phase. Another area of item overlap involved one item (item 211 on the traditional SNAP: "People sometimes tell me to slow down and 'take it easy'"), which is traditionally scored on both Positive

Temperament and Workaholism. In this case, preliminary correlational analyses were conducted using the calibration sample to determine the scale to which the item best correlated; results indicated that item 211 correlated more strongly with Workaholism (r = .53) than with Positive Temperament (r = .20). Thus, item 211 was assigned to the Workaholism scale for all subsequent analyses related to SNAP-CAT construction.

Unidimensionality Assumption

Prior to IRT parameter estimation, multiple methods were used in the calibration sample to assess scale unidimensionality. First, the internal consistencies of each scale were assessed. Cronbach's alpha coefficients and average inter-item correlations (AICs) are listed in Table 2.2. These were generally quite good (median alpha = .84, range = .77 to .92, and median AIC = .22, range = .15 to .34) and suggested that the items within each scale cohere well. Second, traditional item-level principal components analyses were conducted, and the ratios of the first to the second eigenvalues were examined for each scale. Although consistent interpretive standards do not exist for this index, higher ratios generally suggest a greater degree of scale unidimensionality. The median ratio was 4.7 (range = 3.2 to 8.9), suggesting that the first unrotated factor of the average SNAP scale accounts for almost five times the variance of the second factor. Such data point further to the unidimensionality of these scales.

Unfortunately, traditional linear factor analysis (i.e., the type implemented by most statistical programs) can result in spurious factors when applied to dichotomous data (Waller et al., 1996). Thus, a number of factor analytic methods have been proposed to deal with the non-linear item-trait regressions that can result from dichotomously scored items (Mislevy, 1986). Some of these methods rely on factor analyses of

tetrachoric correlation matrices rather than standard Pearson product-moment correlations (e.g., Waller, 1995, 2002). Tetrachoric correlations are estimates of the correlations that would have been observed had the variables been measured on a continuous metric, as opposed to being measured dichotomously. Other methods exist that are more computationally demanding (e.g., Wilson, Wood, & Gibbons), but a number of studies (Knol & Berger, 1991; Waller et al., 1996) have suggested that the simpler methods (i.e., those based on factor analyses of tetrachoric correlations) produce factor loadings that are highly similar to those produced by the more complex methods. Thus, in the present study, scale unidimensionality was assessed using MicroFACT (Waller, 1995, 2002), which is a software program which employs the simpler methods described above. For each scale, MicroFACT was programmed to fit a one-factor model to a matrix of tetrachoric item correlations. Table 2.2 includes two indices of model fit: the Root Mean Squared Residual (RMSR) and the Goodness of Fit Index (GFI). The mean RMSR value was .086 (range = .058 to .122), with lower values indicative of better model-to-data fit. The mean GFI value was .963 (range = .942 to .986), and all but two scales were in the excellent range (i.e., above .950; Hu & Bentler, 1999). The remaining two values, .942 (Entitlement) and .943 (Dependency), were (a) clearly in the adequate range (i.e., between .900 and .950; Hu & Bentler, 1999), and (b) nearly in the excellent range. Taken together, the ratios of eigenvalues, internal consistency estimates, and non-linear factor analyses support the unidimensionality of the SNAP scales and thus provide evidence of their appropriateness for IRT applications.

Item Parameter Estimation

IRT item parameters were estimated from calibration sample item responses using BILOG 3.1 (using program defaults, Mislevy & Bock, 1990), a computer program that implements the marginal maximum likelihood method (MML; Bock & Aiken, 1981) of item parameter estimation. The MML method of estimation is an efficient and effective method for IRT item calibration (Bock & Aiken, 1981; Mislevy & Bock, 1990). BILOG can compute parameter estimates for the one-, two-, or three-parameter models. Given the considerations described above regarding personality data, the two-parameter logistic IRT model was used for all item parameter estimation in this project. Item parameters for all SNAP scales are listed in Tables B1 through B15 in Appendix B. In addition, to examine item-level goodness-of-fit, these tables include *root mean square standardized posterior residuals (RMSSPR)* for each item. Standardized posterior residuals are item-level differences between the *actual* probability of item endorsement (based on real data) at selected values along theta and the *predicted* probability of item endorsement calculated from item characteristic curves (ICCs) at the same levels of theta. RMSSPR is an index that summarizes standardized posterior residuals across multiple points of the theta continuum for a given item. BILOG also computes Chi-square tests of item fit, but because Chi-square tests are generally undependable for shorter scales such as those in this application, RMSSPR values are considered preferable (Mislevy & Bock, 1990). Firm interpretive cut-off values do not exist for RMSSPR; however, the BILOG manual (Mislevy & Bock, 1990) suggests that (a) smaller values of RMSSPR are associated with better item fit, and (b) values greater than 2.0 suggest some failure of item fit.

Results revealed significant scale-level variability for RMSSPR, with the mean

RMSSPR value exceeding 2.0 on 5 of 15 scales. Impulsivity, for instance, yielded the fewest problematic items (10.5%) according to this criterion, whereas 87.5% of Self-harm items were above the problematic threshold. Despite the surprisingly large number of poorly fitting items on some scales, all items, regardless of item fit, were included in the SNAP-CAT in order to maintain consistency with the traditional P&P version of the SNAP. To illustrate the concept of item-fit, ICCs for two poorly fitting items (Aggression item 20 [AGG020, "I would go out of my way to avoid a fight."]: RMSSPR = 4.382; and Self-harm item 15 [SHM015, "I have never given any thought to killing myself."]: RMSSPR = 8.416) and two well-fitting items (Manipulativeness item 9 [MAN009, "I am quite willing to bend the truth if it will benefit me."]: RMSSPR = 0.601; and Exhibitionism item 13 [EXH013, "I like to turn heads when I walk into a room."]: RMSSPR = 0.792) are presented in Figure 2.1. The difference between poorly- and well-fitting items generally can be attributed to one or two marked differences between the observed data and the fitted ICC at some point along the theta continuum. For SHM015, for example, it is clear that the ICC fits the observed data reasonably well at all but the lowest two points, whereas the ICC for MAN009 provided relatively good fit at all theta points.

Measurement Precision

Another way to assess the applicability of the SNAP for IRT applications is to examine the measurement precision associated with each scale as a function of theta. As described above, test information curves (TICs) are functions that describe the range along the trait continuum where measurement is most and least precise for a given test. In addition, the standard error of measurement (SEM) in IRT applications can be

calculated differentially along the trait continuum and is equal to the inverse square root of test information. Thus, knowing the TIC for a given test makes it possible to compute differential standard errors of measurement for any possible trait level. TICs and SEM curves for each SNAP scale, based on the calibration sample, are presented in Figure 2.2. In each graph, the solid line represents test information, and the dashed line is the SEM. Most curves (i.e., Aggression, Dependency, Detachment, Disinhibition, Eccentric Perceptions, Impulsivity, Manipulativeness, Mistrust, and Workaholism) peaked between theta = +0.5 and theta = +1.0, which is appropriate given that (a) the SNAP was designed to measure people in the abnormal range, and (b) most items on these scales are keyed in the pathological direction. Two scales (i.e., Positive Temperament and Propriety) yielded TIC peaks between theta = -0.5 and theta = -1.0, which is understandable for Positive Temperament given that low scores on this scale are considered pathological. Propriety, however, was not designed to have a pathological pole, but these data suggest that the SNAP provides more information at the low end of this trait. The remaining three scales (i.e., Entitlement, Exhibitionism, and Negative Temperament) peaked at or near zero, suggesting that the item content of these scales is more neutrally representative of the broad trait continuum. Taken together, these curves reveal that SNAP scales generally provide the most psychometric information (i.e., yield the lowest standard errors of measurement) for examinees who score in the pathological direction of each trait dimension.

Computerized Simulation Study

A key decision to be considered in any CAT application is when to terminate the test. Should the test terminate when a pre-specified number of items has been

administered, when the SEM of the trait estimate falls below a pre-specified limit, when reasonably informative items no longer exist for a given examinee, or when some combination of these rules has been satisfied? To aid in the determination of a reasonable termination rule for the SNAP-CAT, computerized CAT simulations were conducted using simulated response data. Using simulated response data offers several advantages over real data for computerized CAT simulations (W. P. Vispoel, personal communication, May 9, 2000). First, using simulated data permits comparisons between "true" trait levels (i.e., the trait levels from which the simulated data are derived) with estimated trait levels (i.e., the trait levels estimated by the computerized CAT simulation software), whereas a simulation using real response data is restricted to trait level estimates which may not mirror true trait levels. Second, simulated data allows measurement precision estimates (such as test information or standard error of measurement) to be keyed to true trait levels, whereas real-data simulations produce precision estimates that are referenced to estimated rather than true trait levels. Finally, using simulated data permits the researcher to create an large number of replicates at each pre-specified trait level, whereas real-data simulations often provide very few cases at the high and low ends of the trait dimension.

Computerized simulations were conducted on all SNAP scales and involved three steps: (a) creation of the simulated response data, (b) CAT simulations, and (c) post-hoc analyses to help make informed decisions regarding the termination rules for each scale. For each scale, data were simulated for seven pre-specified "true" trait levels: theta = -3, -2, -1, 0, 1, 2, and 3. To do this, the ICC for each item was used to find the probability of a keyed response given the true theta. To determine item endorsement, this probability

value was then compared to a number sampled from a uniform distribution (i.e., 0 to 1); if the sampled number was less than or equal to the probability level, the item was answered in the keyed direction; if the sampled number was greater than the probability level, the item was answered in the non-keyed direction. This process was repeated for all items in each scale and for each of the seven thetas listed above. One hundred replicates were produced at each theta level, yielding a total sample size of 700 simulated examinees.

The simulated data sets were then subjected to computerized CAT simulations using POSTSIM (Assessment Systems Corporation, 1999), a simple off-the-shelf computerized simulation program that can simulate several common adaptive testing algorithms. For each SNAP scale, POSTSIM was instructed to simulate tests at all test lengths (i.e., ranging from one item to the maximum number of items on each scale). These simulations produced data files that included detailed response patterns, estimated thetas, and precision estimates (i.e., cumulative test information) for each examinee at all tests lengths. These data were exported to SAS (SAS Institute, 1990) for post-hoc analyses to examine test fidelity (i.e., correlations between shortened- and full-scale thetas) across test lengths, key the precision estimates to "true theta," and create data files for the final simulation step. Figure 2.3 shows test fidelity plotted as a function of test length and indicates that all SNAP scales yielded uniformly high short-scale to true-theta correlations of greater than .90 after only six items were administered. These data provided confidence that adaptively shortened versions of SNAP scales could yield highly accurate approximations of full-scale scores.

Response data exported from SAS were then manipulated in a spreadsheet to

examine various termination algorithms. First, to ensure that trait levels were assessed with a minimum level of precision, a minimum number of items to be administered was established for each scale. To do this, TICs were produced and plotted at every test length for each scale. After experimenting with various criteria, item minimums were chosen that yielded an information loss of approximately 33% at the peak of the TIC. Using this criterion yielded mean item savings of 56.2% (range = 50.0% to 63.2%) across scales, which was nearly double the mean loss of information of 31.6% (range = 26.4% to 35.7%). Resultant minimums are presented in Table 2.3. Consistent with the test fidelity analyses reported above, adaptive scores based on these minimums were highly correlated (mean r = .93; range = .92 to .96) with full-scale scores. Thus, these item-presentation minimums served as the first step of the adaptive testing algorithm for the SNAP-CAT.

Next, in order to assess various termination rules, trait SEMs and item information were calculated from cumulative information at every test length. Termination criteria were assessed that induced test termination when (a) the SEM around theta dropped below a given value, (b) the item information afforded by the remaining items dropped below a given value, or (c) some combination of (a) and (b) has been satisfied. Various values for the SEM and item information thresholds were considered; ultimately, values were chosen that yielded a profitable balance between item savings and precision loss. After experimenting with countless variations, a combined termination rule was selected that specified that each test should terminate when either (a) the SEM of the trait estimate drops below 0.40, or (b) items with conditional information estimates greater than 0.10 no longer exist. Based on these criteria, item

savings and information loss for each scale are presented in Table 2.4. Assuming a uniform distribution underlying theta scores (i.e., equivalent numbers of examinees across the theta continuum), this termination rule yielded mean item savings of 42.1% (range = 34.3% to 50.0%) with an associated mean precision loss of 22.4% (range = 16.1% to 33.3%). Assuming a normal distribution underlying theta scores, which is more likely to approximate the actual savings one would gain if real participants were tested, the combined termination rule yielded mean item savings of 36.5% (range = 17.2% to 51.0%) with an associated mean precision loss of 18.1% (range = 8.0% to 28.8%). Notably, Propriety yielded significantly less impressive item savings (17.2%) than all other scales when a normal distribution was assumed (the next lowest scales, Eccentric Perceptions and Entitlement, each yielded 28.8% savings).

SNAP-CAT Construction

The item parameters and termination criteria described above were used to construct the SNAP-CAT with MicroCAT (Assessment Systems Corporation, 1996), which is a DOS-based software package that allows the user to program and administer CATs. Using the Minnesota Computerized Adaptive Test Language (MCATL; Assessment Systems Corporation, 1996), MicroCAT was programmed to administer all 15 SNAP scales adaptively. Unfortunately, MicroCAT was more fickle and quirky than was originally envisioned, and created several programming challenges. Thus, the SNAP-CAT construction procedures described in the following paragraphs reflect the end-result of an arduous process.

Figure 2.4 is a flow chart for the SNAP-CAT. The SNAP-CAT was designed as 15 independent tests—plus two 6-item testlets to measure Rare Virtues—that were linked

together by simple programming steps. The original plan was to program the computer to administer the scales in a random order, but this proved impossible given MicroCAT's memory limitations and the large size of this application. Thus, to vary the presentation order, these scales were grouped into three blocks. The program was instructed to start with an introductory module which described the test instructions and asked examinees for their ID numbers, which included a randomized digit (either 1, 2, or 3) that determined where the test would start. After completing one block of scales, the SNAP-CAT was instructed to go on to the next, in order. After all three blocks were administered, the SNAP-CAT was programmed to display a closing screen that directed examinees to raise their hand to receive the next part of the experiment protocol.

Some theta estimation procedures (e.g., maximum likelihood methods) result in infinite theta estimates for "perfect" (i.e., all-true responding) or zero (i.e., all-false responding) scores. Because such scores are certainly possible with SNAP scales, this problem was ameliorated by using a Bayesian method known as *expected a posteriori* (EAP) estimation (Assessment Systems Corporation, 1996; Mislevy & Bock, 1990) for deriving thetas. However, EAP estimation procedures require that the mean and variance of the population prior distribution be specified in advance. A normal prior distribution was assumed, and the mean and standard deviation of the distribution were set to 0 and 1, respectively. Within each scale, the SNAP-CAT was programmed to administer all items adaptively using a *maximum information* item selection strategy, which means that items were selected for administration that provided the most psychometric information at the examinee's current trait estimate.

The SNAP-CAT was programmed to start each test by administering an item of

median difficulty. Next, theta is estimated, based on the item response, using EAP estimation procedures, and the computer then searches through the remaining scale items for the one that provides the most information at the current trait estimate. That item is then administered, followed by theta estimation and assessment of the termination rule. The cycle of item selection, theta estimation, and termination rule assessment is repeated until the termination rule described above (i.e., administer items beyond the minimum until either the SEM is less than 0.40 or the conditional information associated with any remaining items is less tan 0.10) is satisfied; once met, the adaptive theta estimate, SEM, test time, and the number of items administered are recorded, and the computer then resumes presenting items adaptively until all have been administered. All items were administered so that (a) traditional raw scores could be calculated to provide a validity comparison that controls for computerized administration, and (b) the remaining validity scales (i.e., Deviance, VRIN, TRIN, and DRIN) could be scored at a later date if desired. At the end of each scale, the full-scale theta, SEM, raw score, and test time are recorded, and the test proceeds onto the next scale according to the flow chart in Figure 2.4.

The SNAP-CAT interface was designed to be very simple. An example of a typical item presentation screen appears in Figure 2.5 (the image was inverted for ease of printing). Item presentation screens have black backgrounds and yellow text. Items are centered horizontally in the upper half of the screen. Below the item, the words “TRUE” and “FALSE” are also centered horizontally. In order to minimize rapid-fire or random responding, neither is highlighted when a new item is presented; rather, a solid highlighting bar appears when the examinee presses the space bar or one of the arrow keys. Examinees are instructed to use the space bar or arrow keys to highlight the TRUE

or FALSE options or the blank space below them. The blank space is designed to be the computerized equivalent of leaving an item blank; when examinees select this option, their response is scored in the non-keyed direction (just as an omitted item would not be scored on a traditional paper-and-pencil test). Item omits are tracked and recorded for later decision-making regarding the validity of the protocol. Once the examinee highlights the desired response, the response is recorded, and the next item is selected based on the algorithm described above.

Prior to formal testing, several rounds of informal pilot testing were conducted in order to watch for and fix bugs in the SNAP-CAT administration software. After fixing all problems, live testing began.

Table 2.1: Demographic Characteristics of the Calibration Sample.

		Sex (%)		Ethnicity (%)			Age (years)	
Sample	*n*	M	F	W	B	O	*M*	*SD*
College Samples:								
1	561	38.5	61.5	86.4	5.3	8.3	19.6	3.6
2	378	36.0	64.0	90.1	0.0	9.9	19.7	2.8
3[a]	292	43.5	56.5				19.7	2.2
4	238	37.8	62.2	91.4	4.1	4.5	19.6	5.9
5[a]	197	31.0	69.0				19.6	2.5
6[a]	220	30.9	69.1				19.0	1.4
Community Samples:								
7	561	41.5	58.5	82.6	7.4	10.0	39.3	15.9
8[a]	173	43.9	56.1				31.1	6.4
9	75	52.0	48.0	57.3	29.3	13.4	33.8	10.7
Patient Samples:								
10	108	51.9	48.1	82.4	10.2	7.4	33.0	8.7
11	141	49.7	50.3	97.0	0.0	3.0	32.9	10.5
12	106	23.6	76.4	96.2	1.9	1.9	34.2	10.5
13	125	58.4	41.6	64.8	24.0	11.2	38.6	8.7
14	136	24.3	75.7	89.0	5.1	5.9	41.4	11.1
15	162	20.4	79.6	98.1	1.9	0.0	34.7	11.4
16[a]	522	56.1	43.9				33.1	8.8
Total Sample:	3995	40.8	59.2	86.3	6.1	7.6	28.1	12.0

Note. M = Male, F = Female, W = White, B = Black, O = Other.

[a] Ethnicity data not available.

Table 2.2: Assessment of Unidimensionality in the Calibration Sample.

Scale (items)	α	AIC	Ratio of Eigenvalues	RMSR	GFI
Negative Temperament (28)	.92	.29	8.9	.060	.987
Mistrust (19)	.87	.26	5.9	.069	.980
Manipulativeness (20)	.81	.18	4.4	.071	.971
Aggression (20)	.87	.25	5.1	.086	.972
Self-harm (16)	.89	.34	4.4	.122	.961
Eccentric Perceptions (15)	.81	.22	5.2	.057	.986
Dependency (18)	.82	.20	3.6	.111	.943
Positive Temperament (26)	.88	.22	4.1	.093	.955
Exhibitionism (16)	.83	.23	4.2	.097	.961
Entitlement (16)	.77	.17	3.2	.105	.942
Detachment (18)	.86	.25	5.8	.072	.980
Disinhibition (35)	.86	.15	4.0	.085	.934
Impulsivity (19)	.81	.18	4.1	.078	.963
Propriety (20)	.79	.16	3.7	.071	.965
Workaholism (18)	.82	.20	3.4	.105	.942
Mean	**.84**	**.22**	**4.7**	**.086**	**.963**

Note. α = Cronbach's alpha coefficient, AIC = average inter-item correlation, RMSR = root mean squared residual, GFI = Goodness of Fit Index.

Table 2.3: Summary of Minimum Item Analyses in the Simulation Study.

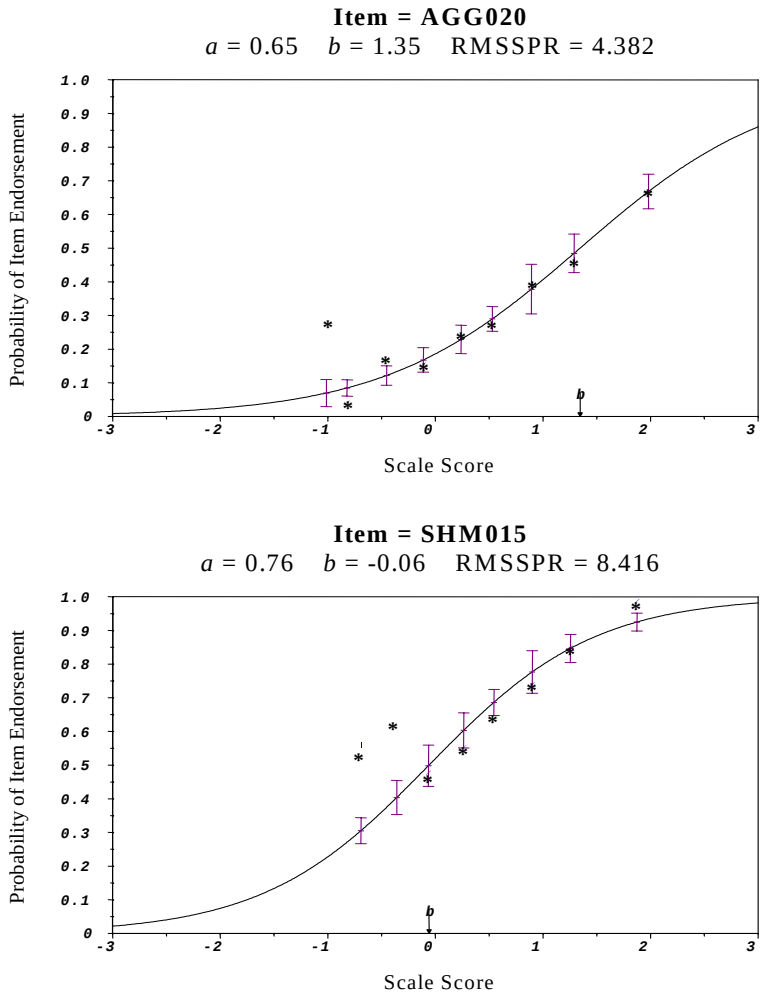
Scale (items)	Minimum	% Item Savings	% Info Loss	Fidelity
Negative Temperament (28)	14	50.0	32.2	.96
Mistrust (19)	8	57.9	35.7	.93
Manipulativeness (20)	9	55.0	35.6	.93
Aggression (20)	10	50.0	33.7	.93
Self-Harm (16)	8	50.0	30.6	.93
Eccentric Perceptions (15)	7	53.3	28.6	.92
Dependency (18)	8	55.6	33.5	.93
Positive Temperament (26)	10	61.5	32.3	.94
Exhibitionism (16)	7	56.3	29.5	.93
Entitlement (16)	7	56.3	26.4	.94
Detachment (18)	7	61.1	30.9	.94
Disinhibition (35)	15	57.1	33.0	.95
Impulsivity (19)	7	63.2	28.6	.92
Propriety (20)	9	55.0	33.7	.93
Workaholism (18)	7	61.1	30.2	.93
Mean		**56.2**	**31.6**	**.93**

Note. Minimum = minimum number of items to be administered, % Info Loss = percentage loss of information at peak of test information curve, Fidelity = correlation between true theta and scores obtained using the minimum test lengths.

Table 2.4: Simulated Item Savings and Loss of Information, Assuming Normal and Uniform Distributions.

	Normal Distribution		Uniform Distribution	
Scale (items)	% Item Savings	% Loss of Info	% Item Savings	% Loss of Info
Negative Temperament (28)	50.0	28.8	50.0	24.0
Mistrust (19)	42.8	26.0	43.6	31.4
Manipulativeness (20)	33.2	14.8	37.9	16.5
Aggression (20)	30.2	17.5	39.3	19.2
Self-Harm (16)	39.7	19.1	38.4	17.3
Eccentric Perceptions (15)	28.8	14.5	37.1	18.3
Dependency (18)	36.0	16.2	40.5	14.9
Positive Temperament (26)	51.0	27.7	48.9	31.2
Exhibitionism (16)	31.1	13.3	37.5	16.1
Entitlement (16)	28.8	6.9	42.0	19.1
Detachment (18)	49.5	24.1	49.2	24.9
Disinhibition (35)	47.2	26.7	48.2	29.5
Impulsivity (19)	31.4	17.1	43.6	33.3
Propriety (20)	17.2	8.0	34.3	18.9
Workaholism (18)	30.6	10.9	40.5	20.9
Mean	**36.5**	**18.1**	**42.1**	**22.4**

Figure 2.1: Item Characteristic Curves and Observed Probabilities (*) for Two Poorly- and Well-fitting Items.



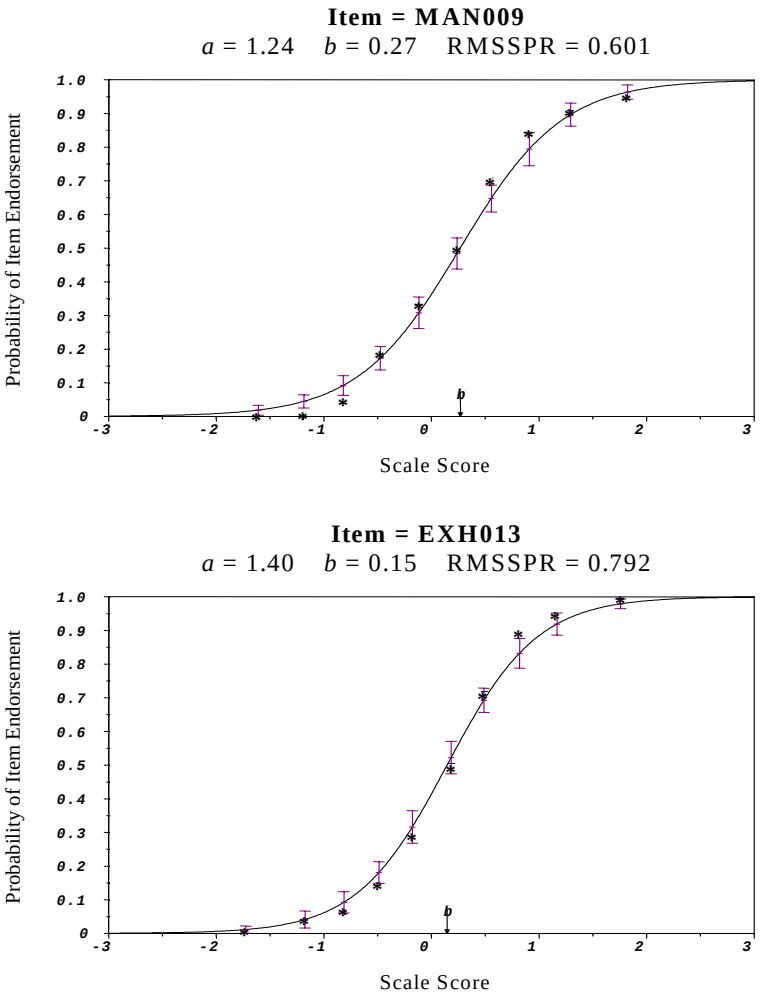


Figure 2.2: Test Information and Measurement Error for SNAP Scales in the Calibration Sample.

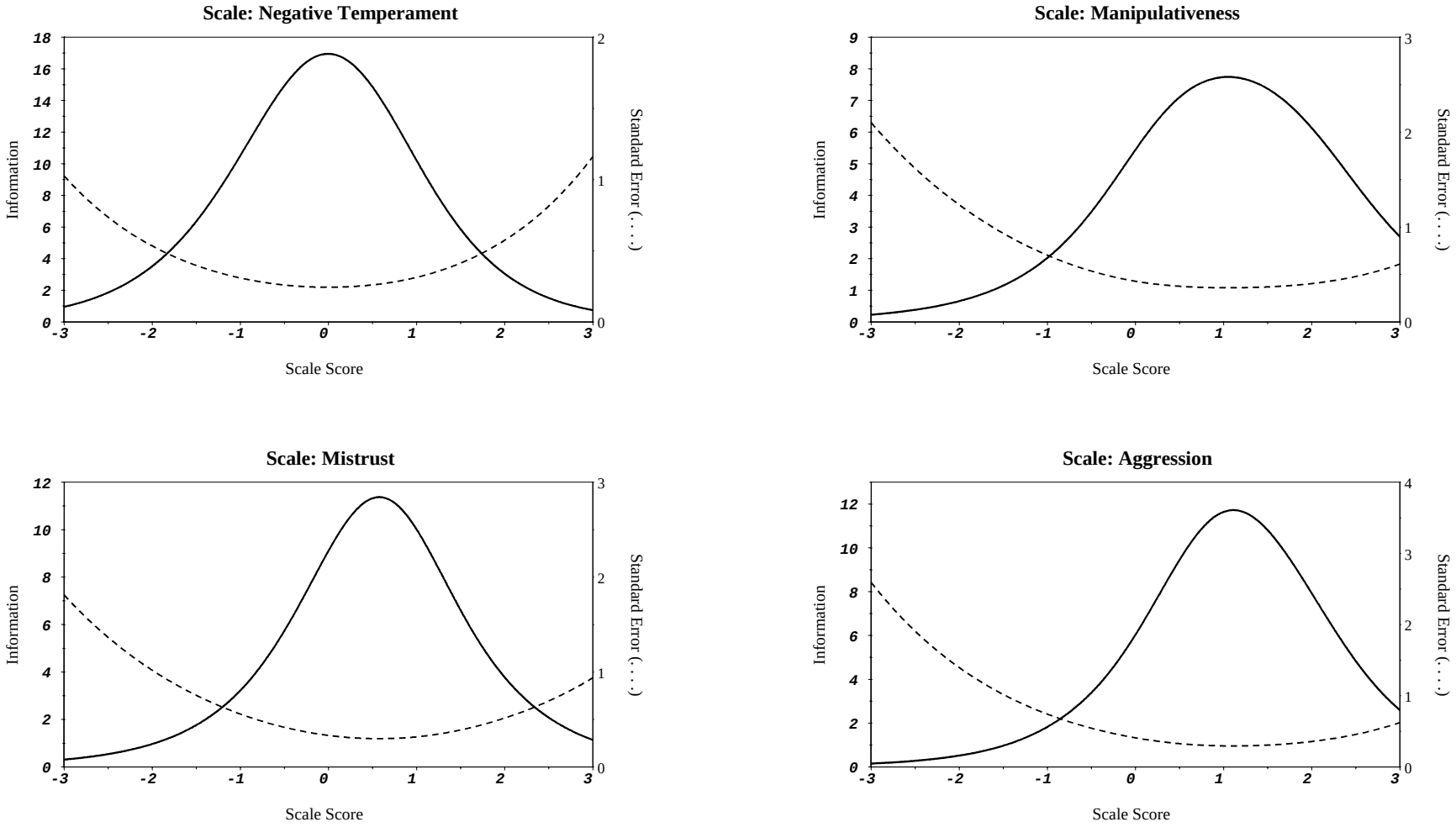
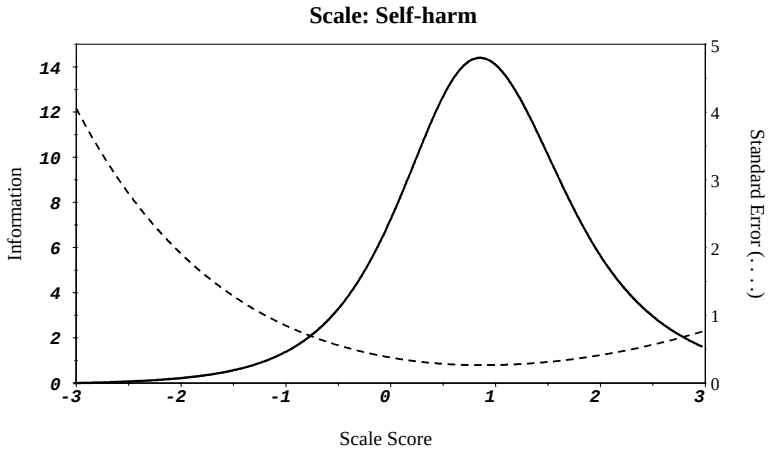
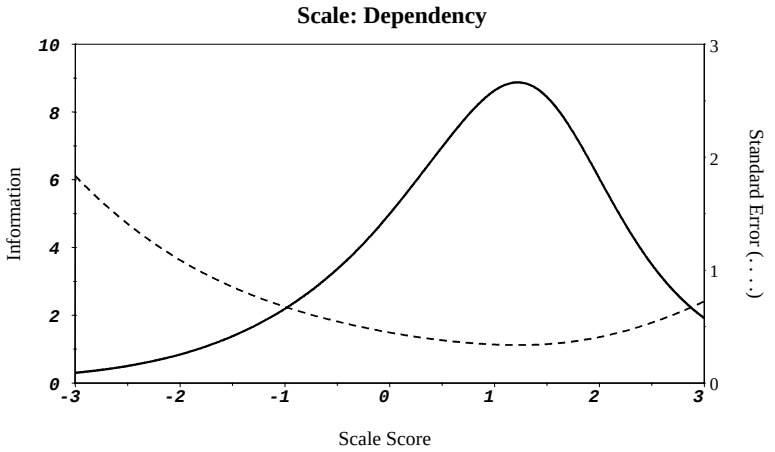
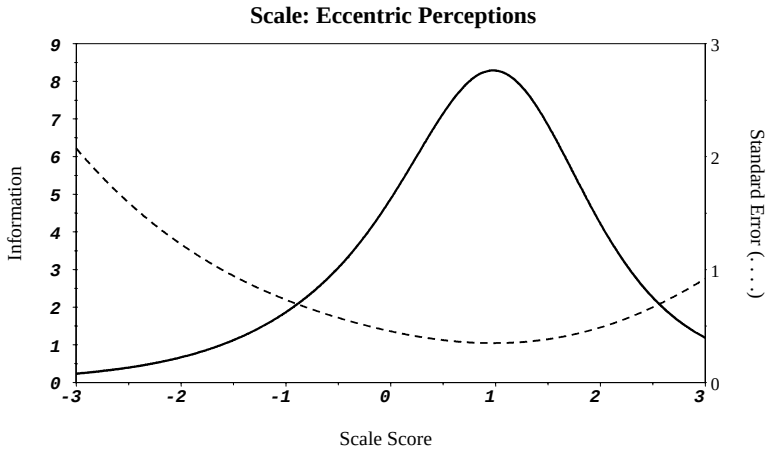


Figure 2.2—continued







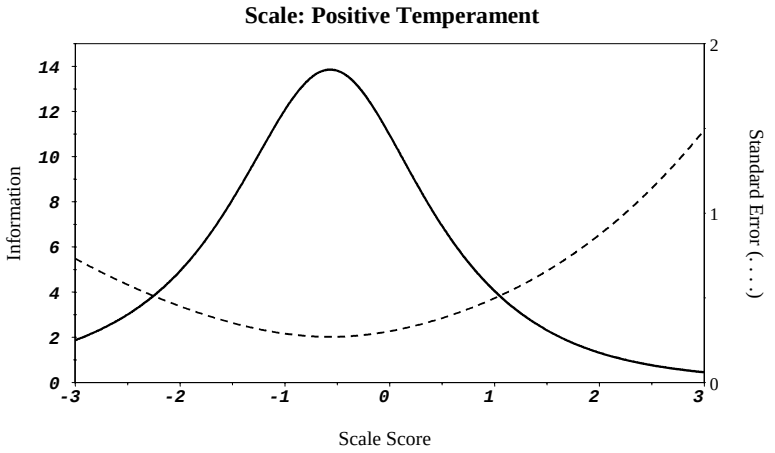
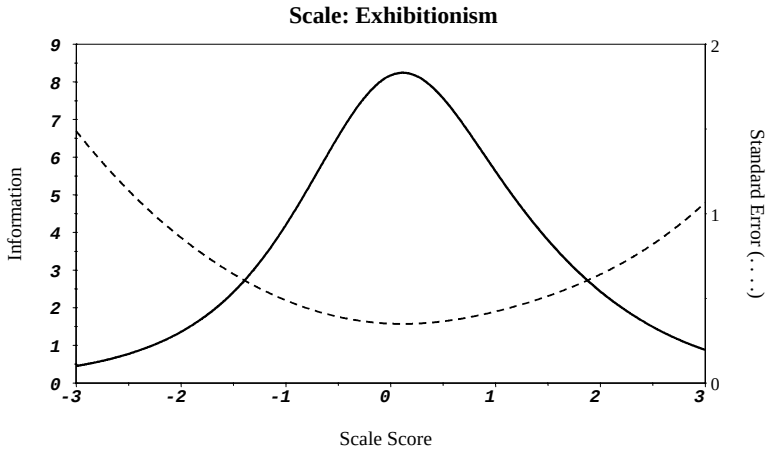
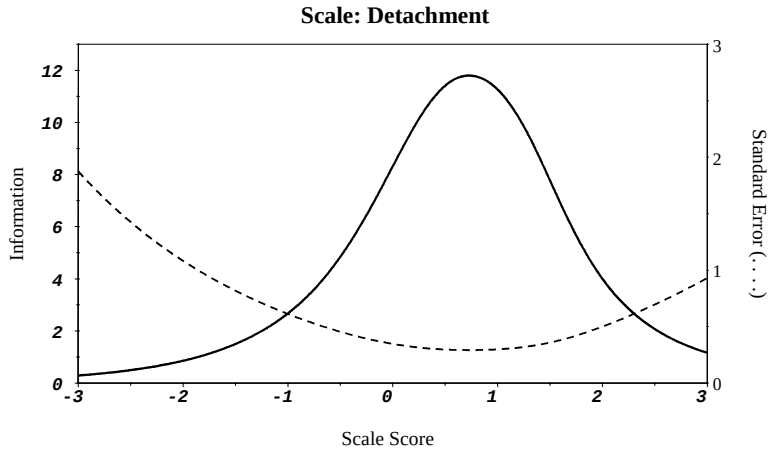
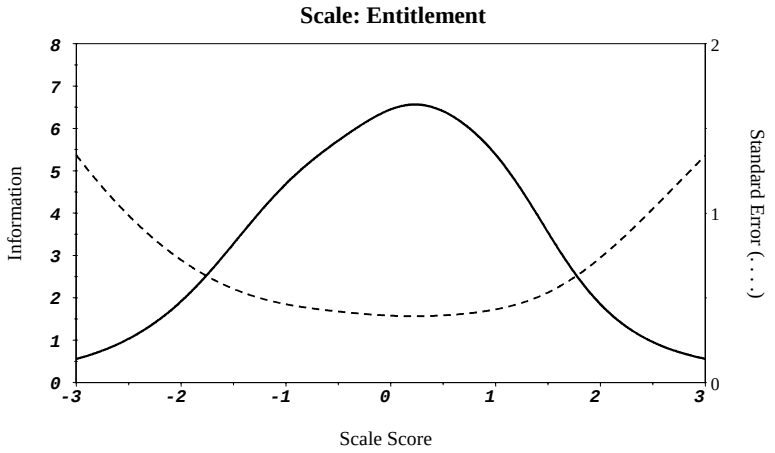


Figure 2.2-continued







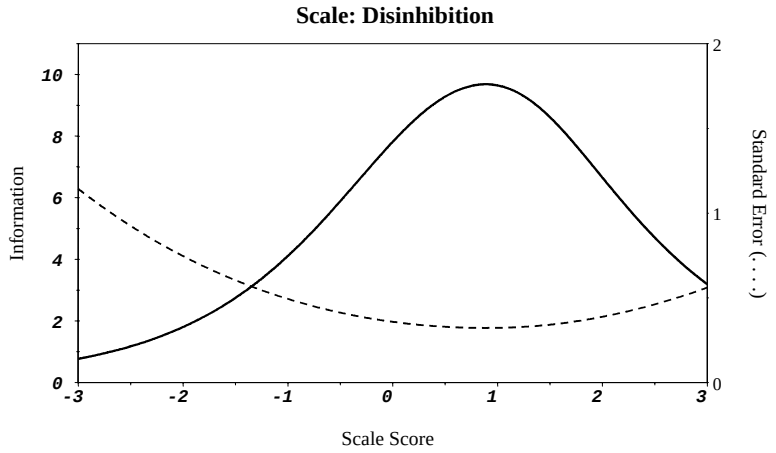
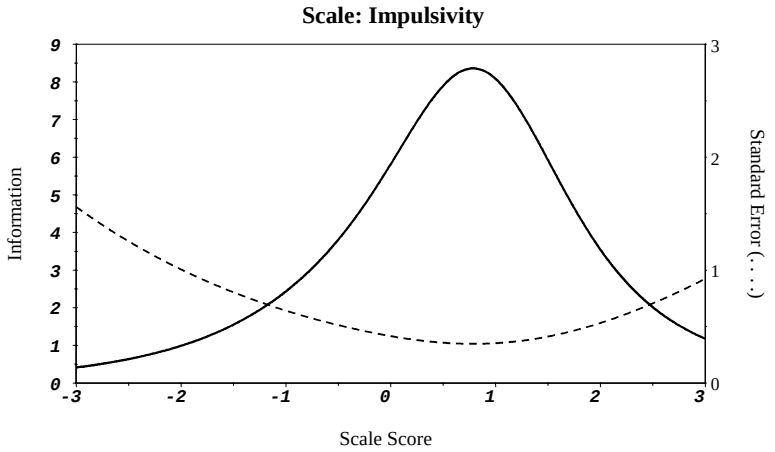
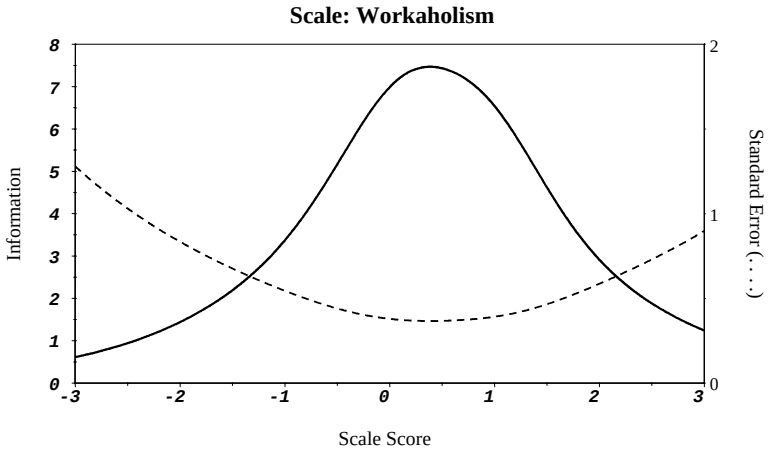


Figure 2.2-continued





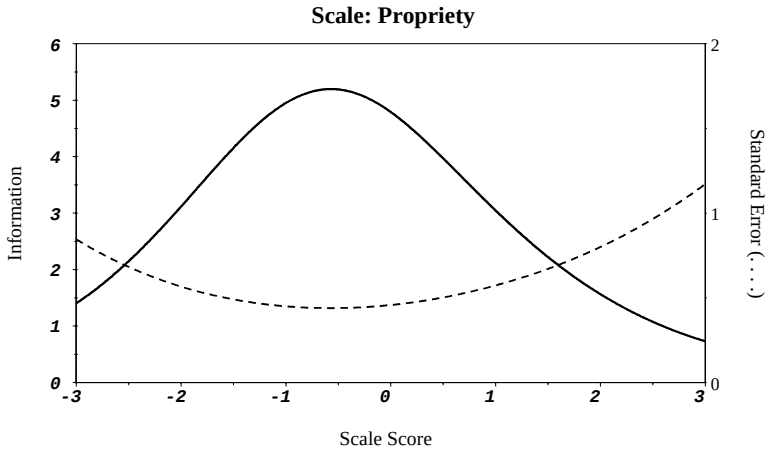


Figure 2.3: Test Fidelity as a Function of Test Length in the Simulation Study.

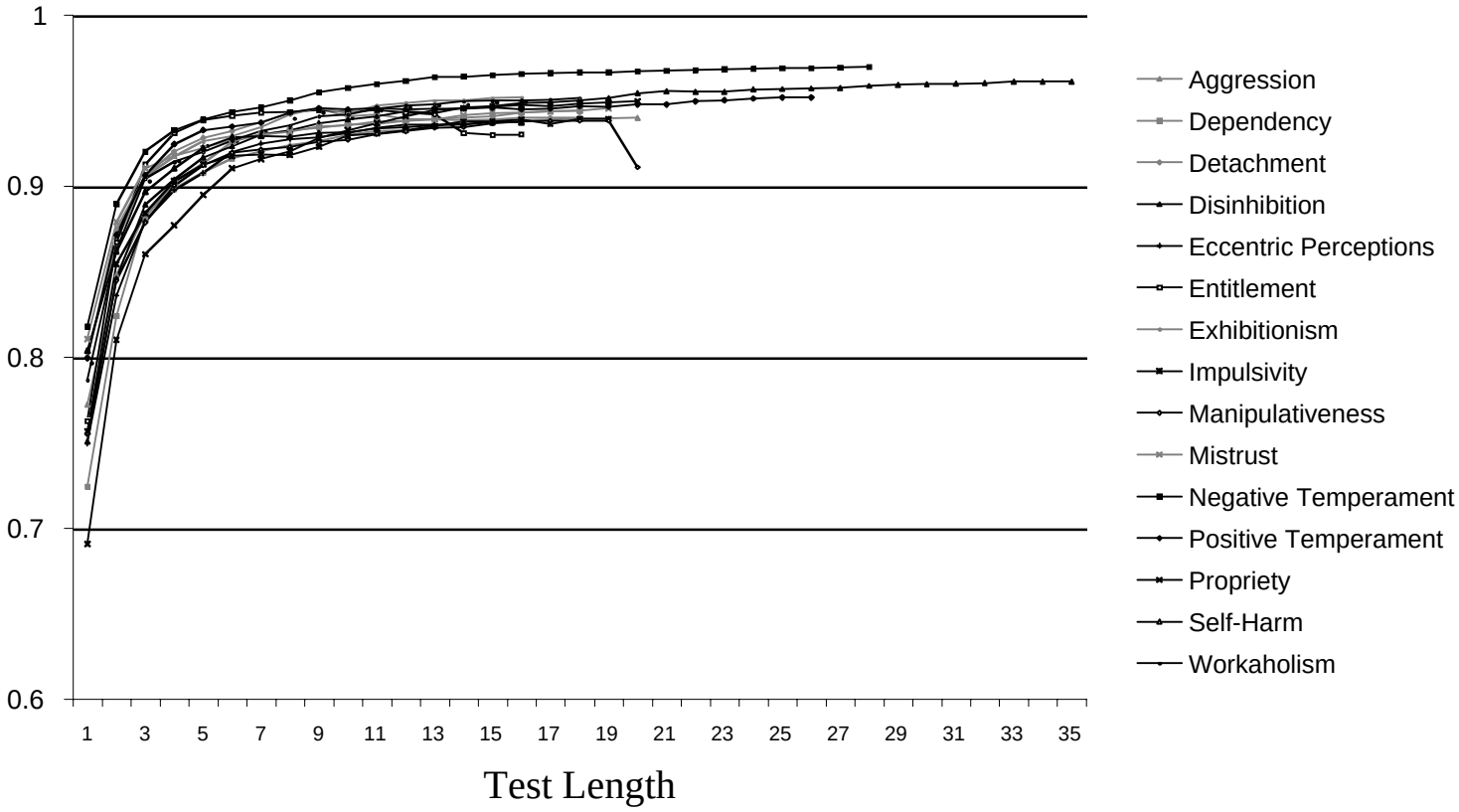


Figure 2.4: SNAP-CAT Organizational Flow Chart.

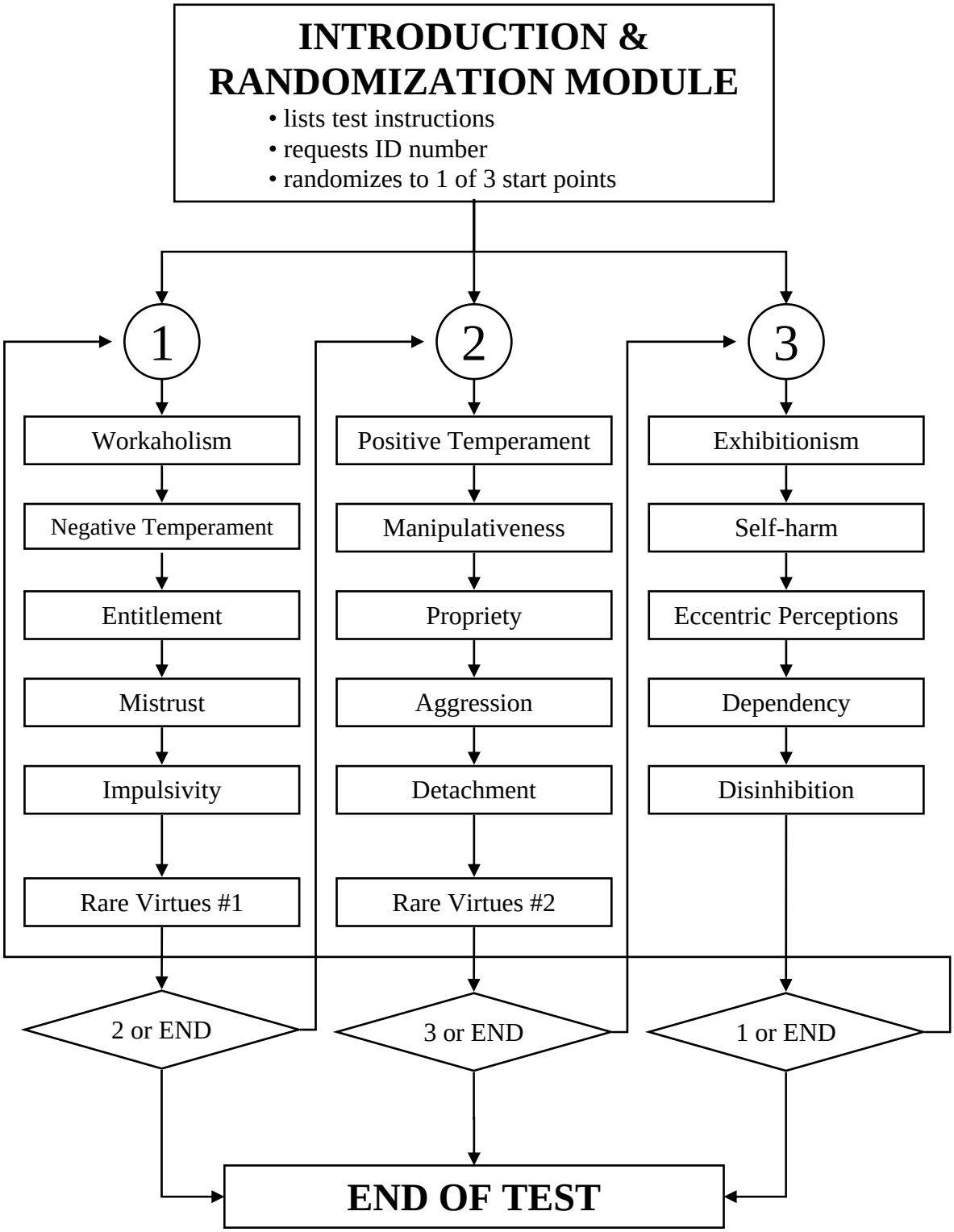
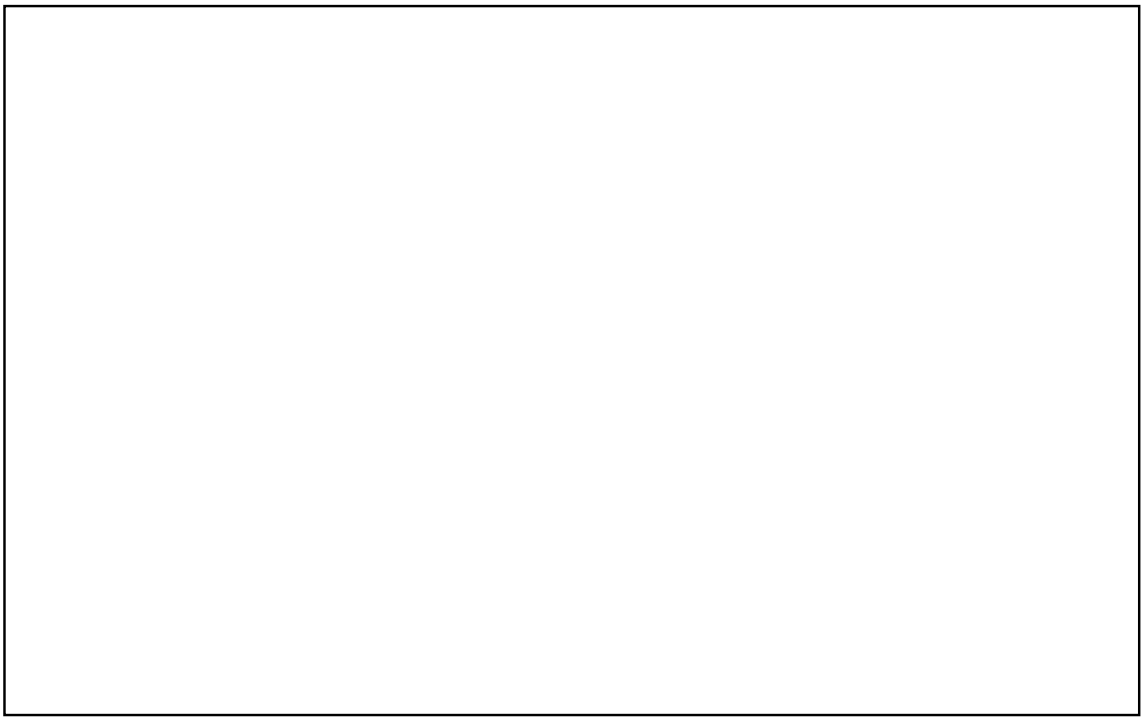


Figure 2.5: Example of SNAP-CAT Item Presentation Screen.



CHAPTER 3
VALIDATION OF SNAP-CAT

Method

Participants

A sample of 491 research participants (67.4% female; 88.2% white; mean age = 19.2 years) was recruited from undergraduate psychology courses offered at the University of Iowa. Course credit and extra credit were offered as compensation for participation. Of these participants, 423 (86.2%) elected to return to the second session one week later. Participants who skipped more than 10 items on either the SNAP or SNAP-CAT were excluded from further analyses. This rule yielded sample sizes of 480 and 413 for Times 1 and 2, respectively. Also, to simplify the presentation and interpretation of results, all subsequent data analyses were restricted to those participants who completed both assessments (N = 413). Demographic characteristics of the final sample appear in Table 3.1. In summary, the sample was 68.5% female, 88.6% white, and 61.5% first-year students. The mean age was 19.2 (SD = 1.3). The demographic characteristics of the final sample did not differ appreciably from those of the full sample, which suggests that participant attrition and exclusion due to excessive item omissions did not markedly influence sample composition.

The rectangular testing room was arranged to accommodate 12 participants. At the far end of the room, six computers were arranged in two rows of three, with the backs of the computers touching. In the front half of the room, participants sat on two sides of a large conference sample, and barricades were placed in the middle of the table to mimic the effect of having a computer in front of participants. As participants arrived for the

first session, they were randomly assigned to one of two blocks of approximately equal size, since some sessions were not full. One block was randomly assigned to the traditional paper-and-pencil version of the SNAP (SNAP-PP); the other block completed the computerized version (SNAP-CAT). Participants were invited back for a second session one week later, but they were blind to which version they would be asked to complete until they arrived for the second session. Half of the returning blocks were randomly assigned to complete the same version of the SNAP, and half were asked to complete the opposite version. This assignment protocol yielded the following four groups of participants:

(1) *P-P* (*n* = 106). Participants in this group completed the paper-and-pencil version of the SNAP twice.

(2) *P-C* (*n* = 105). Those in this condition completed the SNAP-PP first and the SNAP-CAT second.

(3) *C-P* (*n* = 102). Participants assigned to this condition completed the SNAP-CAT first and the SNAP-PP second.

(4) *C-C* (*n* = 100). These participants completed the SNAP-CAT twice.

Table 3.1 includes demographic characteristics for these four groups. Visual inspection of these data indicates that randomization successfully yielded demographically equivalent groups.

Testing Procedures

During Session 1, participants in all four groups completed a general demographic information sheet, the 44-item version of the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991; Benet-Martinez & John, 1998), a state form of the Positive and Negative

Affect Schedule-Expanded version (PANAS-X; Watson & Clark, 1994, completed both pre- and post-SNAP), and either the SNAP-PP or the SNAP-CAT (decided by random assignment of blocks). One week later, at Session 2, returning participants completed either the SNAP-PP or SNAP-CAT (depending on original group assignment) as well as the PANAS-X (again completed both pre- and post-SNAP), and the Eysenck Personality Questionnaire-Revised (Eysenck & Eysenck, 1991). During Session 1, participants completed measures in the following order: (a) Demographic Information Sheet, (b) PANAS-X (pre-SNAP), (c) SNAP-PP or -CAT, (d) PANAS-X (post-SNAP), and (e) BFI. At Session 2, measures were completed in the following order: (a) PANAS-X (pre-SNAP), (b) SNAP-PP or –CAT, (c) PANAS-X (post-SNAP), and (d) EPQ. Completion times for the SNAP-PP and SNAP-CAT were recorded. At the end of Session 2, participants were asked several questions regarding administration mode preference (i.e., SNAP-PP vs. SNAP-CAT), and were provided further information about the study. All measures other than the SNAP-CAT were completed in a paper-and-pencil format, and participants recorded their responses on scannable answer sheets.

Measures

SNAP-CAT and SNAP-PP

The methods used to create the computerized adaptive version of the SNAP were described in Chapter 2. The traditional SNAP was described in Chapter 1. However, in order to provide for fair comparisons between the SNAP-CAT and SNAP, a special P&P version of the SNAP (SNAP-PP) was constructed which contained only those items that were in the SNAP-CAT item pool. Thus, items that loaded only on scales that were not assessed by the SNAP-CAT (i.e., the diagnostic scales) were not included in the SNAP-

PP. The resultant measure contained 297 items that were listed in standard booklet order (minus the diagnostic scale items that were removed). Items were presented in 12-point Times New Roman typeface in a booklet designed to look very similar to the traditional SNAP item booklet. Approximately 30 items were presented on each page in two columns separated by a vertical line. Instructions printed on the front cover of the booklet asked participants to record their responses on a separate scannable answer sheet by filling in consecutively numbered circles marked either "T" or "F."

In the traditional P&P format, SNAP scales have been shown to be internally consistent (median alpha = .81) and temporally stable (median 2-month retest = .79; Clark, 1993). In addition, the SNAP has demonstrated satisfactory levels of convergent and discriminant validity in a variety of settings. The SNAP has been shown to be meaningfully related to the Five-Factor Model of personality (Clark, 1993; Clark et al., 1994; Reynolds & Clark, 2001), the "Big Three" personality traits (Clark, 1993), state and trait mood measures (Clark, 1993), other measures of personality disorder (Clark et al., 1996), and the MMPI-2 (Clark, 1993; Vittengl et al., 1999).

Big Five Inventory

The Big Five Inventory (BFI; Benet-Martinez & John, 1998; John, Donahue, & Kentle, 1991) is a 44-item scale that uses a 5-point Likert-type rating scale, ranging from 1 (strongly disagree) to 5 (strongly agree), and provides scores on the domains of the Five-Factor Model of personality (Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness). Benet-Martinez and John (1998) reported Cronbach's alpha coefficients of .84, .88, .82, .79, and .81 for the traits listed above, respectively, in a sample of 711 English-speaking participants. They also report on unpublished data

indicating good convergence between BFI scales and two established measures of the Five Factor Model.

Eysenck Personality Questionnaire-Revised

The Eysenck Personality Questionnaire-Revised (EPQ-R; Eysenck & Eysenck, 1991) is a 100-item revision of the EPQ (Eysenck & Eysenck, 1975) that provides scores on three broad factors of personality: Neuroticism, Extraversion, and Psychoticism. The major revisions of the EPQ-R involve the Psychoticism scale, which is now measured by 19 of the original 25 items plus 13 new items. The items are answered using a yes/no response format. Published internal consistency reliabilities range from .78 to .90 (Eysenck & Eysenck, 1991).

Positive and Negative Affect Schedule-Expanded Form

The Positive and Negative Affect Schedule-Expanded Form (PANAS-X; Watson & Clark, 1994) is a mood rating form with 60 words or phrases describing a range of affective states. In addition to two general mood dimensions (Positive and Negative Affect), the measure provides scores on 11 specific affect scales, including four core negative affects (Fear, Hostility, Guilt, and Sadness), three core positive affects (Joviality, Self-Assurance, and Attentiveness) and four other affective states (Shyness, Fatigue, Serenity, Surprise). In this study, the PANAS-X was administered using "state" instructions (i.e., how do participants feel *in the moment*). According to the PANAS-X manual (Watson & Clark, 1994), the general and specific affect scales are internally consistent, with alphas ranging from .72 to .90 across samples and instruction sets.

Demographic information sheet

Participants recorded their age, sex, ethnicity, and year in school on this questionnaire. In addition, at the end of Session 2, participants responded to several questions designed to gauge their preferred mode of test administration during this study.

Data Analyses and Results

Analyses were conducted to assess test characteristics such as item and time savings associated with adaptive item administration, examine the psychometric equivalence of scores derived from the SNAP-PP and SNAP-CAT, evaluate the internal and external validity as well as the structural equivalence of the SNAP-CAT, and ascertain the experiential equivalence of the two test forms. Given the large number of statistical analyses and multiple comparisons, the significance level for all hypothesis tests was conservatively set to .01. In addition, Bonferroni corrections were applied when appropriate to guard against a Type I error. Finally, to maintain consistency across scoring methods, omitted SNAP-PP and SNAP-CAT item responses in the final sample of 413 participants were replaced with responses in the non-keyed direction.

Test Characteristics

Item and Time Savings

To test directly the effect of testing mode (i.e., computerized vs. paper-and-pencil) and scoring method (i.e., raw vs. adaptive scoring) on *item* and *time* savings, participants in the two computerized and two paper-and-pencil groups were collapsed into single groups, separately for Times 1 and 2. For each of the following analyses, paper-and-pencil (PP), computerized adaptive (CA), and computerized full-scale (CF) administrations were compared. CF refers to administration of *all* scale items on the

computer (recall that all remaining items were administered in the computerized groups after the termination rule was satisfied). PP vs. CF and PP vs. CA contrasts were conducted with standard between-subjects *t* tests, and the CF vs. CA comparisons were completed using within-subjects paired *t* tests.

Figure 3.1 reveals significant overall *time* savings associated with CF and CA administration of the SNAP. A paired *t* test revealed a significant decrease in overall completion time from Time 1 (M = 25.7 minutes, SD = 6.7) to Time 2 (M = 21.5 minutes, SD = 6.0), $t(412) = 22.52$. At Time 1, mean completion times for the PP, CF, and CA administrations were 30.7 (SD = 5.0), 20.6 (SD = 3.6), and 12.9 (SD = 2.5) minutes, respectively; all differences were significant, $t(411) = 23.27$, $t(411) = 45.13$, and $t(204) = 67.04$, for the PP-CF, PP-CA, and CF-CA comparisons, respectively. Likewise, at Time 2, mean completion times for the PP, CF, and CA administrations were 26.1 (SD = 4.4), 16.8 (SD = 3.2), and 10.4 (SD = 2.2) minutes, respectively; again, all differences were significant, $t(411) = 24.10$, $t(411) = 45.17$, and $t(204) = 66.14$, for the PP-CF, PP-CA, and CF-CA comparisons, respectively. Thus, the SNAP-CAT yielded overall *time* savings of 57.8% and 60.1% at Times 1 and 2, respectively, when compared to the SNAP-PP, and 37.3% and 38.1% at Times 1 and 2, respectively, when compared to the full SNAP administered on the computer.

Item and *time* savings were further quantified at the scale-level in the computerized groups. Results of these analyses appear in Tables 3.2 and 3.3 for Times 1 and 2, respectively. Mean scale-level *item* savings were 36.3% and 36.7% at Times 1 and 2, respectively; similarly, mean *time* savings were 37.7% and 38.2% at Times 1 and 2, respectively. All *item* savings and *time* savings were statistically significant, *t*s(204)

ranged from 12.33 to 151.32 at Time 1, *t*s(201) ranged from 13.32 to 130.22 at Time 2. Positive Temperament yielded the most *item* savings (50.6% and 51.8% at Times 1 and 2, respectively) and the most *time* savings at Time 1 (51.3%), with Disinhibition yielding the most *time* savings at Time 2 (53.5%). Propriety consistently yielded the least *item* savings (8.5% and 14.2% at Times 1 and 2, respectively) and *time* savings (7.6% and 13.8% at Times 1 and 2, respectively). Propriety was a clear outlier (the next least efficient scale, Self-harm, yielded *item* and *time* savings of 24.6% and 28.2%, respectively, at Time 1); this result was predictable given the performance of Propriety in the simulation study described in Chapter 2.

Termination Criteria

Table 3.4 details the frequencies with which each termination criterion was satisfied in the computerized participants. The overall termination pattern was reasonably consistent across scales and revealed that the vast majority of participants satisfied one of the two termination criteria (i.e., SEM < 0.40 or conditional item information < 0.10 for all remaining items) before completing all items. In fact, on 10 of 15 scales at Time 1 and 11 of 15 scales at Time 2, no participants reached the end of the scale during the adaptive administration phase. Another revealing feature of the table is the termination result for Propriety. Unlike all other scales, the SEM of Propriety trait estimates never dropped below the termination threshold. Instead, most terminations occurred when informative items no longer existed. These data help explain why item and time savings were markedly lower for Propriety than for all other scales.

Tables 3.5 and 3.6 further describe the scale-level response patterns of participants, separately for Times 1 and 2. Most scales (i.e., Negative Temperament,

Mistrust, Manipulativeness, Aggression, Eccentric Perceptions, Dependency, Positive Temperament, Entitlement, Detachment, and Disinhibition) yielded a skewed pattern whereby most participants terminated near the established item presentation minimums. Three scales—Exhibitionism, Impulsivity, and Workaholism—followed a more uniform item presentation pattern. Finally, two scales—Self-harm and Propriety—deviated markedly from these patterns in unique ways. The reasons for these deviant response patterns will be explored fully in Chapter 4.

Psychometric Equivalence

Items responses were initially scored in the following four ways: (a) full-scale thetas, (b) adaptively-derived thetas, (c) traditional raw scores, and (d) "true scores" (i.e., on the raw score metric) estimated from adaptively-derived thetas. These methods were chosen to examine the effect, if any, of using different scoring methods on mean-level differences across groups. Full-scale thetas were calculated on *all item responses* in each scale using the EAP estimation methods described above (Assessment System Corporation, 1996; Mislevy & Bock, 1990). SNAP-PP thetas were calculated a posteriori using BILOG, whereas SNAP-CAT thetas were calculated using MicroCAT during test administration. Adaptively-derived thetas were calculated on SNAP-CAT adaptive responses using EAP estimation methods and were calculated and stored by MicroCAT when the termination criterion was satisfied for each participant. Traditional raw scores were calculated by simply summing keyed item responses, for all items within each scale, in a manner consistent with the SNAP manual for both the SNAP-PP and SNAP-CAT. Finally, estimated true scores (e.g., see Hambleton, Swaminathan, & Rogers, 1991) were calculated from SNAP-CAT adaptively-derived thetas by summing the keyed response

probabilities (i.e., calculated from the ICC function) across *all items* in a given test, given the adaptively-derived theta. Estimated trues scores are on the raw score metric and can be directly compared with raw scores computed with traditional methods.

Repeated Measures Analyses

To test for mean-level differences among groups, a series of repeated measures ANOVAs was conducted with follow-up between-subjects contrasts computed separately for Time 1 and Time 2 participants. Each repeated measures ANOVA included main effects for Group (P-P, C-P, P-C, and C-C), Sex (male vs. female), and Time (Session 1 vs. 2), as well as Group*Sex and Group*Time interactions. Dependent variables for all ANOVAs were SNAP-PP or SNAP-CAT scale scores, computed using each of the four scoring methods described above. Notably, adaptively-derived thetas and estimated true scores were calculated in the computerized cells only, and these were compared with full-scale thetas and traditional raw scores, respectively, in the paper-and-pencil cells. Table 3.7 contains a summary of significant effects for all repeated measures ANOVAs.

The Group effect, which averaged participants across Sex and Time, was modeled in order to examine whether random assignment successfully yielded equivalent groups in terms of these personality variables. Thus, Group effects generally were not expected to be significant, and results indicated that the Group effect was significant for only two scales: Aggression and Self-harm. Bonferroni-corrected follow-up tests indicated that participants in the C-P group scored significantly higher those in the P-P group only on Aggression traditional raw scores, $t(405) = 3.17$. Self-harm follow-up tests revealed that the P-P group scored significantly lower than the P-C and C-C groups using three scoring methods—estimated true scores, $ts(405) = 4.60$ and 3.76, respectively, full-scale thetas,

*t*s(405) = 6.25 and 6.85, respectively, and adaptively-derived thetas, *t*s(405) = 6.45 and 7.23, respectively—as well as the C-P group with two scoring methods—full-scale thetas, *t*(405) = 3.99, and adaptively-derived thetas, *t*(405) = 4.28. The relatively small number of group main effects suggests that the randomization scheme assigned participants to groups in a reasonably equivalent manner with respect to personality variables.

Based on previous SNAP data (e.g., Clark, 1993), significant Sex effects were expected for several scales. In the present study, males scored significantly higher than females, using all scoring methods, on Aggression, *t*s(405) ranged from 5.23 to 5.50, Detachment, *t*s(405) ranged from 2.82 to 3.69, Disinhibition, *t*s(405) ranged from 5.48 to 5.51, Impulsivity, *t*s(405) ranged from 2.83 to 3.60, and Manipulativeness, *t*s(405) ranged from 5.18 to 5.50. Likewise, female participants scored significantly higher than males, using all scoring methods, on Dependency, *t*s(405) ranged from 3.60 to 3.77, and Propriety, *t*s(405) ranged from 3.54 to 3.76. In addition, females scored significantly higher than males, using only traditional raw and estimated true scores, on Negative Temperament, *t*s(405) = 2.79 and 2.86, respectively. These findings are congruent with the sex differences identified in previous studies (e.g., Clark, 1993) and were largely consistent across scoring methods.

Time effects were modeled to account for mean score shifts across Groups. Significant effects were identified for several scales. Time 1 scores were significantly higher than Time 2 scores, using all scoring methods, for Aggression, *t*s(405) ranged from 2.97 to 4.22, Eccentric Perceptions, *t*s(405) ranged from 4.33 to 5.11, Negative Temperament, *t*s(405) ranged from 2.99 to 4.31, and Self-harm, *t*s(405) ranged from 4.30 to 5.02. Mistrust yielded higher Time 1 traditional raw scores, full-scale thetas, and

adaptively-derived thetas, *t*s(405) ranged from 2.81 to 3.77, and Dependency estimated true scores and adaptively-derived thetas were significantly higher at Time 1, *t*s(405) = 2.81 and 3.24, respectively. Time 2 scores were significantly higher than Time 1 scores, using all scoring methods, on Positive Temperament, *t*s(405) ranged from 4.38 to 5.67, and Propriety, *t*s(405) ranged from 4.02 to 5.36. In addition, Entitlement traditional raw scores and full-scale thetas were higher at Time 2, *t*s(405) = 2.97 and 3.04, respectively. Interestingly, the differences between Time 1 and Time 2 scores were uniformly indicative of decreased pathology at retest, which is consistent with previous studies of mean trait-level over time (Schubert, 1975; Windle, 1954, 1955).

The Group*Sex interaction was modeled to determine whether Sex effects were distributed evenly across Groups. This interaction was significant only for Aggression using traditional raw and estimated true scoring. Given that only 1 of 15 scales and only 2 significant tests out of a possible 60 (3.3%) produced a significant Group*Sex interaction, it was concluded that this isolated difference was likely due to random error; thus, data were averaged across Sex for subsequent analyses.

The Group*Time interaction was modeled to assess for consistent testing mode effects. The four Groups and two Times yielded eight Group*Time cells. As described above, at each Time level, two groups completed the SNAP-PP and two groups completed the SNAP-CAT, which constituted a built-in replication test of significant effects. Complete descriptive statistics for this interaction appear in Appendix C Tables C1, C2, C3, and C4, for traditional raw scores, estimated true scores, full-scale thetas, and adaptively-derived thetas, respectively. Significant interactions were found for five scales: Dependency, Positive Temperament, Propriety, Workaholism, and Self-harm.

Bonferroni-corrected follow-up tests were conducted; Table 3.8 includes a summary of all significant effects. Of the five scales that yielded significant interactions, four were characterized by minor differences among two or three means that did not replicate meaningfully. For example, for Dependency, in the C-P group higher thetas were obtained by computerized compared to non-computerized participants. However, the P-C group, the counter-balanced partner to the C-P group, did not yield this effect, suggesting that the difference likely was due to a factor other than testing mode (e.g., random error, problems with randomization, etc.). Follow-up tests for Positive Temperament, Propriety, and Workaholism yielded similar inconsistent patterns.

However, the fifth scale with a significant interaction, Self-harm, yielded a somewhat more complex picture in which computerized means tended to be higher than P&P means using estimated true scores and all thetas, *t*s(405) ranged from 3.48 and 14.75. Interestingly, however, this pattern was not observed with traditional raw scores, which suggests that the differences may be related to score conversion to and from the theta metric. In summary, these Group*Time interactions revealed slight evidence of differential means across testing modes, with these findings highly inconsistent across cells. Follow-up contrasts were conducted to examine more directly the effect of testing mode.

Follow-up Testing Mode Contrasts

To assess the effect of testing mode (i.e., SNAP-PP vs. SNAP-CAT) on scores, follow-up between- and within-subjects contrasts were conducted by collapsing participants in the two computerized and two paper-and-pencil groups into single groups, separately for Times 1 and 2. For each of the analyses, the scores resulting from paper-

and-pencil (PP), computerized full-scale (CF), and computerized adaptive (CA) item administration modes were compared. PP vs. CF and PP vs. CA contrasts were conducted with standard between-subjects *t* tests, and the CF vs. CA comparisons were completed using within-subjects paired *t* tests. Tests were Bonferroni-corrected to account for multiple comparisons. Complete descriptive statistics for these comparisons appear in Appendix C Tables C5, C6, C7, and C8, for scores on the Time 1 raw score metric, Time 2 raw score metric, Time 1 theta metric, and Time 2 theta metric, respectively.

Significant Bonferroni-corrected contrasts are presented in Table 3.9. The between-subjects contrasts revealed that only three scales—Propriety, Self-harm, and Workaholism—yielded significantly different means across modes, and only the Self-harm differences replicated across time. For example, at Time 2, Propriety CF and CA means were significantly higher than PP means on both the raw score and theta metrics, *t*s(411) ranged from 2.95 to 3.20, but a similar result did not emerge at Time 1, which suggests that the finding is subsample-specific and may have resulted from some failure of random assignment. Similarly, the significant Time 1 Workaholism results did not replicate at Time 2. Self-harm, however, produced an interesting and somewhat consistent pattern of means. CA (but *not* CF) estimated true scores were consistently and significantly higher than PP traditional raw scores, *t*s(411) = from 2.92 to 5.30 at Times 1 and 2, respectively. On the theta metric, however, both CA *and* CF means were greater than PP thetas at Times 1 and 2, *t*s(411) ranged from 7.30 to 10.17. The consistent discrepancy between raw and theta scoring suggests that conversion to and from the theta metric may artificially inflate Self-harm scores.

Given the greater statistical power associated with within-subjects tests, the CF vs. CA comparisons revealed more significant differences. Within-subject differences were identified for eight scales, but again, most did not replicate. Three scales—Entitlement (raw and theta metrics), Propriety (raw metric only), and Workaholism (raw metric only)—yielded higher means for CA administration at Time 1 only, *t*s(204) ranged from 3.17 to 6.59; two scales—Detachment (raw metric only) and Negative Temperament (raw and theta metrics)—produced higher means for CA administration at Time 2 only, *t*s(201) ranged from 4.37 to 6.11; and two scales—Dependency (raw metric only) and Disinhibition (raw and theta metrics)—generated higher means for CF administration at Time 2 only, *t*s(201) ranged from 3.45 to 3.61. Finally, only Self-harm yielded consistent differences, with CA estimated true scores higher than CF traditional raw scores at Times 1 and 2, t(204) = 9.48 and t(201) = 7.41, respectively.

In summary, with the notable exception of Self-harm, no between- and within-subjects differences replicated across testing sessions, suggesting that random error or subsample differences likely caused many of the differences. To interpret the magnitude of the above-identified differences, *T*-score differences (i.e., scaled to the mean and standard deviation of normative sample, $M = 50$, $SD = 10$) were computed. These calculations revealed that all mean-level differences, except for Self-harm, translated to *T*-score differences of 3.2 or less, and most differences were less than 1.2 *T*-score points. Such differences would not significantly alter clinical or research interpretation of the scales. However, *T*-score differences associated with Self-harm were as high as 4.7, which could potentially affect scale interpretation.

Validity Scales

As described briefly in Chapter 1, the SNAP includes five validity scales—Rare Virtues, Deviance, Variable Response Inconsistency (VRIN), True Response Inconsistency (TRIN), and Desirable Response Inconsistency (DRIN)—and an overall Invalidity Index, that were designed to detect several common response patterns (e.g., defensive, exaggerated, or inconsistent responding). Although these scales are not amenable to IRT or CAT, they were included in the validation study to test for testing mode effects on response bias or inconsistency. Between-subjects *t* tests were conducted comparing P&P to computerized item presentation formats, separately at Times 1 and 2. Descriptive statistics for these analyses appear in Table 3.10. Results revealed no replicable validity scale differences across testing modes. At Time 1, Deviance and the Invalidity Index, *t*s(411) = 2.83 and 2.42, respectively, were significantly higher in the computerized participants; however, these differences did not replicate at Time 2. Thus, the time savings associated with computerized administration did not appear to yield replicable responses biases or inconsistent responding.

Variance Differences

Another important element of psychometric equivalence of test forms is equality of scale variances across modes. Thus, equality of variances was assessed for each of the cross-mode contrasts described above (i.e., CA vs. PP, CF vs. PP). Scale standard deviations appear in Tables C5, C6, C7, and C8, in Appendix C, for scores on the Time 1 raw metric, Time 2 raw metric, Time 1 theta metric, and Time 2 theta metric, respectively. Visual inspection of the raw metric standard deviations revealed that 14 of the 15 CA variances were lower than the PP variances, at both Times 1 and 2. Of these differences, seven (Disinhibition, Entitlement, Exhibitionism, Impulsivity,

Manipulativeness, Propriety, and Self-harm) were significant or near-significant at Time 1, *F*s'(207, 204) ranged from 1.40 to 1.91, and eight (add Aggression and Dependency and remove Exhibitionism from the list above) were significant at Time 2, *F*s'(210, 201) ranged from 1.44 to 1.80. No CF raw score variances were significantly different from the PP variances. On the theta metric, only Self-harm and Disinhibition yielded significantly different variances. For Self-harm, CA and CF variances were lower than PP variances at Time 1, *F*s'(207, 204) = 2.24 and 2.07, respectively, and Time 2, *F*s'(210, 201) = 2.10 and 1.81, respectively. Finally, Disinhibition's CA variance was significantly lower than the PP variance at Time 1 only, *F*'(207, 204) = 1.50.

Taken together, these findings suggest that traditional raw scoring (i.e., generated from P&P administration) resulted in higher score variances than estimated true scores generated from adaptive-test thetas, and these variance differences largely disappeared when all scores were converted to the theta metric. To further aid in interpretation of these results, a posteriori analyses were conducted to compare PP variances to those of estimated true scores calculated from *full-scale thetas* in the computerized groups. Results revealed that, again, estimated true score variances were generally lower than PP traditional raw score variances at both Times 1 and 2. At Time 1, six scales (Disinhibition, Entitlement, Exhibitionism, Manipulativeness, Propriety, and Self-harm) yielded significantly lower true score variances, *F*s'(207, 204) ranged from 1.43 to 1.84. Likewise, at Time 2, six scales (Aggression, Disinhibition, Entitlement, Impulsivity, Manipulativeness, and Propriety) yielded significantly lower true score variances, *F*s'(210, 201) ranged from 1.48 to 1.76. These results largely parallel the initial analyses described above and point generally to lower variances when thetas, both adaptively-

derived and full-scale, were converted to estimated true scores. However, none of these results was unexpected, given that EAP theta estimation generally results in lower standard deviations than other methods such as maximum likelihood estimation (Mislevy and Bock, 1990). The significance and ramifications of this issue are discussed in Chapter 4.

Test Stability

To determine whether the SNAP-PP and SNAP-CAT can be considered parallel forms, test-retest correlational analyses were conducted both within- and between-modes; these results appear in Table 3.11. Retest coefficients were calculated on paper-and-pencil traditional raw scores and full-scale thetas, computerized traditional raw scores, and adaptively-derived thetas calculated during computerized administration. The correlations obtained in the P-P group represent typical paper-and-pencil retest correlations and served as the baseline from which to compare the cross-form correlations obtained in the P-C and C-P groups, as well as the retest correlation in the C-C group. Test-retest correlations were generally high across groups and methods of scoring, again indicating good equivalence between test forms. In the P-P group, mean retest correlations were .88 (range = .76 to .93) and .87 (range = .75 to .93) for raw and theta scoring, respectively. In the C-P group, mean retest correlations were .85 (range = .74 to .91) and .82 (range = .71 to .88) for raw and theta scoring, respectively. In the P-C group, mean retest correlations were .88 (range = .81 to .93) and .85 (range = .77 to .90) for raw and theta scoring, respectively. Finally, in the C-C group, mean retest correlations were .88 (range = .80 to .94) and .84 (range = .73 to .89) for raw and theta scoring, respectively. These retest correlations were remarkably consistent across

groups, suggesting good equivalence between testing modes. Retest correlations for adaptively-derived thetas were consistently, but only slightly, lower than those for all raw scores and paper-and-pencil full-scale thetas.

Correlations between raw scores and thetas were calculated to study scoring method equivalence; these appear in Table 3.12. It is notable that, because they were computed on the same set of item responses, these coefficients are appreciably higher than the retest correlations described above. The mean raw-to-theta correlation for computerized participants was .94 at Times 1 and 2 (ranges = .90 to .95, and .89 to .97, at Times 1 and 2, respectively). Likewise, the mean raw-to-theta correlation for paper-and-pencil participants was .98 at Times 1 and 2 (ranges = .91 to .99, and .96 to .99, at Times 1 and 2, respectively). Taken together, these findings suggest that the slightly lower retest coefficients in the computerized groups are likely related to measurement error associated with adaptively-derived thetas (which, of course, are based on fewer items than raw scores or P&P thetas and may be based on different sets of items when both administrations are computer-based).

Internal Consistency

Standard internal consistency analyses appear in Table 3.13. Whereas Cronbach's alpha coefficients are not usually calculable for adaptively administered scales (i.e., because not all items and different sets of items are administered to all examinees), I chose to compute them here, based on all item responses, to evaluate whether grouping items, as was done on the SNAP-CAT, had any effect on scale internal consistency. The mean computerized alphas were .82 (range = .73 to .90) and .85 (range = .78 to .92) at Times 1 and 2, respectively. Similarly, the mean P&P alphas were .83 (range = .79 to

.92) and .85 (range = .81 to .92) at Times 1 and 2, respectively. These alphas are consistent with those found in the calibration sample (see Table 2.2) as well as those published in the SNAP manual (Clark, 1993). Thus, neither item grouping nor computerized administration appeared to affect internal consistency.

Internal and External Validity

Internal Structure

Test equivalence also requires that both forms of a given test generate similar internal and external correlational structures. To assess internal structure, scale-level principal factor analyses were conducted, with Varimax rotation (i.e., maintaining orthogonal factors), across different testing modes and scoring methods. The scree plot as well as previous factor analytic studies of the SNAP suggested that three factors should be extracted (eigenvalues ranged from 3.15 to 3.38, 1.79 to 2.25, 1.06 to 1.75, and 0.67 to 0.72 for the first four factors, respectively). Factor loadings for Times 1 and 2 appear in Tables 3.14 and 3.15, respectively. Before reporting the results, it is important to note that there is an important difference between the present factor analyses and those typically conducted on the SNAP: the version of Disinhibition used here contains overlapping variance with several others scales (e.g., Impulsivity, Manipulativeness, and Propriety). Typically, SNAP factor analyses are conducted with a “pure” version of Disinhibition that contains no item overlap. That was not possible here because the adaptively-derived thetas were based on subsets of Disinhibition items that varied across individuals. Thus, it was impossible to extract the overlapping variance. So, to directly test structural similarity, non-corrected Disinhibition scores were used for all analyses, including those in which creation of a pure version would have been possible.

Despite this major difference, the factor loadings were reasonably comparable with those identified elsewhere (e.g., Clark, 1993). Across administration modes, scoring methods, and time, the SNAP scales formed three general factors that were labeled *Negative Affectivity (NA)*, *Positive Affectivity (PA)*, and *Disinhibition vs. Constraint (DvC)*. The NA factor included consistent and strong loadings for Negative Temperament, Mistrust, Aggression, Self-harm, and Eccentric Perceptions, Manipulativeness (which split with DvC), and Detachment (which split with PA). Dependency, which usually loads less strongly on the first factor, did the same here. PA included strong and relatively stable loadings for Positive Temperament, Exhibitionism, Entitlement, and Detachment. In addition, Workaholism and Propriety loaded moderately on PA in the computerized groups on both the raw and theta metrics. Finally, the DvC factor yielded high loadings for Disinhibition, Impulsivity, Propriety, Workaholism, and Manipulativeness (which split with NA). Two observations are notable. First, the structure was remarkably similar from Time 1 to Time 2. Second, likely owing to the overlapping variance associated with Disinhibition, loadings for Disinhibition on the DvC factor were greater than those usually found (e.g., Clark, 1993).

To quantify the structural similarity of these loading matrices, *congruence coefficients* (Tucker, 1951) were computed and appear in Table 3.16. In general, congruence coefficients range from –1.0 to +1.0 and are interpreted in a manner similar to Pearson correlation coefficients. Values at or above +.90 generally indicate good factor congruence across structures. The pattern of coefficients was consistent across factors at Times 1 and 2. Two observations are noteworthy. The first is that all coefficients are .90 or above, suggesting that the factor structure of the SNAP was

comparable across modes of administration and scoring methods. Second, within-mode coefficients were uniformly excellent, with mean coefficients of .99 and .98 for raw-to-theta structural convergence within the P&P and computerized modes, respectively. The cross-mode coefficients, while slightly lower (mean = .94), were still excellent. Congruence coefficients also were computed between Time 1 and Time 2 factor loadings, yielding uniformly excellent cross-session congruence (all coefficients were greater than .97) for all factor matrices. Thus, in summary, the internal covariance structure appeared to replicate well across sessions, testing modes, and scoring metrics.

Convergent and Discriminant Validity

To assess external structural similarity, correlations were computed among SNAP scales and two established measures of personality—the BFI and EPQ-R—which have been shown to correlate with the SNAP in meaningful and predictable ways (e.g., Clark, 1993; Clark et al., 1994; Reynolds & Clark, 2001). Correlations with the BFI appear in Tables 3.17 and 3.18 for Times 1 and 2, respectively. Again, similar to the internal findings described above, the SNAP scales correlated similarly with BFI scales across testing modes and scoring metrics. Although it is outside the scope of this paper to comment on each SNAP correlate, it is clear that the correlations were orderly and interpretable in the context of previous SNAP studies. For example, as expected, the Neuroticism scale of the BFI correlated significantly with several SNAP scales, including Negative Temperament, Mistrust, Aggression, and Self-harm. Further, BFI Agreeableness correlated meaningfully with SNAP Aggression, Manipulativeness, and Mistrust, as well as others. Third, as expected, BFI Conscientiousness was strongly related to the scales within the SNAP's Disinhibition factor.

Correlations with the EPQ-R appear in Tables 3.19 and 3.20 for Times 1 and 2, respectively. Once again, scoring metric and testing mode had little apparent effect on the correlations, and these correlations were orderly and meaningful. For instance, EPQ-R Extraversion correlated strongly with several SNAP scales, including Positive Temperament, Exhibitionism, and Detachment (negatively). Also, EPQ-R Psychoticism was predictably related to the Manipulativeness, Disinhibition, Impulsivity, and Propriety (negatively) scales of the SNAP. Visual inspection of the correlation matrices indicated that both the BFI and EPQ-R findings were quite stable across sessions.

To quantify the similarity of these correlational structures across testing modes and scoring methods, a unique approach was undertaken that utilized structural equation modeling techniques. Specifically, adapting and modifying procedures from a recent longitudinal study of personality change (Robins, Fraley, Roberts, & Trzesniewski, 2001), the correlations among the SNAP, BFI, and EPQ-R scales (a total of 24 scales) were modeled in EQS (Bentler & Wu, 1995) using a series of single-indicator confirmatory factor analyses (CFAs). The CFAs were constructed such that each of 24 factors was modeled by a single scale (i.e., yielding as many factors as variables). The variances of each factor were fixed at 1.0, and those of the residuals were fixed at 0. Initially, to test correlational similarity, intercorrelations among all 24 factors (i.e., scales) were modeled and *constrained to be identical* across four correlation matrices that were calculated using SNAP (a) paper-and-pencil traditional raw scores, (b) paper-and-pencil full-scale thetas, (c) computerized traditional raw scores, and (d) adaptively-derived thetas calculated during computerized administration. Analyses were completed separately for Times 1 and 2. Unfortunately, the 24x24 correlation matrices were too

large and created so many constraints that EQS was unable to converge on a clean solution.

Thus, the design was modified to break the task into smaller, more computationally manageable chunks. In the new procedure, each SNAP scale was assessed individually along with the nine correlates, creating a 10x10 matrix of correlations including a single SNAP scale, five BFI scales and four EPQ-R scales. Next, in each of 15 CFAs (i.e., one for each SNAP scale) conducted at both Times 1 and 2, four 10x10 correlation matrices (i.e., one for each testing mode/scoring metric combination described above) were analyzed. Correlations between the SNAP scale and correlates (i.e., a total of nine correlations) were constrained across matrices, but the intercorrelations among the correlates were permitted to vary. The goodness-of-fit of each CFA was evaluated by visually inspecting three fit indices—the overall model χ^2, the standardized root-mean-square residual (SRMR), and the Bentler-Bonett Normed Fit Index (NFI). Following the recent recommendations of Hu & Bentler (1999), fit was considered to be excellent if (a) the overall model χ^2 was not significant (which often can be difficult to achieve with large sample sizes), (b) SRMR was .08 or less, and (c) NFI was .95 or greater. Following each CFA were Lagrange Multiplier (LM) tests to determine whether the release of any constraints would improve model fit significantly. This procedure yielded interpretable results and was repeated for Time 2 results.

The resultant fit indices appear in Table 3.21. At Times 1 and 2, all model χ^2 statistics were non-significant, and all NFI and SRMR values were in the excellent range. Thus, the overall convergent and discriminant structure of the SNAP did not appear to be affected significantly by differences in testing mode or scoring metric. However, the LM

tests revealed four constraints (one at Time 1 and three at Time 2) that, if released, would improve fit significantly. At Time 1, the LM tests suggested that the parameters constraining the correlation between Eccentric Perceptions and BFI Agreeableness could be released, $\chi^2\Delta(1) = 8.61$. At Time 2, the LM tests revealed that the parameters constraining the correlations of Aggression to BFI Conscientiousness, Mistrust to EPQ-R Neuroticism, and Positive Temperament to BFI Openness, $\chi^2\Delta s(1) = 7.83$, 6.77, and 6.31, respectively. In all four cases, the correlational differences that gave rise to the LM test results appeared to be related to administration mode. However, these results were not stable. None of the significant LM tests replicated from Time 1 to Time 2 or vice versa.

To summarize, both internal and external analyses of structure on the SNAP-PP and SNAP-CAT revealed excellent convergence among covariance structures calculated across administration modes and scoring methods.

Experiential Equivalence

Effects on State Mood

In light of suggestions that experiential equivalence be considered when comparing two different forms of the same test (Honaker, 1988; Hofer & Green, 1985), state mood was assessed with the PANAS-X before and after each SNAP-PP and SNAP-CAT administration. Between-subjects analyses of covariance (ANCOVA) were conducted, separately for Times 1 and 2, with computerized vs. P&P group status entered as the independent variable, pre-SNAP PANAS-X score as the covariate, and post-SNAP PANAS-X score as the dependent variable. Such an analysis accounted for pre-test mood while assessing for changes in mood as a function of testing mode. Pre- and post-SNAP means for the PANAS-X appear in Tables 3.22 and 3.23 for Times 1 and 2, respectively.

Across both testing sessions, only three significant differences were identified for post-SNAP PANAS-X means. At Time 1, Guilt was significantly higher in the computerized groups (relative to the pre-SNAP mean) than the paper-and-pencil group, $F(1, 404) = 7.48$. Visual inspection of the means suggested that the cause for this result is that pre-SNAP P&P Guilt scores were higher than post-SNAP P&P scores as well as pre- and post-SNAP computerized group Guilt scores. At Time 2, Fatigue and Shyness scores were higher in the computerized groups (relative to the pre-SNAP means) than in the P&P groups, Fs(1, 404) = 9.4 and 11.74, respectively. Regarding Fatigue, scores in the P&P groups dropped from pre- to post-SNAP, whereas they increased in the computerized groups. This result in interestingly counterintuitive given that P&P administration is significantly more time-consuming. With Shyness, the result seems to be related to a drop in scores from pre- to post-SNAP in the P&P groups. Overall, the mean effects described above were quite small, none of the significant effects replicated from Time 1 to 2, or vice versa, and all appearedx to be somewhat random (i.e., not interpretable) in nature. Thus, state mood did not appear to fluctuate appreciably as a function of testing mode.

Mode Preferences

A final aspect of experiential equivalence is the degree to which participants experienced the test interfaces differentially. Participants in the C-P and P-C groups were asked at the end of the study which mode they preferred and why they preferred it. Of the 207 participants in those two conditions, 87.0% preferred the computerized condition, and they gave several reasons. Figure 3.2 includes a bar chart detailing the frequencies of the reasons that were provided. Multiple responses were permitted. Notably, of the 187

open-ended responses to this query, 48.7% and 24.1% indicated that the computerized version was faster and easier, respectively, than the P&P version. Comparable results for the 13% of participants who preferred the paper-and-pencil version appear in Figure 3.3. The most frequently reported reason to prefer P&P administration (63.9% of the responses) was dislike of not being able to change answers on the computerized version.

Table 3.1: Demographic Characteristics of Validation Sample.

Variable	Sample					
	P-P (n = 106)	C-P (n = 105)	P-C (n = 102)	C-C (n = 100)	Combined (N = 413)	
Age						
M	19.0	19.2	19.3	19.1	19.2	
SD	1.2	1.4	1.5	1.1	1.3	
Gender						
Female	69	72	70	72	283	(68.5%)
Male	37	33	32	28	130	(31.5%)
Ethnicity						
Asian	3	3	3	1	10	(2.4%)
Black	3	3	2	4	12	(2.9%)
Latino	2	3	4	2	11	(2.7%)
Native American	1	0	0	3	4	(1.0%)
White	92	96	92	86	366	(88.6%)
Other/Multi-racial	5	0	1	4	10	(2.4%)
Class Status						
First-year	69	60	61	64	254	(61.5%)
Sophomore	19	21	22	18	80	(19.4%)
Junior	10	14	11	11	46	(11.1%)
Senior	8	10	8	7	33	(8.0%)

Table 3.2: Time 1 Adaptive Item and Time Savings in the Computerized Groups.

	Items		Time (sec.)		
Scale (items)	Adaptive *M* (*SD*)	Savings (%)	Adaptive *M* (*SD*)	Total[a] *M* (*SD*)	Savings (%)
Negative Temperament (28)	14.4 (1.3)	48.6	51.7 (12.8)	102.1 (22.6)	49.3
Mistrust (19)	11.8 (4.0)	38.0	44.8 (18.3)	73.9 (14.9)	39.4
Manipulativeness (20)	13.0 (2.6)	35.0	54.7 (17.7)	91.5 (19.7)	40.1
Aggression (20)	12.7 (3.2)	36.7	40.5 (15.6)	63.7 (14.2)	36.4
Self-harm (16)	12.1 (3.3)	24.6	37.5 (12.6)	52.3 (10.8)	28.2
Eccentric Perceptions (15)	10.0 (2.9)	33.4	48.6 (17.3)	71.6 (16.8)	32.1
Dependency (18)	11.2 (3.0)	37.7	42.6 (15.4)	68.2 (17.5)	37.6
Positive Temperament (26)	12.8 (3.7)	50.6	44.9 (17.6)	92.2 (21.1)	51.3
Exhibitionism (16)	10.3 (2.6)	35.4	35.2 (17.6)	55.3 (19.3)	36.4
Entitlement (16)	10.4 (1.6)	35.1	32.7 (10.9)	56.3 (13.2)	41.9
Detachment (18)	9.8 (2.8)	45.7	36.4 (14.1)	68.6 (15.6)	47.0
Disinhibition (35)	17.9 (3.6)	48.8	71.7 (23.6)	150.4 (29.7)	52.3
Impulsivity (19)	12.6 (3.4)	33.6	51.3 (19.4)	80.0 (17.3)	35.9
Propriety (20)	18.3 (2.0)	8.5	80.7 (20.5)	87.4 (19.2)	7.6
Workaholism (18)	12.3 (3.0)	31.9	49.9 (15.6)	71.6 (15.2)	30.4
Mean		**36.2**			**37.7**

Note. $n = 205$.

[a] Total time and time savings are based on computerized administration of all items.

Table 3.3: Time 2 Adaptive Item and Time Savings in the Computerized Groups.

	Items		Time (sec.)		
Scale (items)	Adaptive *M* (*SD*)	Savings (%)	Adaptive *M* (*SD*)	Total[a] *M* (*SD*)	Savings (%)
Negative Temperament (28)	14.5 (1.5)	48.3	43.0 (11.2)	85.2 (18.2)	49.5
Mistrust (19)	12.1 (3.9)	36.3	38.4 (15.8)	60.9 (14.1)	37.0
Manipulativeness (20)	12.5 (2.4)	37.6	41.1 (12.8)	71.5 (14.7)	42.5
Aggression (20)	12.6 (3.2)	36.8	34.5 (12.3)	54.4 (12.8)	36.7
Self-harm (16)	12.4 (3.3)	22.8	32.5 (10.9)	43.7 (8.9)	25.7
Eccentric Perceptions (15)	10.0 (2.9)	33.6	37.0 (14.5)	55.0 (14.7)	32.8
Dependency (18)	11.3 (3.2)	37.0	36.1 (13.3)	56.8 (13.8)	36.4
Positive Temperament (26)	12.5 (3.3)	51.8	37.1 (14.2)	77.1 (18.7)	51.9
Exhibitionism (16)	10.6 (2.6)	33.9	30.1 (18.0)	45.7 (18.1)	34.1
Entitlement (16)	10.4 (1.6)	35.3	25.6 (7.6)	44.5 (10.1)	42.4
Detachment (18)	9.5 (2.7)	47.2	29.5 (13.8)	57.1 (13.2)	48.3
Disinhibition (35)	17.6 (3.4)	49.6	56.2 (17.3)	120.9 (24.5)	53.5
Impulsivity (19)	12.7 (3.3)	33.4	41.3 (15.9)	65.3 (15.9)	36.8
Propriety (20)	17.2 (2.9)	14.2	60.1 (18.1)	69.7 (16.9)	13.8
Workaholism (18)	12.0 (2.9)	33.3	40.3 (14.2)	59.5 (13.4)	32.3
Mean		**36.7**			**38.2**

Note. $n = 202$.

[a] Total time and time savings are based on computerized administration of all items.

Table 3.4: Percentage of Termination Types in Computerized Group Participants.

	Time 1 (*n* = 205)			Time 2 (*n* = 202)		
Scale	SEM	Info	Max	SEM	Info	Max
Negative Temperament	91	9	0	89	11	0
Mistrust	63	30	7	60	32	8
Manipulativeness	54	46	0	52	48	0
Aggression	56	43	1	45	55	0
Self-harm	47	53	0	44	56	0
Eccentric Perceptions	48	40	12	42	49	9
Dependency	53	47	0	48	52	0
Positive Temperament	63	37	0	59	41	0
Exhibitionism	69	31	0	62	38	0
Entitlement	62	38	0	59	41	0
Detachment	47	53	0	51	49	0
Disinhibition	78	22	0	77	23	0
Impulsivity	46	54	0	46	52	2
Propriety	0	86	14	0	90	10
Workaholism	57	43	0	51	49	0

Note. SEM = test terminated when standard error of measurement dropped below 0.40, Info = test terminated when conditional item information dropped below 0.10, Max = all items administered.

Table 3.5: Frequency Distribution: Number of Items Adaptively Administered in Time 1 Computerized Groups (Combined) by Scale.

	Number of Participants														
# of Items	NT	MS	MN	AG	SH	EP	DP	PT	EX	EN	DT	DS	IM	PR	WK
7						56			19		58		15		
8		83			73	49	60		53		50		14		17
9		4	14		4	7	13		27	91	3		10	1	34
10		22	4	93	6	10	32	115	23	42	6		29	2	27
11		10	66	20	1	13	19	5	27	15	39		25	2	21
12		6	30	15	9	14	21	4	7	19	5		13	1	19
13		11	15	10	3	17	8	5	8	35	20		9		8
14	184	10	16	3		15	8	15	12	3	6		20	4	17
15	1	2	10	8	109	24	11	6	29		11	95	6	14	18
16	7	8	25	24			20	10			6	23	37	6	19
17	3	25	8	6			13	10			1	9	6	7	25
18	3	10	16	10				5				7	20	12	
19		14	1	14				17				5	1	127	
20	6			2				3				6		29	
21	1							10				5			
22												18			
23												19			
24												7			
25												6			
26												5			

Note. n = 205. NT = Negative Temperament, MS = Mistrust, MN = Manipulativeness, AG = Aggression, SH = Self-harm, EP = Eccentric Perceptions, DP = Dependency, PT = Positive Temperament, EX = Exhibitionism, EN = Entitlement, DT = Detachment, DS = Disinhibition, IM = Impulsivity, PR = Propriety, WK = Workaholism.

Table 3.6: Frequency Distribution: Number of Items Adaptively Administered in Time 2 Computerized Groups (Combined) by Scale.

	Number of Participants														
# of Items	NT	MS	MN	AG	SH	EP	DP	PT	EX	EN	DT	DS	IM	PR	WK
7						64			18		57		8		
8		63			66	36	69		45		58		21		16
9		8	17		5	10	9		19	90	6		11	5	38
10		29	1	100	1	13	26	106	28	44	5		27	1	27
11		10	81	11	3	7	14	14	18	19	36		19	13	23
12		7	33	9	8	13	13	2	19	8	5		14	2	21
13		10	16	19	1	21	9	4	18	38	15		7		8
14	180	15	6	2		20	5	26	15	3	6		35	9	21
15	1	4	10	9	118	18	24	6	22		6	100	4	27	12
16	2	8	24	17			19	15			8	16	30	9	11
17	6	25	7	4			14	10				15	7	10	25
18	5	7	6	16								6	15	15	
19		16	1	15				12				8	4	90	
20	4							1				7		21	
21	4							5				7			
22												15			
23												12			
24								1				5			
25												7			
26												3			
27												1			

Note. n = 202. NT = Negative Temperament, MS = Mistrust, MN = Manipulativeness, AG = Aggression, SH = Self-harm, EP = Eccentric Perceptions, DP = Dependency, PT = Positive Temperament, EX = Exhibitionism, EN = Entitlement, DT = Detachment, DS = Disinhibition, IM = Impulsivity, PR = Propriety, WK = Workaholism.

Table 3.7: Summary of Repeated Measures ANOVA Tests.

	Group				Sex				Time				Group*Sex				Group*Time			
Scale	RS	ES	FT	AT	RS	ES	FT	AT	RS	ES	FT	AT	RS	ES	FT	AT	RS	ES	FT	AT
Negative Temperament	-	-	-	-	**Y**	**Y**	-	-	**Y**	**Y**	**Y**	**Y**	-	-	-	-	-	-	-	-
Mistrust	-	-	-	-	-	-	-	-	**Y**	-	**Y**	**Y**	-	-	-	-	-	-	-	-
Manipulativeness	-	-	-	-	**Y**	**Y**	**Y**	**Y**	-	-	-	-	-	-	-	-	-	-	-	-
Aggression	**Y**	-	-	-	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	-	-	-	-	-	-
Self-harm	-	**Y**	**Y**	**Y**	-	-	-	-	**Y**	**Y**	**Y**	**Y**	-	-	-	-	-	**Y**	**Y**	**Y**
Eccentric Perceptions	-	-	-	-	-	-	-	-	**Y**	**Y**	**Y**	**Y**	-	-	-	-	-	-	-	-
Dependency	-	-	-	-	**Y**	**Y**	**Y**	**Y**	-	**Y**	-	**Y**	-	-	-	-	-	-	**Y**	**Y**
Positive Temperament	-	-	-	-	-	-	-	-	**Y**	**Y**	**Y**	**Y**	-	-	-	-	**Y**	**Y**	-	-
Exhibitionism	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Entitlement	-	-	-	-	-	-	-	-	**Y**	-	**Y**	-	-	-	-	-	-	-	-	-
Detachment	-	-	-	-	**Y**	**Y**	**Y**	**Y**	-	-	-	-	-	-	-	-	-	-	-	-
Disinhibition	-	-	-	-	**Y**	**Y**	**Y**	**Y**	-	-	-	-	-	-	-	-	-	-	-	-
Impulsivity	-	-	-	-	**Y**	**Y**	**Y**	**Y**	-	-	-	-	-	-	-	-	-	-	-	-
Propriety	-	-	-	-	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	-	-	-	-	**Y**	**Y**	**Y**	**Y**
Workaholism	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**Y**	-	**Y**

Note. **Y** = effect significant ($p < .01$), RS = traditional raw scores for both modes, ES = estimated true scores for computerized mode (compared with P&P raw scores), FT = full-scale thetas for both modes, AT = adaptively derived thetas in computerized mode (compared with P&P full-scale thetas).

Table 3.8: Group*Time Cell Means for Significant Effects Only.

Scale Method	P-P T1(A)	P-P T2(B)	C-P T1(C)	C-P T2(D)	P-C T1(E)	P-C T2(F)	C-C T1(G)	C-C T2(H)	Significant Differences*
Dependency									
Full-scale thetas	-.03	-.03	.23	.01	.04	.05	.21	.15	D < C
Adaptively derived thetas[a]	-.03	-.03	.25	.01	.04	.03	.24	.13	D < C
Positive Temperament									
Traditional raw scores	18.9	19.7	17.7	17.7	17.1	18.7	18.9	19.5	E < B, F
Estimated true scores[b]	18.9	19.7	17.7	17.7	17.1	18.6	19.0	19.4	E < B, F
Propriety									
Traditional raw scores	10.9	11.0	10.9	11.1	11.5	12.7	11.2	12.0	E < F
Estimated true scores[b]	10.9	11.0	11.1	11.1	11.5	12.5	11.5	12.1	E < F
Full-scale thetas	-.11	-.07	-.14	-.10	-.02	.23	-.04	.12	E < F
Adaptively derived thetas[a]	-.11	-.07	-.14	-.10	-.02	.23	-.03	.12	E < F
Workaholism									
Estimated true scores[b]	6.7	6.9	7.7	7.3	6.1	7.1	7.5	7.5	E < C, F
Adaptively derived thetas[a]	-.18	-.17	.06	-.07	-.32	-.11	-.03	-.03	E < C, F

Table 3.8—continued

Scale Method	P-P T1(A)	P-P T2(B)	C-P T1(C)	C-P T2(D)	P-C T1(E)	P-C T2(F)	C-C T1(G)	C-C T2(H)	Significant Differences*
Self-harm									
Estimated true scores[b]	1.4	1.3	2.8	1.7	2.7	2.9	2.6	2.4	A < C, E, F, G; B < C, E, F, G, H; D < C, F
Full-scale thetas	-.60	-.68	-.02	-.57	-.28	.03	-.04	-.11	A, B < C, E, F, G, H; D < C, F, G, H; E < F
Adaptively derived thetas[a]	-.60	-.68	-.01	-.57	-.28	.02	-.03	-.11	A, B, D < C, E, F, G, H; E < F

Note. Standard deviations appear in Tables B1 to B4. P-P = paper-and-pencil at Times 1 and 2 (n = 106), C-P = computerized at Time 1 and paper-and-pencil at Time 2 (n = 105), P-C = paper-and-pencil at Time 1 and computerized at Time 2 (n = 102), C-C = computerized at Times 1 and 2 (n = 100), T1 = Time 1, T2 = Time 2.

* ts(405) ranged from 3.48 to 14.75, $p < .01$ (Bonferroni-corrected).

[a] in computerized mode (compared with P&P full-scale thetas). [b] in computerized mode (compared with P&P traditional raw scores).

Table 3.9: Follow-up Test Means for Significant Effects Only.

	Time 1			Time 2			B-S Differences*		W-S Differences*	
Scale / Scoring Metric	PP	CF	CA	PP	CF	CA	Time 1	Time 2	Time 1	Time 2
Dependency										
Raw	5.4	6.2	6.1	5.4	5.8	5.5				CF > CA
Detachment										
Raw	4.5	4.3	4.5	4.2	4.2	4.7				CA > CF
Disinhibition										
Raw	13.5	12.5	12.2	13.2	13.1	12.4				CF > CA
Theta	.19	.14	.10	.18	.18	.13				CF > CA
Entitlement										
Raw	8.4	8.0	8.6	8.7	8.4	8.6			CA > CF	
Theta	.22	.21	.27	.34	.26	.28			CA > CF	
Negative Temperament										
Raw	13.7	14.1	14.3	13.1	13.3	13.9				CA > CF
Theta	-.08	-.03	-.02	-.18	-.15	-.10				CA > CF
Propriety										
Raw	11.2	11.0	11.3	11.1	12.3	12.3		CA, CF > PP	CA > CF	
Theta	-.07	-.09	-.09	-.09	.18	.18		CA, CF > PP		

Table 3.9—continued

Scale / Scoring Metric	Time 1			Time 2			B-S Differences*		W-S Differences*	
	PP	CF	CA	PP	CF	CA	Time 1	Time 2	Time 1	Time 2
Self-harm										
Raw	2.1	1.9	2.7	1.5	2.0	2.7	CA > PP	CA > PP	CA > CF	CA > CF
Theta	-.45	-.03	-.02	-.62	-.04	-.04	CA, CF > PP	CA, CF > PP		
Workaholism										
Raw	6.4	7.3	7.6	7.1	7.1	7.3	CA > PP		CA > CF	
Theta	-.25	-.01	.02	-.12	-.10	-.07	CA > PP			

Note. Standard deviations were omitted for clarity but appear in Tables B5 to B8. PP = paper-and-pencil, CF = computerized full-scale, CA = computerized adaptive, B-S = between-subjects, W-S = within subjects. Time 1 *n*s = 208 and 205 for paper-and-pencil and computerized groups, respectively. Time 2 *n*s = 211 and 202 for paper-and-pencil and computerized groups, respectively.

* $p < .01$ (Bonferroni-corrected).

Table 3.10: Descriptive Statistics for SNAP Validity Scales by Testing Mode.

	Time 1			Time 2	
	P&P	Comp		P&P	Comp
Scale	*M* (*SD*)	*M* (*SD*)		*M* (*SD*)	*M* (*SD*)
Rare Virtues	2.1 (1.7)	2.2 (2.2)		2.2 (1.8)	2.2 (1.8)
Deviance	2.3 (1.9)	**2.9 (2.2)**	*	2.4 (2.0)	2.6 (2.1)
VRIN	5.1 (2.1)	5.3 (2.1)		4.4 (2.2)	4.8 (2.1)
TRIN	18.4 (2.4)	18.3 (2.7)		18.7 (2.3)	19.2 (2.4)
DRIN	19.1 (2.0)	19.3 (2.1)		19.2 (1.8)	18.9 (2.0)
Invalidity Index	14.4 (4.1)	**15.6 (4.4)**	*	13.5 (4.4)	13.7 (4.4)

Note. P&P = paper-and-pencil groups (*n*s = 208 and 211 at Times 1 and 2, respectively), Comp = computerized groups (*n*s = 205 and 202 at Times 1 and 2, respectively), VRIN = Variable Response Inconsistency, TRIN = True Response Inconsistency, DRIN = Desirable Response Inconsistency.

* $p < .01$. Higher means within time level appear in **boldface**.

Table 3.11: Test-retest Correlations, by Group.

	P-P		C-P		P-C		C-C		*M*	
Scale	R	θ	R	θ	R	θ	R	θ	R	θ
Negative Temperament	.93	.93	.79	.75	.88	.84	.85	.87	.87	.86
Mistrust	.83	.84	.91	.86	.90	.88	.92	.83	.89	.85
Manipulativeness	.90	.91	.83	.83	.91	.89	.87	.84	.88	.87
Aggression	.90	.86	.88	.86	.90	.85	.80	.73	.88	.83
Self-harm	.82	.76	.88	.78	.91	.77	.94	.77	.90	.77
Eccentric Perceptions	.89	.86	.84	.78	.86	.84	.82	.75	.85	.81
Dependency	.88	.87	.74	.71	.82	.80	.87	.87	.83	.82
Positive Temperament	.86	.86	.87	.86	.81	.82	.90	.89	.86	.86
Exhibitionism	.92	.91	.87	.82	.90	.87	.91	.87	.90	.87
Entitlement	.76	.75	.82	.82	.85	.80	.81	.81	.81	.80
Detachment	.89	.84	.87	.83	.93	.89	.92	.88	.91	.86
Disinhibition	.91	.91	.91	.88	.92	.90	.92	.88	.92	.89
Impulsivity	.90	.88	.86	.86	.83	.79	.88	.86	.87	.85
Propriety	.87	.87	.81	.81	.87	.86	.87	.86	.86	.85
Workaholism	.87	.87	.83	.82	.85	.81	.86	.85	.85	.84
Mean	**.88**	**.87**	**.85**	**.82**	**.88**	**.85**	**.88**	**.84**	**.87**	**.85**

Note. All correlations are significant ($p < .001$). R = raw score metric, θ = theta metric. Paper-and-pencil thetas are full-scale, computerized raw scores are traditional, and computerized thetas are adaptively-derived. P-P = paper-and-pencil at Times 1 and 2 ($n = 106$), C-P = computerized at Time 1 and paper-and-pencil at Time 2 ($n = 105$), P-C = paper-and-pencil at Time 1 and computerized at Time 2 ($n = 102$), C-C = computerized at Times 1 and 2 ($n = 100$).

Table 3.12: Raw-to-theta Correlations, by Testing Mode and Time.

	Comp		P&P	
Scale	Time 1	Time 2	Time 1	Time 2
Negative Temperament	.95	.96	.98	.98
Mistrust	.95	.96	.98	.98
Manipulativeness	.93	.94	.98	.98
Aggression	.94	.92	.97	.98
Self-harm	.90	.89	.91	.96
Eccentric Perceptions	.91	.94	.98	.98
Dependency	.95	.95	.98	.98
Positive Temperament	.93	.92	.97	.97
Exhibitionism	.94	.96	.99	.99
Entitlement	.93	.92	.97	.97
Detachment	.93	.94	.98	.97
Disinhibition	.91	.92	.99	.98
Impulsivity	.93	.91	.98	.98
Propriety	.98	.97	.99	.99
Workaholism	.95	.97	.98	.98
Mean	**.94**	**.94**	**.98**	**.98**

Note. All correlations are significant ($p < .001$). Paper-and-pencil thetas are full-scale, and computerized thetas are adaptively-derived. Comp = computerized groups (*n*s = 205 and 202 at Times 1 and 2, respectively), P&P = paper-and-pencil groups (*n*s = 208 and 211 at Times 1 and 2, respectively).

Table 3.13: Cronbach's Alpha Coefficients, by Testing Mode and Time.

	Comp		P&P	
Scale (items)	Time 1	Time 2	Time 1	Time 2
Negative Temperament (28)	.90	.92	.92	.92
Mistrust (19)	.87	.89	.83	.86
Manipulativeness (20)	.79	.82	.80	.84
Aggression (20)	.85	.85	.88	.89
Self-harm (16)	.82	.84	.83	.85
Eccentric Perceptions (15)	.78	.84	.80	.84
Dependency (18)	.84	.84	.82	.84
Positive Temperament (26)	.89	.89	.88	.89
Exhibitionism (16)	.83	.86	.84	.86
Entitlement (16)	.74	.78	.78	.81
Detachment (18)	.85	.87	.85	.85
Disinhibition (35)	.77	.80	.83	.82
Impulsivity (19)	.82	.82	.82	.84
Propriety (20)	.73	.81	.79	.82
Workaholism (18)	.83	.87	.82	.85
Mean	**.82**	**.85**	**.83**	**.85**

Note. Computerized group alphas calculated on all items within each scale. Comp = computerized groups (*n*s = 205 and 202 at Times 1 and 2, respectively), P&P = paper-and-pencil groups (*n*s = 208 and 211 at Times 1 and 2, respectively).

Table 3.14: Time 1 Factor Loadings of SNAP Scales on Three Principal Factors.

	NA				PA				DvC			
	P&P		Comp		P&P		Comp		P&P		Comp	
Scale	R	θ	R	θ	R	θ	R	θ	R	θ	R	θ
Negative Temperament	**.69**	**.71**	**.68**	**.70**	-.06	-.06	.14	.02	-.06	-.07	.03	.04
Mistrust	**.78**	**.77**	**.67**	**.61**	-.10	-.08	.12	-.08	.01	.03	.13	.19
Manipulativeness	**.47**	**.46**	**.37**	.32	.29	.30	.24	.08	**.57**	**.58**	**.64**	**.67**
Aggression	**.60**	**.62**	**.40**	**.35**	.04	.03	.12	.00	.18	.21	.30	**.37**
Self-harm	**.59**	**.56**	**.63**	**.51**	-.34	-.31	-.14	-.27	.16	.23	.19	.26
Eccentric Perceptions	**.43**	**.43**	.27	**.37**	.14	.13	.30	.17	.10	.09	.11	.02
Dependency	.22	.24	.28	**.36**	-.02	-.04	-.05	-.06	.00	.00	.01	-.01
Positive Temperament	-.28	-.31	**-.50**	-.34	**.71**	**.72**	**.65**	**.69**	-.16	-.15	-.09	-.09
Exhibitionism	.00	.00	-.20	-.12	**.63**	**.63**	**.58**	**.55**	.20	.23	.27	.29
Entitlement	.17	.10	.00	-.01	**.60**	**.60**	**.51**	**.42**	-.11	-.13	.11	.04
Detachment	**.44**	**.44**	**.63**	**.49**	**-.53**	**-.50**	-.21	**-.39**	-.09	-.09	-.03	-.03
Disinhibition	.28	.27	.17	.12	.18	.15	.05	.04	**.91**	**.91**	**.93**	**.88**
Impulsivity	.17	.17	.01	.03	.18	.16	.01	.05	**.79**	**.78**	**.76**	**.69**
Propriety	.13	.12	-.02	.14	.25	.25	**.36**	.31	**-.59**	**-.59**	**-.39**	**-.43**
Workaholism	.07	.06	.13	.13	.32	.34	**.45**	**.36**	**-.56**	**-.53**	-.22	-.15

Note. Correlations above ±.35 are presented in **boldface**. NA = Negative Affectivity, PA = Positive Affectivity, DvC = Disinhibition vs. Constraint, P&P = paper-and-pencil groups (*n* = 208), Comp = computerized groups (*n* = 205), R = based on traditional raw scores, θ = based on full-scale thetas for the P&P mode and adaptively-derived thetas for the computerized mode. Disinhibition includes overlapping variance with several other scales.

Table 3.15: Time 2 Factor Loadings of SNAP Scales on Three Principal Factors.

	NA				PA				DvC			
	P&P		Comp		P&P		Comp		P&P		Comp	
Scale	R	θ	R	θ	R	θ	R	θ	R	θ	R	θ
Negative Temperament	**.63**	**.65**	**.69**	**.70**	-.13	-.15	.05	-.06	-.04	-.05	-.06	-.08
Mistrust	**.71**	**.69**	**.80**	**.79**	.01	-.03	.07	-.06	.02	.01	.06	.07
Manipulativeness	**.60**	**.63**	**.39**	**.43**	**.42**	**.39**	.21	.12	**.38**	**.37**	**.66**	**.62**
Aggression	**.46**	**.48**	**.60**	**.55**	.12	.08	.10	-.07	.15	.15	.25	.25
Self-harm	**.56**	**.53**	**.60**	**.50**	-.27	-.29	-.23	-.30	.15	.22	.18	.23
Eccentric Perceptions	.32	.34	**.45**	**.47**	.17	.13	.29	.14	.06	.02	.11	.02
Dependency	**.38**	**.39**	.15	.22	-.08	-.13	-.06	-.03	-.03	-.03	-.01	.03
Positive Temperament	**-.35**	-.31	-.29	-.19	**.72**	**.74**	**.71**	**.74**	-.22	-.21	-.09	-.10
Exhibitionism	.01	.06	-.19	-.09	**.60**	**.62**	**.58**	**.60**	.09	.09	**.35**	.34
Entitlement	.16	.13	.04	-.04	**.57**	**.54**	**.58**	**.59**	-.26	-.29	.12	.02
Detachment	**.47**	**.46**	**.62**	**.55**	**-.49**	**-.52**	-.32	**-.38**	-.13	-.11	-.09	-.07
Disinhibition	**.41**	**.45**	.18	.25	.27	.27	.06	.07	**.82**	**.79**	**.95**	**.87**
Impulsivity	.22	.24	.01	.05	.17	.15	.03	.15	**.82**	**.80**	**.83**	**.76**
Propriety	.06	.07	.06	.07	.22	.21	**.38**	.28	**-.61**	**-.61**	**-.40**	**-.46**
Workaholism	.09	.08	.21	.22	.26	.24	**.48**	**.39**	**-.64**	**-.61**	**-.37**	**-.37**

Note. Correlations above ±.35 are presented in **boldface**. NA = Negative Affectivity, PA = Positive Affectivity, DvC = Disinhibition vs. Constraint, P&P = paper-and-pencil groups (n = 211), Comp = computerized groups (n = 202), R = based on traditional raw scores, θ = based on full-scale thetas for the P&P mode and adaptively-derived thetas for the computerized mode. Disinhibition includes overlapping variance with several other scales.

Table 3.16: Factor Convergence Coefficients across Administration Modes and Scoring Metrics.

	Time 1			Time 2			
Factor/Matrix Type	1	2	3	1	2	3	T1-T2
Negative Affectivity							
1. P&P traditional raw scores							.98
2. P&P full-scale thetas	.99			.99			.97
3. Comp traditional raw scores	.94	.95		.94	.93		.97
4. Comp adaptively-derived thetas	.96	.96	.98	.95	.95	.99	.97
Positive Affectivity							
1. P&P traditional raw scores							.98
2. P&P full-scale thetas	.99			.99			.98
3. Comp traditional raw scores	.91	.92		.93	.92		.99
4. Comp adaptively-derived thetas	.97	.97	.94	.94	.95	.96	.99
Disinhibition vs. Constraint							
1. P&P traditional raw scores							.98
2. P&P full-scale thetas	.99			.99			.98
3. Comp traditional raw scores	.95	.95		.91	.90		.99
4. Comp adaptively-derived thetas	.93	.94	.99	.93	.93	.99	.98

Note. T1-T2 = congruence coefficients computed between Time 1 and Time 2 factor loadings, P&P = paper-and-pencil, Comp = computerized.

Table 3.17: Time 1 Correlations between SNAP scales and Big Five Inventory in combined paper-and-pencil and computerized samples.

	Neuroticism				Extraversion				Conscientiousness				Agreeableness				Openness			
Scale	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ
Negative Temperament	**.83**	**.83**	**.76**	**.71**	-.23	-.26	-.25	-.25	-.22	-.22	-.16	-.13	**-.38**	**-.37**	-.27	-.24	-.10	-.09	-.15	-.11
Mistrust	**.44**	**.45**	**.37**	**.37**	-.31	-.31	-.23	-.25	-.21	-.22	-.25	-.22	**-.48**	**-.49**	**-.40**	**-.39**	.01	.03	-.12	-.15
Manipulativeness	.08	.09	.13	.17	.05	.08	-.07	-.09	**-.44**	**-.43**	**-.44**	**-.39**	**-.45**	**-.45**	**-.45**	**-.48**	-.01	.00	-.04	-.09
Aggression	**.43**	**.43**	.34	**.38**	-.09	-.11	-.11	-.12	-.21	-.24	-.14	-.14	**-.64**	**-.66**	**-.62**	**-.59**	-.04	-.03	-.07	-.11
Self-harm	**.42**	**.42**	**.45**	**.37**	**-.37**	**-.36**	-.30	-.29	**-.40**	**-.45**	-.33	**-.35**	-.32	**-.35**	-.23	-.28	.01	-.02	-.09	-.07
Eccentric Perceptions	.14	.14	.06	.10	-.04	-.04	.00	-.04	-.21	-.20	-.14	-.10	-.23	-.21	.09	.12	.28	.28	.27	.18
Dependency	.32	.34	.29	.34	-.07	-.10	-.15	-.19	-.20	-.22	-.15	-.13	.03	.04	.14	.12	-.28	-.29	-.32	-.31
Positive Temperament	-.34	**-.37**	**-.46**	**-.44**	**.66**	**.68**	**.63**	**.66**	**.37**	**.36**	**.35**	.27	.23	.23	**.35**	**.35**	**.36**	**.35**	**.41**	**.37**
Exhibitionism	-.11	-.10	-.16	-.11	**.55**	**.53**	**.54**	**.48**	-.08	-.10	.03	-.02	-.14	-.15	-.03	-.08	.17	.17	.26	.21
Entitlement	-.03	-.07	-.23	-.25	.25	.27	.24	.28	.20	.23	.08	.14	-.10	-.04	-.13	-.04	.19	.17	.23	.25
Detachment	.25	.26	.36	.31	**-.72**	**-.71**	**-.69**	**-.68**	-.16	-.16	-.16	-.13	-.31	-.33	**-.52**	**-.43**	-.06	-.04	-.06	-.04
Disinhibition	-.01	.00	.05	.05	.12	.10	.06	.07	**-.69**	**-.67**	**-.62**	**-.54**	**-.35**	**-.35**	-.26	-.27	-.01	.02	-.04	-.03
Impulsivity	.00	.04	-.02	.03	.18	.19	.16	.20	**-.63**	**-.63**	**-.53**	**-.44**	-.20	-.23	-.11	-.12	.05	.06	.03	.03
Propriety	.12	.12	-.03	-.04	.00	.00	.11	.09	**.47**	**.48**	.32	.30	.08	.09	.24	.27	-.04	-.04	-.09	-.09
Workaholism	.00	.02	.06	.03	.06	.08	.05	.10	**.55**	**.55**	**.44**	**.45**	.06	.06	-.12	-.10	.20	.18	.16	.17

Note. Correlations above ±.18 are significant ($p < .01$). Correlations above ±.35 are presented in **boldface**. PR = paper-and-pencil ($n = 204$) traditional raw scores, Pθ = paper-and-pencil full-scale thetas, CR = computerized ($n = 202$) traditional raw scores, Cθ = computerized adaptively derived thetas.

Table 3.18: Time 2 Correlations between SNAP scales and Big Five Inventory in combined paper-and-pencil and computerized samples.

	Neuroticism				Extraversion				Conscientiousness				Agreeableness				Openness			
Scale	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ
Negative Temperament	**.80**	**.79**	**.76**	**.74**	-.27	-.27	-.25	-.25	-.21	-.21	-.14	-.11	-.33	-.33	-.34	-.28	-.20	-.20	-.13	-.10
Mistrust	.34	**.36**	**.41**	**.42**	-.20	-.20	-.31	-.30	-.22	-.22	-.21	-.24	**-.47**	**-.47**	**-.45**	**-.47**	-.04	-.05	.01	-.04
Manipulativeness	.15	.14	.05	.09	-.04	-.03	.03	.02	**-.43**	**-.42**	**-.44**	**-.40**	**-.50**	**-.50**	**-.44**	**-.46**	-.02	-.03	.02	.00
Aggression	.34	**.37**	**.41**	**.45**	-.04	-.05	-.17	-.15	-.07	-.08	-.29	-.33	**-.59**	**-.61**	**-.59**	**-.55**	-.07	-.08	.06	.00
Self-harm	**.38**	**.38**	**.42**	**.35**	-.28	-.27	-.32	-.31	-.33	**-.37**	-.31	-.34	-.24	-.28	-.29	-.29	-.18	-.17	.09	.09
Eccentric Perceptions	.03	.05	.10	.11	-.05	-.08	-.09	-.14	-.19	-.17	-.10	-.08	-.01	.00	-.22	-.21	.32	.31	.23	.18
Dependency	.32	.33	.30	.33	-.21	-.25	-.01	-.04	-.20	-.20	-.13	-.12	-.02	-.03	.18	.18	-.29	-.29	-.31	-.29
Positive Temperament	**-.39**	**-.39**	**-.37**	**-.38**	**.59**	**.62**	**.60**	**.65**	.29	.30	**.36**	.29	.31	.30	.27	.29	**.44**	**.42**	.23	.22
Exhibitionism	-.17	-.17	-.13	-.11	**.49**	**.47**	**.59**	**.55**	-.12	-.11	.02	-.03	-.08	-.10	-.05	-.06	.20	.19	.21	.19
Entitlement	-.11	-.13	-.21	-.28	.21	.23	.24	.33	.16	.19	.17	.22	-.11	-.04	-.11	-.01	.20	.19	.18	.15
Detachment	.34	.34	.31	.32	**-.72**	**-.73**	**-.71**	**-.69**	-.07	-.07	-.15	-.16	**-.39**	**-.40**	**-.46**	**-.48**	-.21	-.21	.08	.05
Disinhibition	.04	.04	-.04	-.01	.03	.05	.14	.09	**-.64**	**-.61**	**-.63**	**-.55**	**-.38**	**-.40**	-.25	-.25	-.04	-.04	.01	.00
Impulsivity	-.03	.01	-.07	-.02	.11	.11	.22	.27	**-.63**	**-.59**	**-.57**	**-.48**	-.25	-.27	-.12	-.14	.02	.03	.13	.15
Propriety	.04	.04	-.03	-.05	.07	.08	.07	.07	**.40**	**.39**	**.36**	**.38**	.11	.13	.10	.14	-.04	-.03	-.10	-.10
Workaholism	.03	.04	-.04	-.03	.06	.07	.08	.11	**.54**	**.55**	**.50**	**.49**	.02	.02	-.01	.01	.08	.08	.22	.19

Note. Correlations above ±.18 are significant ($p < .01$). Correlations above ±.35 are presented in **boldface**. PR = paper-and-pencil ($n = 207$) traditional raw scores, Pθ = paper-and-pencil full-scale thetas, CR = computerized ($n = 199$) traditional raw scores, Cθ = computerized adaptively derived thetas.

Table 3.19: Time 1 Correlations between SNAP and Eysenck Personality Questionnaire-Revised in paper-and-pencil and computerized samples.

	Neuroticism				Extraversion				Psychoticism				Lie			
Scale	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ
Negative Temperament	**.83**	**.82**	**.76**	**.75**	-.22	-.23	-.24	-.25	.15	.14	.14	.13	-.24	-.23	-.16	-.13
Mistrust	**.53**	**.54**	**.52**	**.47**	-.22	-.21	-.25	-.26	.30	.31	**.37**	**.35**	-.22	-.23	-.16	-.15
Manipulativeness	.24	.26	.18	.20	.16	.18	.02	.00	**.54**	**.52**	**.53**	**.52**	**-.58**	**-.59**	**-.44**	**-.44**
Aggression	**.38**	**.40**	.18	.23	-.07	-.07	-.08	-.09	**.37**	**.39**	**.36**	.32	-.26	-.26	-.24	-.23
Self-harm	**.43**	**.41**	**.50**	**.45**	-.31	-.29	-.33	-.30	.32	**.35**	**.35**	**.39**	-.10	-.13	-.18	-.19
Eccentric Perceptions	.34	.34	.31	.30	.04	.04	.04	.01	.22	.22	.24	.19	-.19	-.16	-.06	-.02
Dependency	**.40**	**.42**	.29	.34	-.06	-.08	-.20	-.23	-.04	-.04	-.04	-.02	-.18	-.18	-.13	-.16
Positive Temperament	-.24	-.28	-.29	-.31	**.66**	**.67**	**.71**	**.72**	-.16	-.17	-.18	-.17	-.03	.00	.10	.09
Exhibitionism	-.07	-.06	-.10	-.10	**.58**	**.56**	**.54**	**.49**	.10	.11	.14	.15	-.29	-.30	-.10	-.14
Entitlement	.06	.01	-.06	-.12	.31	.34	.18	.19	.00	-.04	.17	.15	-.11	-.07	-.04	-.03
Detachment	.23	.23	.29	.27	**-.67**	**-.64**	**-.66**	**-.64**	.19	.18	.32	.27	.01	.01	-.08	-.06
Disinhibition	.13	.15	.11	.13	.25	.24	.16	.15	**.69**	**.68**	**.65**	**.63**	**-.50**	**-.50**	**-.41**	**-.37**
Impulsivity	.07	.09	.03	.05	.31	.29	.22	.22	**.56**	**.56**	**.51**	**.52**	**-.35**	-.34	-.17	-.20
Propriety	.09	.09	.07	.05	.07	.07	.15	.13	**-.51**	**-.52**	**-.50**	**-.50**	.15	.15	.27	.26
Workaholism	-.06	-.04	.14	.10	.06	.07	.02	.09	-.24	-.26	.00	-.03	.14	.14	.11	.07

Note. Correlations above ±.18 are significant ($p < .01$). Correlations above ±.35 are presented in **boldface**. PR = paper-and-pencil (n = 207) traditional raw scores, Pθ = paper-and-pencil full-scale thetas, CR = computerized (n = 203) traditional raw scores, Cθ = computerized adaptively derived thetas.

Table 3.20: Time 2 Correlations between SNAP scales and Eysenck Personality Questionnaire-Revised in paper-and-pencil and computerized samples.

	Neuroticism				Extraversion				Psychoticism				Lie			
Scale	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ	PR	Pθ	CR	Cθ
Negative Temperament	**.86**	**.86**	**.84**	**.81**	-.25	-.24	-.22	-.22	.21	.21	.11	.08	-.28	-.29	-.12	-.07
Mistrust	**.42**	**.42**	**.62**	**.61**	-.19	-.17	-.24	-.25	**.45**	**.43**	.31	.31	-.24	-.23	-.16	-.17
Manipulativeness	.25	.27	.24	.25	.11	.12	.05	.02	**.57**	**.55**	**.51**	**.51**	**-.54**	**-.54**	**-.52**	**-.50**
Aggression	.20	.25	**.42**	**.44**	.00	-.02	-.10	-.11	**.40**	**.38**	**.36**	.31	-.24	-.23	-.22	-.24
Self-harm	**.40**	**.41**	**.45**	**.38**	-.28	-.28	-.31	-.30	.33	**.38**	**.37**	**.37**	-.21	-.23	-.04	-.05
Eccentric Perceptions	.29	.29	**.35**	.32	.02	.01	.01	.00	.20	.19	.31	.25	-.10	-.10	-.15	-.17
Dependency	**.42**	**.44**	.30	.34	-.15	-.18	-.12	-.13	.01	.03	-.12	-.09	-.10	-.10	-.08	-.12
Positive Temperament	-.31	-.31	-.25	-.28	**.70**	**.70**	**.71**	**.71**	-.24	-.24	-.15	-.13	.06	.08	.06	.05
Exhibitionism	-.06	-.06	-.09	-.08	**.54**	**.53**	**.62**	**.58**	.09	.09	.17	.19	-.17	-.18	-.22	-.19
Entitlement	-.02	-.04	-.03	-.13	.25	.26	.28	.32	.05	-.01	.11	.06	-.09	-.04	-.11	-.09
Detachment	.27	.27	.32	.32	**-.72**	**-.70**	**-.67**	**-.61**	.22	.25	.24	.22	-.02	-.04	.02	.00
Disinhibition	.11	.13	.12	.13	.22	.23	.22	.16	**.73**	**.73**	**.69**	**.67**	**-.46**	**-.45**	**-.43**	**-.41**
Impulsivity	.02	.03	.02	.05	.25	.25	.30	.28	**.64**	**.64**	**.58**	**.53**	-.31	-.30	-.22	-.29
Propriety	.15	.14	.06	.02	.08	.08	.18	.16	**-.50**	**-.50**	**-.52**	**-.55**	.15	.14	.17	.18
Workaholism	.04	.04	.06	.05	.04	.04	.07	.10	-.24	-.25	-.12	-.11	.07	.06	.21	.20

Note. Correlations above ±.18 are significant ($p < .01$). Correlations above ±.35 are presented in **boldface**. PR = paper-and-pencil (n = 208) traditional raw scores, Pθ = paper-and-pencil full-scale thetas, CR = computerized (n = 208) traditional raw scores, Cθ = computerized adaptively derived thetas.

Table 3.21: Fit Indices Testing Correlational Similarity between SNAP-CAT Scales and Validity Measures.

	Time 1			Time 2		
Scale	χ^2	NFI	SRMR	χ^2	NFI	SRMR
Negative Temperament	20.753	.979	.048	14.510	.986	.041
Mistrust	21.419	.979	.026	30.153*	.972	.033
Manipulativeness	17.768	.983	.034	12.182	.989	.021
Aggression	13.878	.987	.028	41.574*	.961	.041
Self-harm	20.899	.979	.022	24.885	.975	.035
Eccentric Perceptions	33.696*	.968	.035	22.377	.979	.025
Dependency	14.809	.986	.021	26.911	.974	.031
Positive Temperament	15.260	.985	.029	22.912*	.978	.032
Exhibitionism	14.696	.985	.028	15.496	.985	.024
Entitlement	31.397	.969	.030	9.172	.991	.021
Detachment	13.372	.987	.030	30.713	.972	.041
Disinhibition	16.673	.985	.040	15.378	.986	.032
Impulsivity	20.434	.981	.037	17.816	.984	.032
Propriety	27.515	.974	.031	11.705	.990	.019
Workaholism	20.769	.981	.031	16.187	.985	.021

Note. All χ^2 tests conducted with 27 degrees of freedom. NFI = Bentler-Bonett Normed Fit Index, SRMR = Standardized Root Mean Squared Residual.

*Lagrange Multiplier test suggested constraint releases to significantly improve fit.

Table 3.22: Time 1 Pre- and Post-SNAP PANAS-X Descriptive Statistics with ANCOVA Results.

	P&P (n = 206)		Comp (n = 201)		
Scale	Pre-SNAP *M* (*SD*)	Post-SNAP *M* (*SD*)	Pre-SNAP *M* (*SD*)	Post-SNAP *M* (*SD*)	$F(1, 404)$
Negative Affect	15.8 (6.5)	14.9 (5.9)	15.2 (4.9)	14.5 (4.9)	
Positive Affect	25.9 (7.5)	24.6 (8.1)	25.3 (6.9)	24.2 (8.1)	
Fear	8.8 (4.0)	8.4 (3.6)	8.6 (3.1)	8.0 (2.7)	
Hostility	8.8 (3.4)	8.5 (3.3)	8.8 (3.4)	8.5 (3.5)	
Guilt	9.3 (4.3)	8.7 (4.2)	8.3 (3.5)	**8.5 (3.8)**	7.48
Sadness	8.8 (4.3)	7.9 (3.9)	8.5 (4.1)	8.1 (4.1)	
Joviality	19.5 (7.3)	18.5 (7.7)	18.5 (6.6)	18.1 (7.2)	
Self-assurance	14.6 (4.8)	13.9 (5.1)	14.2 (4.8)	13.7 (5.3)	
Attentiveness	12.2 (3.0)	11.4 (3.3)	12.5 (3.0)	11.6 (3.4)	
Shyness	6.8 (2.8)	5.9 (2.7)	6.4 (2.4)	5.7 (2.3)	
Fatigue	12.1 (4.1)	11.1 (4.0)	12.4 (4.1)	11.9 (4.4)	
Serenity	10.2 (2.4)	9.2 (2.7)	10.5 (2.4)	10.0 (2.7)	
Surprise	4.1 (1.6)	4.3 (2.0)	3.9 (1.5)	4.4 (2.1)	

Note. Only significant *F*-test results are presented ($p < .01$). Significantly higher post-SNAP means (relative to pre-SNAP means) are presented in **boldface**. P&P = paper-and-pencil groups, Comp = computerized groups.

Table 3.23: Time 2 Pre- and Post-SNAP PANAS-X Descriptive Statistics with ANCOVA Results.

	P&P (n = 207)		Comp (n = 200)		
Scale	Pre-SNAP *M* (*SD*)	Post-SNAP *M* (*SD*)	Pre-SNAP *M* (*SD*)	Post-SNAP *M* (*SD*)	$F(1, 404)$
Negative Affect	14.5 (5.5)	14.1 (5.6)	15.2 (5.5)	14.8 (5.3)	
Positive Affect	24.7 (8.2)	23.4 (8.6)	25.7 (7.7)	24.0 (8.5)	
Fear	8.2 (3.3)	7.9 (3.1)	8.6 (3.2)	8.5 (3.1)	
Hostility	8.6 (3.6)	8.5 (3.8)	8.6 (3.5)	8.4 (3.3)	
Guilt	8.1 (3.2)	8.1 (3.6)	8.4 (3.7)	8.5 (3.8)	
Sadness	7.6 (3.4)	7.4 (3.4)	7.6 (3.6)	7.5 (3.6)	
Joviality	19.4 (7.6)	18.3 (7.8)	20.5 (7.8)	19.3 (8.1)	
Self-assurance	14.3 (5.3)	13.4 (5.4)	14.4 (5.0)	13.4 (5.2)	
Attentiveness	11.4 (3.2)	10.8 (3.6)	11.8 (3.1)	11.0 (3.5)	
Shyness	5.9 (2.5)	5.4 (2.4)	5.8 (2.3)	**5.7 (2.4)**	11.74
Fatigue	10.7 (4.2)	10.4 (4.3)	10.3 (4.0)	**10.9 (4.5)**	9.40
Serenity	9.7 (2.6)	9.0 (3.0)	9.8 (2.6)	9.3 (2.7)	
Surprise	3.9 (1.7)	4.0 (1.7)	4.2 (2.0)	4.2 (1.9)	

Note. Only significant *F*-test results are presented ($p < .01$). Significantly higher post-SNAP means (relative to pre-SNAP means) are presented in **boldface**. P&P = paper-and-pencil groups, Comp = computerized groups.

Figure 3.1: Total Time Comparisons by Testing Mode, Separately for Times 1 and 2.

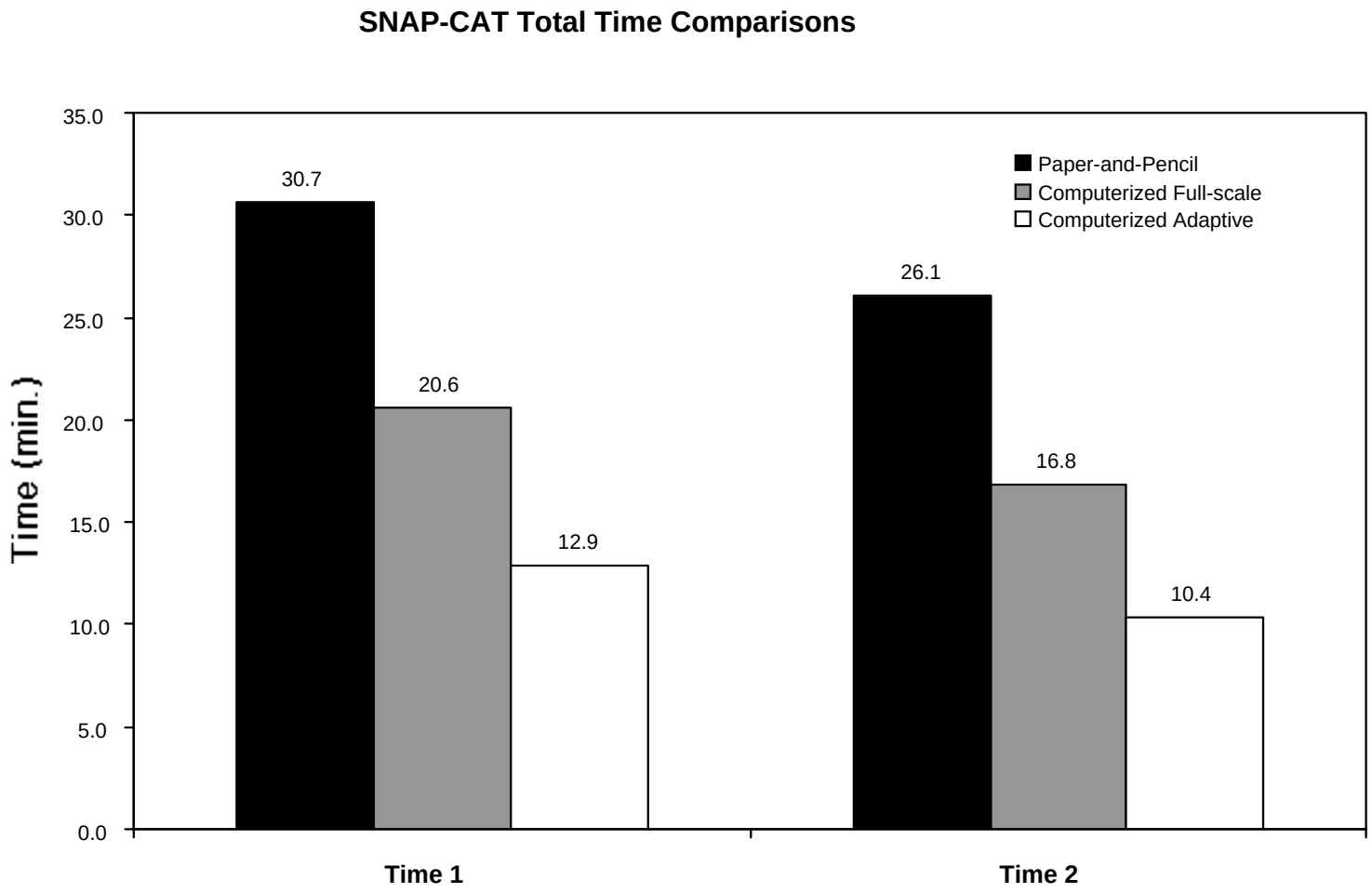


Figure 3.2: Reasons Why Some Participants Preferred SNAP-CAT.

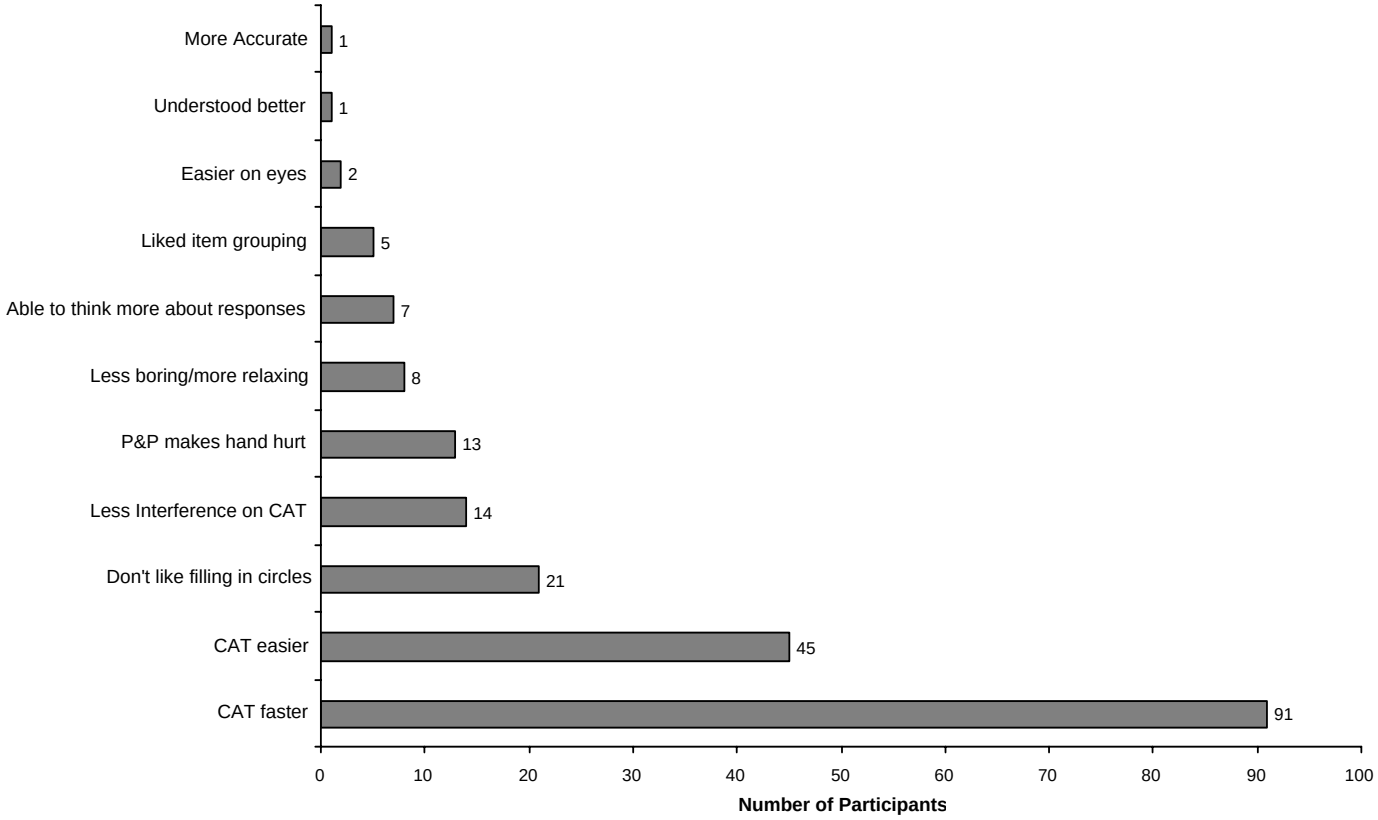
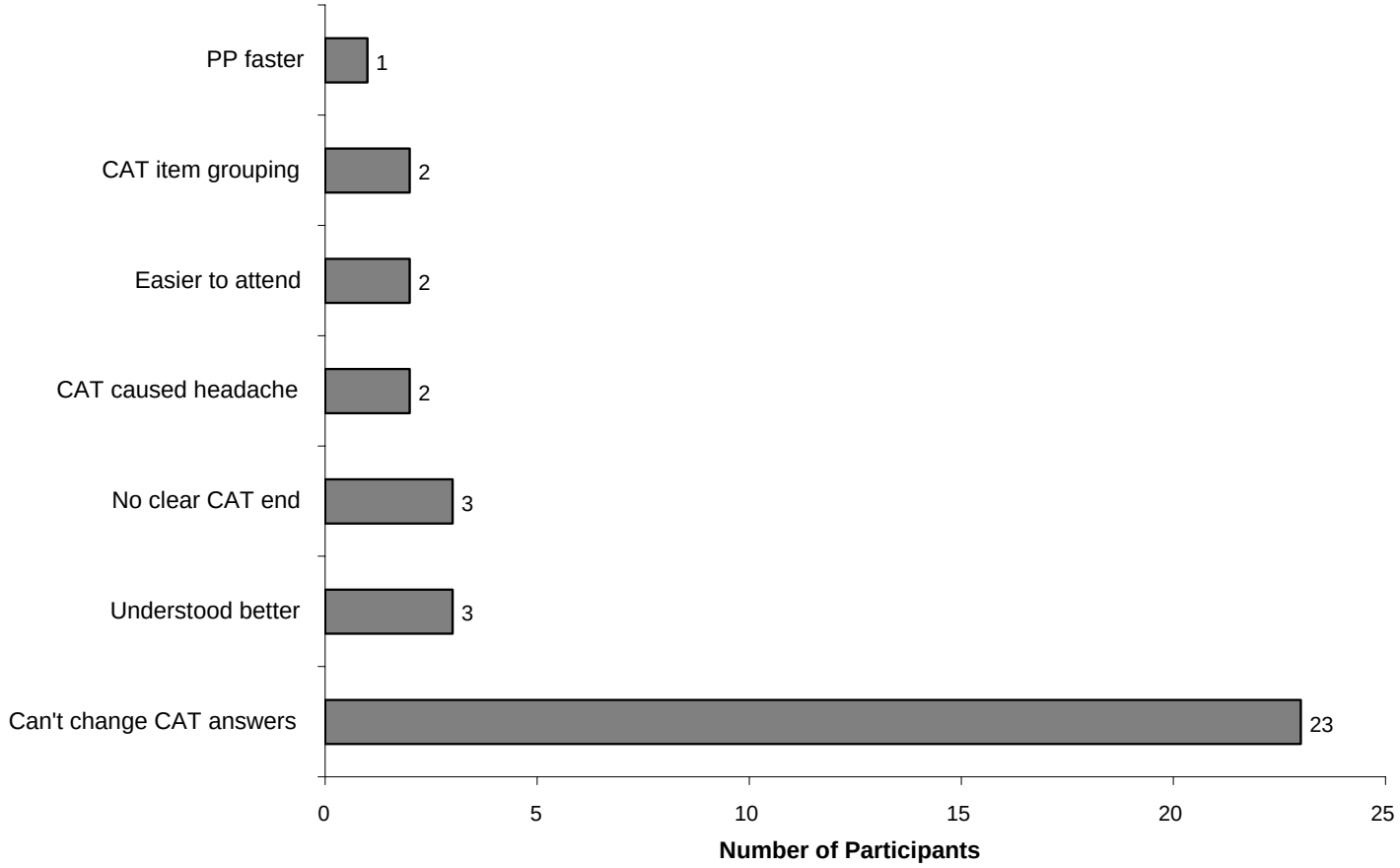


Figure 3.3: Reasons Why Some Participants Preferred SNAP-PP.



CHAPTER 4
DISSCUSION AND CONCLUSIONS

Summary of Findings

The overriding objective of this study was to develop and validate an IRT-based computerized adaptive version of the SNAP, a large multi-scale measure of traits relevant to personality disorder. Given the length of the traditional SNAP, the primary goal within that objective was to create a version of the SNAP that would yield comparable (or superior) psychometric (e.g., descriptive statistics, rank-ordering of scores, internal correlational structure, convergent/discriminant validity) and experiential (e.g., effects on examinee attitudes or moods, etc.) features while administering only a fraction of the items. In brief, the SNAP-CAT yielded significant item and time savings in comparison to both the traditional P&P method and full-scale administration on the computer. Impressively, CAT administration of the SNAP was 57.8% and 60.1% faster than the traditional P&P version, at Times 1 and 2, respectively. Even when controlling for computerized administration, time savings were 37.3% and 38.1%, respectively, and mean item savings across scales were 36.3% and 36.7%, respectively. These savings are consistent with those identified in the present simulation study as well as in previous IRT-Based CAT simulations (Kamakura & Balasubramanian, 1989; Reise & Henson, 2000; Waller, 1999; Waller & Reise, 1989). Also, these savings are better than those typically found in non-IRT CAT applications such as those created on the MMPI-2 and MMPI-A (Forbey et al., 2000; Roper et al., 1991, 1995).

These savings came with little cost to reliability or validity. Descriptive statistics, internal factor structure, and external correlational structure were largely comparable

across testing modes and methods of scoring. Moreover, participants overwhelmingly preferred the computerized version over the P&P version. Thus, the SNAP-CAT seems a viable alternative for SNAP administration. More detailed discussions of the study follow, with a particular emphasis on isolated results that did not conform to the above conclusions.

Development of the SNAP-CAT

The study began with the item calibration phase, during which all scales were subjected to IRT parameter calibration on a large sample of SNAP data from undergraduates, community-dwelling adults, and psychiatric patients. The unidimensionality of the scales was evaluated and verified, and then the two-parameter logistic model (Birnbaum, 1968), the most commonly used model for dichotomous personality items, was fit to all scales. Given the relatively small size of SNAP item pools (compared to those typically found in ability testing), chi-square item fit statistics were not appropriate (Mislevy & Bock, 1990); thus; item-fit was assessed by visual inspection of standardized posterior residuals (i.e., difference between actual item response probabilities and those predicted from the item characteristic curve). The BILOG manual provided an arbitrary threshold of 2.0 for interpretation of standardized posterior residuals, and 5 of 15 scales' mean residual values exceeded this level. However, no empirical studies could be found to validate this threshold. Thus, in order to maintain consistency with the P&P version of the SNAP at this stage, all items were kept in the pool. Longer scales would have permitted more sophisticated measures of item fit (e.g., Mislevy & Bock, 1990; Orlando & Thissen, 2000).

Interestingly, most published studies applying IRT to personality tests (e.g., Cooke & Michie, 1997; Rouse, Finger, & Butcher; Waller & Reise, 1989) have failed to report the goodness-of-fit of their selected IRT models. In a recent article (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001), published after data collection had begun on this study, the authors challenged the assumption apparently held by many personality researchers that IRT models will fit personality data. Chernyshenko et al. subjected item responses from the Sixteen Personality Factor Questionnaire-Fifth Edition (16PF; Conn & Rieke, 1994) and Goldberg's (1997) rating scale for the Five-Factor Model of personality to several common IRT models, including the two-parameter logistic model used in this study. Among many findings, their data suggested that the two- and three-parameter models fit some personality scales better than others, and they concluded that the application of IRT to personality should always be accompanied by assessment of fit. These data suggest that when the SNAP-CAT is reconstructed to account for the results of this study, other item response models (e.g., the three-parameter logistic model or Levine's (1984) maximum likelihood formula scoring model) should be fitted and assessed for relative goodness-of-fit.

Test information was assessed for each scale and revealed that SNAP scales provide the most psychometric discrimination power at the pathological end. This result was not unexpected given that the primary original purpose and design of the SNAP was to measure traits relevant to personality disorder. Looking at these data, it appears that more items, particularly at the non-pathological end, would need to be added to each scale for efficient SNAP-CAT application to individuals within the normal range of

personality. However, the SNAP-CAT item pool appears to be well-suited for use with disordered populations.

Computerized simulations were conducted primarily to assess and establish appropriate termination criteria for the SNAP-CAT. A combined termination algorithm was adopted in which scale administrations were terminated after either the SEM around theta dropped below 0.40 or the conditional item information associated with all remaining scale items was less than 0.10. Assuming a normal distribution underlying test scores, this strategy yielded significant simulated item savings across scales (mean = 36.5%), which was more than double the associated loss of psychometric information (mean = 18.1%).

However, Propriety was a clear outlier in the simulation study, yielding item savings of only 17.2%. The reasons for this were not fully appreciated until after the SNAP-CAT was constructed and tested. A posteriori investigations revealed that the SEM portion of the termination rule was never satisfied in the simulation study. Upon closer inspection, the peak of Propriety's test information curve (TIC) is at approximately 5.2; given that the SEM of a test is equal to the inverse square root of test information, the *lowest possible* SEM that could have been yielded by Propriety is 0.439. Consistent with Propriety's TIC, the lowest simulated SEM was 0.45, occurring at theta = -1.0. Thus, with the SEM never reaching the termination threshold, the CAT simulation continued administering items until all minimally informative items were completely exhausted. As will be discussed shortly, this same pattern occurred in the live testing phase. If this pattern had been noticed before SNAP-CAT construction, a modified termination rule would have been necessary for Propriety. Follow-up simulation analyses

indicated that raising the SEM threshold at 0.50 would have yielded simulated item savings of 35.3% for Propriety. Using a different termination rule likely would result in significantly fewer items being administered to each participant and, thus, might affect the psychometric properties of the scale (e.g., stability, internal structure, convergent/discriminant validity, etc.). However, the equivalence data from the other adaptively administered SNAP scales suggests that adaptively shortened scales do not result in significantly reduced reliability or validity. Nonetheless, future CAT studies of Propriety using alternative termination rules should continue to assess the effect of adaptive administration of its psychometric features.

Given that information is proportionally related to discrimination, relatively low TICs are indicative of scales consisting of poorly discriminating items. Indeed, Propriety (along with Disinhibition) yielded the lowest item discriminations of all scales (mean = 0.623), with all values less than 1.0. While having equivalently low item discriminations, Disinhibition, which a relatively broad and heterogeneous construct, did not yield the same termination problems as Propriety because it has 15 more items than Propriety (all things being equal, longer scales generate more information and smaller standard errors). Thus, future studies are needed to determine how and whether Propriety's item pool can be improved to include items that discriminate better or that add incremental information. One way to test this idea would be to combine Propriety with unique items from other known scales of the same or similar constructs (e.g., the Traditionalism scale of the MPQ). Doing so will, at minimum, increase the scale's length and may identify new items that discriminate more strongly than those in the current pool. One possible explanation for Propriety's relatively low item discriminations and overall test

information is that Propriety may somehow be fundamentally different from other SNAP traits. That is, Propriety, which was designed to assess a dimension of preference for traditional, conservative morality versus rejection of social rules and convention (Clark, 1993), may not be a personality trait as much as a set of attitudes or beliefs about the world. IRT parameterization may have different properties for attitude/belief data, but this question has not been adequately researched in the literature.

Construction of the SNAP-CAT also posed some challenges because the adaptive testing software, MicroCAT, had some unexpected (and unadvertised!) memory limitations that forced changes in the design. Initially, the goal was to have the program administer items adaptively for all scales simultaneously. More specifically, the program would start, for example, with an Aggression item and store the response and conditional theta estimate. However, instead of administering the next Aggression item, the program would then administer the first item from another scale, perhaps Mistrust, again keeping track of the item response and estimated theta. The cycle would continue until all scales had at least one item administered, and then the program would return to Aggression to administer the next adaptively chosen item. After all items were administered adaptively for all scales, the program would then administer all skipped items in order to provide full raw scores for all scales. Unfortunately, such a design created a huge, memory intensive application that completely overtaxed MicroCAT. Thus, the present design, in which scales were administered in their entirety before going on to the next, was engineered. The primary fear with the new design was that grouping items together by scale would create context effects (Knowles, 1988; Steinberg, 1994) that might appreciably alter scores or change scale properties such as internal consistency. This fear was not realized,

however, as scale-level descriptive statistics and internal consistencies were largely consistent across modes.

SNAP-CAT Validation

Atypical Response Patterns

As described above, the SNAP-CAT yielded significant item and time savings when compared to the SNAP-PP as well as to computerized administration of all test items. Item pattern analyses revealed that most scales behaved as predicted with respect to the termination criteria that were established. For most scales, a majority of individuals reached one of the termination criteria shortly after passing the item presentation minimums. A smaller group of scales produced a more uniform distribution of individuals terminating at different scale lengths. Finally, two scales, Propriety and Self-harm, yielded peculiar patterns that deserved further study. For Propriety, 82.0% and 62.4% of participants, at Times 1 and 2, respectively, terminated within three items of the maximum for the scale. A posteriori analyses indicated that the reason for this anomaly lies in two places: (a) the termination data found in Table 3.4, and (b) the simulation study results described above. Just as occurred in the simulation, the live testing data revealed that the SEM never dropped below the termination threshold of 0.40. Thus, participants were clustered near the maximum because Propriety was not capable, given its relatively low item discrimination values, to yield a SEM below the pre-determined termination threshold for the SNAP-CAT as a whole. Thus, as described above, future studies need to be conducted to identify new items for Propriety that discriminate better or, at minimum, add incremental information that would translate into lower standard errors.

Self-harm yielded a bimodal distribution of terminations, with 88.8% and 91.1% of participants terminating after either 8 (the minimum) or 15 (one short of the maximum) items at Times 1 and 2, respectively. This pattern was not predicted from the simulation study. A posteriori pattern analyses revealed that participants terminating after only 8 items scored significantly higher (raw score means = 4.3 and 4.9 at Times 1 and 2, respectively) than those exiting after completing 15 items (means = 0.30 and 0.32 at Times 1 and 2, respectively), $t(180) = 13.78$ and $t(182) = 15.40$, respectively. A mean difference of this magnitude is equivalent to more the 14 *T*-score points, or approximately 1.5 standard deviations. Thus, it appears that participants who endorsed more Self-harm items were able to meet the 0.40 threshold for SEM more easily than those who scored lower, which is understandable given that Self-harm provides very little psychometric information for examinees at the low end of the trait (see the Self-harm's TIC in Figure 2.2).

Self-harm can be broken down into two subordinate scales—Low Self-esteem (7 items) and Suicide Proneness (9 items)—but it was calibrated as a single scale for the SNAP-CAT. One might speculate that these two sources of overlapping variance led to the pattern observed here. To test this hypothesis, several steps were taken. First, Table 2.2 was consulted to determine whether Self-harm showed any evidence, however slight, of being multidimensional. Visual inspection of Self-harm's alpha, average inter-item correlation, ratio of eigenvalues, and non-linear factor analysis results revealed little evidence of multidimensionality. Although the RMSR value was higher for Self-harm than for all other scales, the remaining indices uniformly supported Self-harm's unidimensionality. Second, the subscales were scored using all item responses, and *t*

tests were conducted to see if one or both subscales contributed to the overall Self-harm mean difference between those terminating after 8 and 15 items. Again, results failed to confirm the hypothesis, with both Suicide Proneness and Low Self-esteem means significantly higher for the 8-item group than the 15-item group, *t*s(180) = 9.18 and 11.84, respectively. Finally, Self-harm item-fits (Table B5) were reexamined. Self-harm's mean RMSSPR was the highest of all scales, suggesting some failure of model fit.

Thus, while the reason for the bimodal response pattern is still somewhat unclear, it is possible that poor model fit contributed. Poor model fit, simply stated, can result from any number of factors, including violation of model assumptions (e.g., unidimensionality), selection of an inappropriate item response model (i.e., does not adequately account for actual response probabilities), or small item pools. Future recalibration and simulation testing are needed to examine these variables systematically and assess different ways to calibrate Self-harm items (e.g., with different IRT models or by calibrating the subscales separately).

Psychometric Equivalence

The mean-level analyses revealed largely comparable descriptive statistics across modes of administration and scoring methods. Most significant differences did not replicate across sessions and appeared random in nature. Moreover, in the comparisons between computerized adaptive and P&P raw scoring, which were most directly relevant to the question of equivalence across forms, only three scales—Propriety, Self-harm, and Workaholism—yielded significant differences, always in the direction of higher CAT scores, and only Self-harm's differences replicated across testing sessions. In addition, other than Self-harm, the largest differences were less than 3.2 *T*-score points, and most

were less than 1.2 *T*-score points, which would not likely be significant in a clinical setting. Nevertheless, in the context of the other problems identified above, the mean-level results suggest that IRT parameter calibration was less successful, for these three scales, Self-harm in particular.

Recall the pattern of means for Self-harm described in Chapter 3. At Time 1, Self-harm raw score means for PP, CF, and CA scoring were 2.1, 1.9, and 2.7, and mean thetas were -.45, -.03, and -.02, respectively. On the raw score metric, PP and CF scores were comparable, and CA estimated true scores were greater than both. With thetas, however, CA and CF scores were similar, and both were greater than PP thetas. The pattern was exactly the same at Time 2. This discrepancy suggests that something is happening with the conversion to and from the theta metric that significantly alters scores. Again, poor model fit likely contributed to this problem. Similar patterns were observed for Propriety and Workaholism, but these were not stable across sessions, suggesting that random error or sample fluctuations may be at least a partial explanation. Interestingly, when significant differences were identified across all scales, they were generally in the direction of more pathological responding in the computerized groups. This finding is consistent with two recent meta-analyses (Dwight & Feigelson, 2000; Richman, Kiesler, Weisband, & Drasgow, 1999) which reported that impression management tends to be lower with computerized personality test administration. However, Dwight & Feigelson also reported that the magnitude of this effect decreased markedly in more recent studies, which raises the possibility that mean differences associated with computerized administration have waned as participants have become more familiar or comfortable with computers in general.

As was expected, EAP theta estimation procedures resulted in slightly lower standard deviations for many scales when adaptively-derived estimated true scores (i.e., scores converted to the raw metric from the theta metric) were compared with traditional raw scores. This biased all scores slightly toward the mean, which could be problematic in clinical practice because extreme scores may appear less extreme when EAP methods are used for theta estimation. Other estimation methods such as maximum likelihood procedures would likely ameliorate this problem, but recall that such methods do not converge when scores are zero or "perfect," two real possibilities in any adaptive testing algorithm. One possible way to solve this problem would be to use EAP estimation until zero or perfect scores are no longer possible for a given examinee (i.e., at least one keyed and one non-keyed response have been made), and then use maximum likelihood methods from that point on. Such a hybrid approach likely would eliminate this problem for all but the most extreme examinees. Alternatively, a recently developed Bayesian estimation technique, *essentially unbiased maximum a posteriori* (EU-MAP; Wang, Hanson, & Lau, 1999) theta estimation, has been shown to reduce bias in adaptively-derived thetas. Thus, both of these methods should be considered and tested for the revised SNAP-CAT.

Rank-ordering of scores was also quite stable across sessions, methods of scoring, and modes of administration. Again, Self-harm was the lone consistent anomaly. In all four retest conditions, theta retest coefficients were significantly lower than those for raw scoring, *z*s ranged from 2.3 to 10.3, although they were never lower than .76. On average, the remaining scales yielded retest coefficients either at or very close to the P&P retest coefficients, with only a trend toward slightly lower theta correlations compared to

those of raw scores. Thus, of the SNAP's 15 scales, only one, namely Self-harm, consistently yielded less equivalent rank-ordering of scores and descriptive statistics across groups, and these differences appear to be related more to scoring method than to testing mode. Given that all previous studies of IRT-based CATs have been simulations, retest coefficients associated with adaptively administered personality scales have not appeared in the literature to date. However, two live testing studies conducted on the non-IRT CAT versions of the MMPI-2 (Roper et al., 1991, 1995) reported booklet-to-CAT retest correlations (mean *r*s = .75 and .73 for the basic scales in the 1991 and 1995 studies, respectively) that were somewhat lower than those found in the present study (mean *r*s = .85 and .88 for raw scoring and .82 and .85 for theta scoring in the C-P and P-C groups, respectively). However, the MMPI-2 booklet-to-CAT correlations must be interpreted in the context of equivalently low booklet retest coefficients in the same studies.

Structural Stability

Another important aspect of equivalence is similarity across internal and external covariance structures. Assessment of internal stability involved a series of exploratory factor analyses that that yielded consistent loadings for all scales across scoring methods and modes of presentation. Even Self-harm, which did not fare particularly well in other tests, correlated consistently and meaningfully across matrices. Congruence coefficients were computed to quantify overall structural similarity and revealed that within-mode convergence was slightly better than cross-mode convergence, although both within- and cross-mode convergence were within the acceptable range. Moreover, despite including a non-pure version of Disinhibition, the structure found here was highly consistent with

that identified in past studies (e.g., Clark, 1993; Clark et al., 1996). In particular, even the two scales that split the most across factors in these analyses—Manipulativeness and Detachment—showed similar evidence of splitting in the SNAP manual (Clark, 1993).

Regarding the similarity of convergent and discriminant validity patterns across modes and scoring methods, a unique CFA approach suggested that the overall correlational structure was highly consistent. The analyses revealed only four equality constraints whose release would lead to significantly improved model fit. However, none of these findings replicated across sessions, suggesting that they most likely were random occurrences.

Experiential Features

No consistent effects were identified across testing modes for state mood, a result that supports the experiential equivalence of the two test forms. However, participants clearly had different experiences of and reactions to the test forms. Eighty-seven percent favored the computerized version and stated that certain aspects of the SNAP-CAT were especially salient to them. In particular, most reported that the SNAP-CAT was faster (which was true) and easier to complete than the SNAP-PP. That participants enjoyed the computerized form better is consistent with a recent study (Vispoel, Boo, & Bleiler, 2001) examining equivalence between P&P and computerized versions of the Rosenberg Self-esteem Scale (SES; Rosenberg, 1965). Vispoel et al. found that participants preferred the computerized form and rated it as easier to read, more comfortable and enjoyable to complete, and less fatiguing than the P&P version. Interestingly, however, their data indicated that the computerized version took participants *longer* to complete, perhaps due to a multi-screen presentation format, than the P&P version, and yet they still

enjoyed it more. In the present study as well as in Vispoel et al. (2001), all participants were college students, who might be expected to be quite familiar and comfortable with computers. Generalizability of these findings to other populations (e.g., psychiatric patients or elderly persons) will need to be established empirically.

Conclusions and Future Directions

The results of the simulation and live testing studies suggest that the SNAP-CAT, except for several specific problems, largely centering on the Self-harm scale, is a comparable form of the traditional paper-and-pencil SNAP. The overall results indicate that the SNAP-CAT yielded reasonably comparable descriptive statistics, rank-ordering of scores, internal correlational structure, convergent/discriminant validity, and effects on state mood. Experientially, participants preferred the computerized form, but this preference did not appear to affect test scores. The SNAP-CAT achieved these comparable results despite administering only a fraction of the items and taking far less time than the paper-and-pencil version. In light of calls for more efficient measures of personality and psychopathology, the SNAP-CAT appears to be a viable alternative to traditional testing.

In the future, I plan to reconstruct the SNAP-CAT to incorporate new features and fix problems highlighted by the data obtained in the present study. First, greater attention will be paid to the IRT model used to calibrate item parameters. Multiple models, including the two- and three-parameter logistic models, will be fitted and more thoroughly assessed for fit. Second, based on mean-level and rank-order stability problems identified with thetas for Self-harm, recalibration that splits the scale into its two subscales may be necessary to apply IRT principles successfully. Next, given the

low information afforded by Propriety, the SEM threshold should be studied further and perhaps raised individually for Propriety in order to yield item and time savings more consistent with the other scales. Alternatively, Propriety's item pool could be expanded through item writing in order to identify items that better discriminate along the trait continuum. Finally, whereas the item grouping in the SNAP-CAT did not appear to affect item responses appreciably, consideration will be given to constructing the new SNAP-CAT will be constructed to present items in a less transparent manner in order to guard against any possible problems with local dependence of item responses.

APPENDIX A
CALIBRATION SAMPLE DETAILS

Table A1: Description of Samples that Form the Calibration Sample.

Sample	*n*	Source description
College Samples:		
1	561	Subset of original college norms collected by Clark (1993)
2	378	Data collected for self-peer agreement study (Ready, Clark, Watson, & Westerhouse, 2000)
3	292	Data collected for self-peer and self-parent agreement study (Harlan & Clark, 1999)
4	238	Validity study of personality and neuropsychological measures of executive functions (Ready, Stierman, & Paulsen, 2001)
5	197	Study of behavioral correlates of personality (Wu & Clark, 2002)
6	220	Undergraduate sample selected for higher levels of personality pathology, substance use, and antisocial practices (Casillas & Clark, 2000)
Community Samples:		
7	561	New normative data collected at multiple sites (Clark, Simms, Wu, & Casillas, 2002)
8	173	Sample of adoptee data provided by R. J. Cadoret, University of Iowa, Department of Psychiatry.
9	75	Non-patient controls of a chronic back pain study (Vittengl, Clark, Owen-Salters, & Gatchel, 1999)

Table A1—continued

Sample	*n*	Source description
Patient Samples:		
10	108	Subset of patient data presented in Clark (1993)
11	141	Cumulative sample of University of Iowa Seashore Psychology Clinic patients who completed the SNAP and various other measures (unpublished raw data)
12	106	Mixed in- and outpatient psychiatric patients (Reynolds & Clark, 2001)
13	125	Chronic back pain patient data (Vittengl, Clark, Owen-Salters, & Gatchel, 1999)
14	136	Recurrent depressed patient data (Clark, Vittengl, Jarrett, & Kraft, 2001)
15	162	Mixed in- and outpatient psychiatric patients (Ready & Clark, 2002; Ready, Watson, & Clark, in press)
16	522	Substance abusers participating in substance abuse treatment (Casillas, Clark, & Hall, 2001)

APPENDIX B
SNAP IRT CALIBRATION PARAMETERS

Table B1: Negative Temperament IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
NEG001	241	1.225	0.198	2.689
NEG002	244	1.123	-0.492	1.234
NEG003	245	1.064	0.031	2.202
NEG004	248	1.333	-0.328	1.459
NEG005	250	1.123	-0.720	1.259
NEG006	252	0.902	1.144	2.473
NEG007	259	1.422	-0.042	1.369
NEG008	260	1.096	0.843	2.189
NEG009	264	0.749	-0.504	1.836
NEG010	269	1.374	0.521	2.366
NEG011	273	0.967	0.209	3.115
NEG012	274	0.706	0.818	1.994
NEG013	277	1.143	0.391	0.917
NEG014	281	0.893	-0.939	2.386
NEG015	288	0.812	-1.005	1.377
NEG016	290	0.858	-0.581	1.233
NEG017	294	0.905	0.304	2.126
NEG018	298	0.811	-0.362	0.990
NEG019	301	1.086	-0.257	2.692
NEG020	309	0.826	-0.471	2.408
NEG021	311	0.922	0.140	1.049
NEG022	312	0.652	0.752	1.959
NEG023	316	0.918	0.125	2.483
NEG024	320	1.035	0.005	1.436
NEG025	323	0.805	-1.362	1.637
NEG026	325	1.027	-0.458	1.864
NEG027	331	0.491	-0.475	2.264
NEG028	333	1.095	0.575	2.477
M		**0.977**	**-0.069**	**1.910**
SD		**0.218**	**0.613**	**0.598**

Note. *N* = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B2: Mistrust IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
MIS001	8	0.549	1.498	1.273
MIS002	14	0.649	0.637	2.042
MIS003	27	0.857	0.898	1.186
MIS004	38	1.159	0.224	1.311
MIS005	53	0.945	1.183	1.196
MIS006	59	0.561	0.842	1.263
MIS007	67	1.189	0.437	1.043
MIS008	87	1.528	0.605	0.647
MIS009	106	0.763	0.419	0.517
MIS010	121	0.609	0.516	2.268
MIS011	133	0.756	0.421	0.994
MIS012	144	0.905	0.852	2.064
MIS013	147	0.675	0.753	3.060
MIS014	163	0.569	-0.760	2.483
MIS015	176	0.790	0.573	2.844
MIS016	188	0.944	-0.146	0.341
MIS017	205	0.944	0.713	2.702
MIS018	216	1.340	0.238	2.008
MIS019	224	1.448	0.919	0.929
M		**0.904**	**0.570**	**1.588**
SD		**0.302**	**0.486**	**0.821**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B3: Manipulativeness IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
MAN001	12	0.453	0.172	2.099
MAN002	25	0.668	1.119	1.104
MAN003	33	0.737	1.634	1.089
MAN004	46	0.540	1.923	1.682
MAN005	63	0.097	4.526	3.116
MAN006	76	0.733	0.231	1.456
MAN007	88	0.746	1.077	0.802
MAN008	91	1.116	0.706	1.912
MAN009	102	1.238	0.269	0.601
MAN010	104	0.588	1.247	1.256
MAN011	105	1.129	2.033	0.803
MAN012	119	1.064	1.482	0.513
MAN013	129	1.041	1.315	1.577
MAN014	149	0.887	2.048	1.501
MAN015	159	0.677	0.534	1.515
MAN016	166	0.785	1.506	2.067
MAN017	186	1.014	0.632	1.333
MAN018	200	0.719	1.978	2.339
MAN019	208	0.780	0.446	1.088
MAN020	219	0.523	0.283	0.514
M		**0.777**	**1.258**	**1.418**
SD		**0.274**	**1.006**	**0.664**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B4: Aggression IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
AGG001	2	1.166	1.746	0.807
AGG002	17	0.882	1.252	0.932
AGG003	21	0.913	1.556	1.117
AGG004	24	0.584	0.528	2.715
AGG005	31	1.150	1.243	1.221
AGG006	43	1.027	1.925	0.896
AGG007	56	0.992	1.484	1.844
AGG008	70	1.109	0.983	1.564
AGG009	80	0.695	1.134	3.259
AGG010	96	0.823	0.816	1.414
AGG011	109	0.782	0.141	1.971
AGG012	122	0.761	0.744	3.636
AGG013	141	1.210	1.067	2.084
AGG014	148	0.567	0.910	3.691
AGG015	153	0.930	0.616	1.465
AGG016	165	1.284	0.628	1.125
AGG017	181	0.776	1.151	2.752
AGG018	194	1.361	0.979	2.814
AGG019	212	0.922	1.698	1.267
AGG020	225	0.646	1.347	4.382
M		**0.929**	**1.097**	**2.048**
SD		**0.231**	**0.453**	**1.071**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B5: Self-harm IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
SHM001	13	1.365	0.713	3.342
SHM002	30	0.770	2.284	2.575
SHM003	65	1.121	1.207	3.744
SHM004	77	1.360	1.161	1.888
SHM005	98	1.802	0.705	3.446
SHM006	110	1.057	1.355	3.012
SHM007	139	1.227	0.840	3.066
SHM008	142	0.995	1.356	2.044
SHM009	157	1.516	1.167	4.642
SHM010	161	1.361	0.844	4.314
SHM011	172	0.712	2.258	3.588
SHM012	174	0.782	1.639	6.099
SHM013	190	1.326	0.720	3.096
SHM014	193	1.250	0.302	1.285
SHM015	206	0.764	-0.058	8.416
SHM016	237	1.010	0.802	2.335
M		**1.151**	**1.081**	**3.556**
SD		**0.307**	**0.623**	**1.739**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B6: Eccentric Perceptions IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
ECC001	7	0.549	1.748	2.220
ECC002	20	0.368	1.310	0.577
ECC003	42	1.159	1.081	1.899
ECC004	51	0.770	0.045	1.944
ECC005	60	0.662	0.321	0.787
ECC006	72	1.196	0.865	0.475
ECC007	86	0.636	-0.402	2.171
ECC008	103	0.859	0.247	1.465
ECC009	128	1.064	0.800	0.557
ECC010	150	0.638	0.167	1.702
ECC011	162	1.189	1.289	0.428
ECC012	178	1.126	1.269	1.299
ECC013	199	0.509	1.761	2.956
ECC014	213	0.982	1.129	0.720
ECC015	221	1.285	0.919	1.266
M		**0.866**	**0.837**	**1.364**
SD		**0.296**	**0.637**	**0.773**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B7: Dependency IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
DEP001	16	0.569	1.450	1.320
DEP002	26	0.998	1.594	2.477
DEP003	36	0.931	1.109	0.831
DEP004	40	1.350	1.275	0.882
DEP005	44	0.404	0.315	1.667
DEP006	52	0.922	1.536	1.877
DEP007	62	0.785	1.588	2.447
DEP008	75	0.755	0.306	0.871
DEP009	81	0.723	-0.483	0.475
DEP010	84	1.348	1.614	1.034
DEP011	95	1.043	0.544	2.862
DEP012	100	1.163	0.554	1.878
DEP013	112	0.450	2.165	1.427
DEP014	123	0.601	0.131	4.417
DEP015	136	0.696	0.083	1.749
DEP016	156	0.743	-0.284	1.083
DEP017	197	1.363	1.246	2.075
DEP018	230	0.535	2.562	0.666
M		**0.854**	**0.961**	**1.669**
SD		**0.306**	**0.850**	**0.964**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B8: Positive Temperament IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
POS001	239	0.629	-0.817	2.202
POS002	240	0.394	-1.209	2.876
POS003	242	1.130	-1.033	0.742
POS004	249	1.011	-0.368	2.046
POS005	253	0.760	-0.666	1.251
POS006	256	1.394	-0.603	0.704
POS007	257	1.065	-0.560	1.169
POS008	262	0.881	0.262	2.171
POS009	266	0.535	0.333	2.947
POS010	270	0.987	-0.588	2.026
POS011	276	0.710	-1.526	1.855
POS012	279	0.742	-0.504	1.618
POS013	283	0.583	-0.885	2.262
POS014	284	0.473	-1.099	2.357
POS015	291	1.160	-0.375	1.001
POS016	297	1.366	-0.289	1.091
POS017	300	1.639	-0.578	1.727
POS018	306	0.556	-2.063	1.761
POS019	308	0.583	-1.668	2.191
POS020	317	0.600	-0.561	1.077
POS021	319	0.691	-1.072	1.504
POS022	324	0.483	-1.382	2.521
POS023	328	0.914	0.317	3.286
POS024	336	0.725	-0.409	1.298
POS025	337	0.473	-0.320	2.068
POS026	341	1.233	-0.991	1.503
M		**0.835**	**-0.717**	**1.817**
SD		**0.331**	**0.583**	**0.678**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B9: Exhibitionism IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
EXH001	6	0.643	-1.030	1.461
EXH002	19	0.687	-0.628	1.516
EXH003	28	0.825	-0.486	2.203
EXH004	45	1.228	-0.478	1.432
EXH005	57	0.910	0.614	0.457
EXH006	69	1.200	0.490	2.034
EXH007	82	1.381	-0.041	1.287
EXH008	93	0.591	1.103	2.019
EXH009	108	0.690	1.862	2.032
EXH010	137	0.317	-1.659	2.674
EXH011	145	0.712	0.000	0.857
EXH012	169	0.819	0.898	0.632
EXH013	183	1.400	0.147	0.792
EXH014	195	0.815	0.801	0.519
EXH015	217	0.927	-0.058	1.500
EXH016	223	0.884	1.263	1.758
M		**0.877**	**0.175**	**1.448**
SD		**0.296**	**0.915**	**0.658**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B10: Entitlement IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
ENT001	49	1.125	-0.990	2.163
ENT002	71	0.456	-0.271	1.500
ENT003	83	0.460	0.291	3.004
ENT004	90	0.162	0.954	4.566
ENT005	113	1.148	-0.998	0.666
ENT006	120	0.882	-0.512	2.073
ENT007	125	0.293	0.714	4.679
ENT008	132	0.612	-0.689	2.346
ENT009	143	1.167	0.714	0.492
ENT010	155	1.402	0.269	0.918
ENT011	167	1.496	1.134	2.282
ENT012	171	1.222	0.020	2.177
ENT013	179	0.414	2.216	2.455
ENT014	203	0.464	0.807	3.288
ENT015	215	0.911	-0.724	3.298
ENT016	226	0.930	0.448	4.345
M		**0.821**	**0.211**	**2.516**
SD		**0.417**	**0.884**	**1.299**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B11: Detachment IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
DET001	9	1.394	0.563	1.839
DET002	11	0.626	1.580	1.410
DET003	22	1.886	1.209	1.005
DET004	35	1.251	0.632	3.831
DET005	48	0.940	0.158	1.426
DET006	61	0.692	0.352	3.026
DET007	74	0.983	1.408	0.937
DET008	92	0.632	-0.497	0.703
DET009	101	1.071	0.897	2.969
DET010	131	0.694	1.416	3.408
DET011	140	1.322	0.341	2.003
DET012	158	0.784	0.564	2.635
DET013	170	0.363	2.451	2.702
DET014	177	0.460	2.339	3.313
DET015	182	0.655	-0.744	1.876
DET016	196	1.564	0.570	0.812
DET017	207	0.894	0.096	1.541
DET018	218	1.159	0.588	2.079
M		**0.965**	**0.774**	**2.084**
SD		**0.404**	**0.849**	**0.969**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B12: Disinhibition IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
DIS001	4	0.347	-0.126	1.807
DIS002	25	0.492	1.386	1.975
DIS003	33	0.583	1.911	1.926
DIS004	34	0.614	-0.595	1.095
DIS005	37	0.909	0.321	0.722
DIS006	57	0.521	0.871	1.763
DIS007	88	0.838	1.002	2.099
DIS008	91	0.854	0.790	1.431
DIS009	99	0.501	0.814	1.948
DIS010	117	0.915	0.402	1.895
DIS011	124	0.504	0.715	2.032
DIS012	130	0.737	1.078	2.823
DIS013	154	0.479	1.261	2.033
DIS014	160	0.315	0.603	1.326
DIS015	164	0.398	0.934	1.539
DIS016	168	0.286	-0.571	1.545
DIS017	173	0.755	1.080	1.653
DIS018	186	0.796	0.705	1.459
DIS019	198	0.547	0.197	1.159
DIS020	200	0.565	2.339	2.107
DIS021	204	0.777	1.631	1.228
DIS022	208	0.732	0.458	1.178
DIS023	220	1.212	1.304	1.509
DIS024	232	0.682	0.497	0.803
DIS025	247	0.372	0.187	3.768
DIS026	251	0.479	0.963	2.864
DIS027	254	0.701	0.980	2.260
DIS028	261	0.648	0.878	0.882
DIS029	268	0.683	1.772	2.934
DIS030	272	0.436	1.973	1.014
DIS031	282	0.686	-0.227	1.766
DIS032	307	0.585	1.305	1.155
DIS033	318	0.847	0.376	2.098
DIS034	326	0.577	0.314	0.461
DIS035	329	0.436	0.767	0.874
M		**0.623**	**0.808**	**1.689**
SD		**0.200**	**0.675**	**0.699**

Note. *N* = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B13: Impulsivity IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
IMP001	4	0.461	-0.094	1.699
IMP002	37	0.861	0.337	0.525
IMP003	41	0.685	0.295	0.869
IMP004	58	1.002	0.760	0.487
IMP005	66	0.862	0.439	0.515
IMP006	89	1.186	1.161	0.368
IMP007	99	0.406	0.964	2.032
IMP008	114	0.400	0.875	0.795
IMP009	126	0.431	1.402	0.743
IMP010	130	1.082	0.887	0.545
IMP011	138	0.569	0.166	1.299
IMP012	146	1.411	0.749	0.954
IMP013	154	0.647	1.021	1.596
IMP014	173	1.285	0.842	0.772
IMP015	185	0.689	0.245	0.709
IMP016	189	0.518	0.202	2.657
IMP017	198	0.645	0.183	1.034
IMP018	228	0.300	0.485	1.222
IMP019	236	0.455	0.990	1.604
M		**0.731**	**0.627**	**1.075**
SD		**0.328**	**0.408**	**0.608**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B14: Propriety IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
PRP001	3	0.381	-1.222	1.054
PRP002	10	0.791	-1.360	0.544
PRP003	32	0.631	1.085	0.595
PRP004	34	0.660	0.572	0.758
PRP005	50	0.679	-0.702	0.827
PRP006	64	0.819	-0.876	0.788
PRP007	78	0.693	-1.433	0.377
PRP008	85	0.402	-0.552	0.533
PRP009	94	0.453	-0.526	0.279
PRP010	115	0.630	-0.786	1.130
PRP011	124	0.651	-0.594	1.598
PRP012	135	0.519	0.764	0.820
PRP013	151	0.751	-0.808	2.169
PRP014	160	0.884	-0.282	1.271
PRP015	184	0.540	-0.997	1.430
PRP016	201	0.777	-0.550	1.509
PRP017	202	0.494	-1.102	2.392
PRP018	210	0.712	0.024	2.288
PRP019	222	0.603	0.206	1.018
PRP020	231	0.388	2.608	0.666
M		**0.623**	**-0.327**	**1.102**
SD		**0.149**	**0.980**	**0.625**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

Table B15: Workaholism IRT Parameters

Item	SNAP Item	Item Discrimination	Item Difficulty	RMSSPR
WRK001	1	0.602	1.559	1.225
WRK002	18	1.558	0.984	0.749
WRK003	29	0.467	-1.539	2.810
WRK004	47	0.858	1.300	2.907
WRK005	54	0.682	-0.282	3.349
WRK006	68	0.599	-0.371	1.575
WRK007	79	0.713	2.420	1.348
WRK008	111	1.386	0.053	2.247
WRK009	116	0.718	-0.217	1.812
WRK010	127	0.557	0.698	2.872
WRK011	168	1.407	0.188	1.428
WRK012	180	0.582	0.244	3.004
WRK013	187	0.999	0.908	7.358
WRK014	192	0.882	-0.049	3.524
WRK015	211	0.797	-0.139	3.745
WRK016	214	0.633	-1.305	1.182
WRK017	227	0.412	1.831	2.004
WRK018	234	0.458	0.727	2.823
M		**0.795**	**0.389**	**2.553**
SD		**0.339**	**1.025**	**1.498**

Note. N = 3995. RMSSPR = Root Mean Square Standardized Posterior Residual.

APPENDIX C
SUPPLEMENTAL TABLES

Table C1: Descriptive Statistics (Traditional Raw Scores for All Modes) for Trait and Temperament Scales by Group and Time.

	P-P Group		C-P Group		P-C Group		C-C Group	
Scale	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)
Negative Temperament	13.0 (7.6)	12.5 (7.7)	13.9 (7.1)	13.6 (7.0)	14.5 (7.2)	13.7 (7.7)	14.2 (6.8)	12.8 (7.0)
Mistrust	6.0 (3.7)	5.5 (4.0)	6.2 (4.8)	6.2 (5.1)	7.0 (4.8)	6.7 (5.1)	6.5 (4.8)	6.2 (5.2)
Manipulativeness	6.4 (4.1)	6.5 (4.4)	5.8 (4.0)	5.8 (4.5)	6.1 (3.8)	6.1 (4.2)	5.6 (3.7)	5.6 (4.0)
Aggression	3.9 (4.0)	3.7 (4.2)	5.6 (4.9)	5.1 (5.0)	4.6 (4.8)	4.6 (4.5)	4.2 (3.2)	3.7 (3.3)
Self-harm	1.4 (2.1)	1.3 (2.1)	1.9 (2.5)	1.7 (2.8)	2.7 (3.2)	2.3 (2.9)	1.9 (2.7)	1.6 (2.6)
Eccentric Perceptions	5.1 (3.4)	4.4 (3.7)	4.9 (3.4)	4.4 (3.7)	5.7 (3.6)	5.4 (3.7)	5.3 (3.4)	5.0 (3.9)
Dependency	5.3 (3.8)	5.4 (4.2)	6.2 (4.0)	5.4 (3.8)	5.5 (3.8)	5.5 (3.9)	6.2 (4.1)	6.0 (4.1)
Positive Temperament	18.9 (5.3)	19.7 (5.3)	17.7 (6.0)	17.7 (6.3)	17.1 (6.2)	18.7 (5.5)	18.9 (5.8)	19.5 (6.0)
Exhibitionism	8.4 (4.0)	8.7 (4.2)	8.7 (3.8)	8.8 (4.3)	8.4 (4.1)	8.9 (4.4)	8.3 (4.1)	8.3 (4.3)
Entitlement	8.5 (3.5)	9.0 (3.7)	8.2 (3.4)	8.3 (3.8)	8.3 (3.5)	8.5 (3.8)	7.8 (3.3)	8.2 (3.4)
Detachment	4.1 (3.4)	3.9 (3.6)	4.6 (3.9)	4.4 (4.2)	4.9 (4.3)	4.8 (4.1)	4.1 (4.0)	3.7 (3.9)
Disinhibition	13.8 (7.3)	13.3 (6.9)	12.4 (6.0)	13.0 (6.8)	13.3 (6.3)	13.5 (6.8)	12.7 (6.2)	12.7 (6.3)
Impulsivity	7.0 (4.5)	6.9 (4.8)	6.4 (4.2)	6.6 (4.3)	6.9 (4.2)	6.5 (4.2)	6.9 (4.3)	6.8 (4.3)
Propriety	10.9 (4.3)	11.0 (4.7)	10.9 (3.5)	11.1 (4.1)	11.5 (3.8)	12.7 (4.0)	11.2 (3.8)	12.0 (4.3)
Workaholism	6.7 (3.6)	6.9 (4.0)	7.5 (3.8)	7.3 (4.2)	6.1 (4.0)	6.8 (4.2)	7.1 (4.1)	7.3 (4.7)

Note. P-P = paper-and-pencil at Times 1 and 2 (n = 106), C-P = computerized at Time 1 and paper-and-pencil at Time 2 (n = 105), P-C = paper-and-pencil at Time 1 and computerized at Time 2 (n = 102), C-C = computerized at Times 1 and 2 (n = 100).

Table C2: Descriptive Statistics (Estimated True Scores for Computerized Mode) for Trait and Temperament Scales by Group and Time.

	P-P Group		C-P Group		P-C Group		C-C Group	
Scale	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)
Negative Temperament	13.0 (7.6)	12.5 (7.7)	14.1 (6.6)	13.6 (7.0)	14.5 (7.2)	14.2 (7.1)	14.6 (6.3)	13.6 (6.7)
Mistrust	6.0 (3.7)	5.5 (4.0)	6.2 (4.3)	6.2 (5.1)	7.0 (4.8)	6.9 (4.8)	6.8 (4.4)	6.4 (4.6)
Manipulativeness	6.4 (4.1)	6.5 (4.4)	5.8 (3.4)	5.8 (4.5)	6.1 (3.8)	5.9 (3.5)	5.7 (3.2)	5.6 (3.4)
Aggression	3.9 (4.0)	3.7 (4.2)	5.4 (4.5)	5.1 (5.0)	4.6 (4.8)	4.3 (4.0)	4.0 (2.7)	3.6 (2.8)
Self-harm	1.4 (2.1)	1.3 (2.1)	2.8 (2.1)	1.7 (2.8)	2.7 (3.2)	2.9 (2.1)	2.6 (2.0)	2.4 (2.0)
Eccentric Perceptions	5.1 (3.4)	4.4 (3.7)	5.1 (3.3)	4.4 (3.7)	5.7 (3.6)	5.2 (3.3)	5.3 (3.1)	5.2 (3.7)
Dependency	5.3 (3.8)	5.4 (4.2)	6.1 (3.6)	5.4 (3.8)	5.5 (3.8)	5.3 (3.2)	6.1 (3.5)	5.6 (3.3)
Positive Temperament	18.9 (5.3)	19.7 (5.3)	17.7 (5.4)	17.7 (6.3)	17.1 (6.2)	18.6 (5.0)	19.0 (5.0)	19.4 (5.2)
Exhibitionism	8.4 (4.0)	8.7 (4.2)	8.7 (3.2)	8.8 (4.3)	8.4 (4.1)	8.8 (3.8)	8.1 (3.5)	8.2 (3.7)
Entitlement	8.5 (3.5)	9.0 (3.7)	8.8 (2.9)	8.3 (3.8)	8.3 (3.5)	8.6 (3.2)	8.4 (3.0)	8.7 (3.1)
Detachment	4.1 (3.4)	3.9 (3.6)	4.7 (3.3)	4.4 (4.2)	4.9 (4.3)	5.2 (3.8)	4.2 (3.5)	4.3 (3.5)
Disinhibition	13.8 (7.3)	13.3 (6.9)	12.3 (5.1)	13.0 (6.8)	13.3 (6.3)	12.7 (5.5)	12.2 (5.0)	12.2 (5.1)
Impulsivity	7.0 (4.5)	6.9 (4.8)	6.5 (3.5)	6.6 (4.3)	6.9 (4.2)	6.3 (3.6)	6.9 (3.8)	6.8 (3.6)
Propriety	10.9 (4.3)	11.0 (4.7)	11.1 (2.9)	11.1 (4.1)	11.5 (3.8)	12.5 (3.2)	11.5 (3.0)	12.1 (3.5)
Workaholism	6.7 (3.6)	6.9 (4.0)	7.7 (3.4)	7.3 (4.2)	6.1 (4.0)	7.1 (3.6)	7.5 (3.7)	7.5 (4.1)

Note. Paper-and-pencil scores are traditional raw scores. P-P = paper-and-pencil at Times 1 and 2 (n = 106), C-P = computerized at Time 1 and paper-and-pencil at Time 2 (n = 105), P-C = paper-and-pencil at Time 1 and computerized at Time 2 (n = 102), C-C = computerized at Times 1 and 2 (n = 100).

Table C3: Descriptive Statistics (Full-scale Thetas for All Modes) for Trait and Temperament Scales by Group and Time.

	P-P Group		C-P Group		P-C Group		C-C Group	
Scale	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)
Negative Temperament	-.17 (.97)	-.26 (1.0)	-.03 (.89)	-.11 (.90)	.01 (.89)	-.11 (.96)	-.04 (.84)	-.20 (.89)
Mistrust	-.14 (.75)	-.25 (.80)	-.11 (.91)	-.18 (.99)	.04 (.89)	-.03 (.98)	-.02 (.91)	-.11 (.98)
Manipulativeness	.21 (.88)	.24 (.94)	.16 (.86)	.10 (.99)	.17 (.83)	.18 (.90)	.13 (.81)	.07 (.88)
Aggression	-.21 (.85)	-.28 (.88)	.15 (.97)	-.01 (1.0)	-.05 (.96)	-.04 (.95)	-.06 (.71)	-.19 (.78)
Self-harm	-.60 (.59)	-.68 (.61)	-.02 (.47)	-.57 (.73)	-.28 (.72)	.03 (.51)	-.04 (.47)	-.11 (.48)
Eccentric Perceptions	.10 (.82)	-.10 (.95)	.12 (.86)	-.07 (.94)	.25 (.84)	.19 (.91)	.22 (.83)	.13 (.99)
Dependency	-.03 (.90)	-.03 (.99)	.23 (.90)	.01 (.89)	.04 (.87)	.05 (.88)	.21 (.94)	.15 (.92)
Positive Temperament	.29 (.87)	.45 (.89)	.11 (.93)	.13 (.98)	.03 (.96)	.28 (.88)	.34 (.95)	.44 (.98)
Exhibitionism	.25 (.93)	.34 (.97)	.37 (.83)	.37 (.99)	.28 (.96)	.39 (1.0)	.23 (.93)	.24 (.98)
Entitlement	.25 (.84)	.40 (.91)	.26 (.84)	.27 (.95)	.20 (.88)	.28 (.95)	.16 (.83)	.25 (.86)
Detachment	-.43 (.73)	-.44 (.80)	-.29 (.85)	-.36 (.93)	-.24 (.92)	-.23 (.91)	-.43 (.90)	-.49 (.93)
Disinhibition	.22 (.94)	.20 (.88)	.14 (.75)	.15 (.90)	.16 (.85)	.22 (.83)	.15 (.76)	.14 (.80)
Impulsivity	.00 (.90)	-.03 (.99)	-.10 (.89)	-.07 (.89)	-.01 (.90)	-.13 (.90)	.01 (.93)	-.02 (.91)
Propriety	-.11 (.89)	-.07 (.99)	-.14 (.73)	-.10 (.84)	-.02 (.79)	.23 (.85)	-.04 (.74)	.12 (.90)
Workaholism	-.18 (.83)	-.17 (.95)	.03 (.86)	-.07 (.96)	-.32 (.93)	-.14 (.97)	-.07 (.95)	-.06 (1.1)

Note. P-P = paper-and-pencil at Times 1 and 2 (n = 106), C-P = computerized at Time 1 and paper-and-pencil at Time 2 (n = 105), P-C = paper-and-pencil at Time 1 and computerized at Time 2 (n = 102), C-C = computerized at Times 1 and 2 (n = 100).

Table C4: Descriptive Statistics (Adaptively derived Thetas in Computerized Mode) for Trait and Temperament Scales by Group and Time.

	P-P Group		C-P Group		P-C Group		C-C Group	
Scale	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)	Time 1 *M* (*SD*)	Time 2 *M* (*SD*)
Negative Temperament	-.17 (.97)	-.26 (1.0)	-.02 (.87)	-.11 (.90)	.01 (.89)	-.06 (.94)	-.01 (.82)	-.14 (.89)
Mistrust	-.14 (.75)	-.25 (.80)	-.11 (.90)	-.18 (.99)	.04 (.89)	.01 (.98)	.01 (.91)	-.09 (.96)
Manipulativeness	.21 (.88)	.24 (.94)	.16 (.85)	.10 (.99)	.17 (.83)	.16 (.88)	.14 (.80)	.09 (.87)
Aggression	-.21 (.85)	-.28 (.88)	.16 (.96)	-.01 (1.0)	-.05 (.96)	-.11 (.96)	-.08 (.69)	-.22 (.74)
Self-harm	-.60 (.59)	-.68 (.61)	-.01 (.46)	-.57 (.73)	-.28 (.72)	.02 (.47)	-.03 (.44)	-.11 (.45)
Eccentric Perceptions	.10 (.82)	-.10 (.95)	.16 (.91)	-.07 (.94)	.25 (.84)	.19 (.91)	.21 (.88)	.12 (1.1)
Dependency	-.03 (.90)	-.03 (.99)	.25 (.92)	.01 (.89)	.04 (.87)	.03 (.86)	.24 (.92)	.13 (.90)
Positive Temperament	.29 (.87)	.45 (.89)	.10 (.90)	.13 (.98)	.03 (.96)	.26 (.85)	.35 (.88)	.44 (.93)
Exhibitionism	.25 (.93)	.34 (.97)	.37 (.82)	.37 (.99)	.28 (.96)	.40 (1.0)	.22 (.91)	.24 (.98)
Entitlement	.25 (.84)	.40 (.91)	.32 (.86)	.27 (.95)	.20 (.88)	.28 (.95)	.21 (.87)	.29 (.89)
Detachment	-.43 (.73)	-.44 (.80)	-.30 (.85)	-.36 (.93)	-.24 (.92)	-.19 (.92)	-.44 (.88)	-.44 (.92)
Disinhibition	.22 (.94)	.20 (.88)	.10 (.74)	.15 (.90)	.16 (.85)	.16 (.79)	.10 (.73)	.09 (.75)
Impulsivity	.00 (.90)	-.03 (.99)	-.08 (.85)	-.07 (.89)	-.01 (.90)	-.13 (.89)	.02 (.91)	.00 (.88)
Propriety	-.11 (.89)	-.07 (.99)	-.14 (.70)	-.10 (.84)	-.02 (.79)	.23 (.83)	-.03 (.73)	.12 (.87)
Workaholism	-.18 (.83)	-.17 (.95)	.06 (.86)	-.07 (.96)	-.32 (.93)	-.11 (.95)	-.03 (.97)	-.03 (1.1)

Note. Paper-and-pencil scores are full-scale thetas. P-P = paper-and-pencil at Times 1 and 2 (n = 106), C-P = computerized at Time 1 and paper-and-pencil at Time 2 (n = 105), P-C = paper-and-pencil at Time 1 and computerized at Time 2 (n = 102), C-C = computerized at Times 1 and 2 (n = 100).

Table C5: Time 1 Descriptive Statistics (Raw Score Metric) for the Trait and Temperament Scales.

	PP		CC		CA	
Scale (no. of items)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
Negative Temperament (28)	13.7	(7.4)	14.1	(6.9)	14.3	(6.5)
Mistrust (19)	6.5	(4.3)	6.3	(4.8)	6.5	(4.3)
Manipulativeness (20)	6.3	(4.0)	5.7	(3.9)	5.7	(3.3)
Aggression (20)	4.3	(4.4)	4.9	(4.2)	4.7	(3.8)
Self-harm (16)	2.1	(2.8)	1.9	(2.6)	2.7	(2.0)
Eccentric Perceptions (15)	5.4	(3.5)	5.1	(3.4)	5.2	(3.2)
Dependency (18)	5.4	(3.8)	6.2	(4.1)	6.1	(3.5)
Positive Temperament (26)	18.0	(5.8)	18.3	(5.9)	18.4	(5.3)
Exhibitionism (16)	8.4	(4.0)	8.5	(3.9)	8.4	(3.3)
Entitlement (16)	8.4	(3.5)	8.0	(3.3)	8.6	(3.0)
Detachment (18)	4.5	(3.9)	4.3	(3.9)	4.5	(3.4)
Disinhibition (35)	13.5	(6.8)	12.5	(6.1)	12.2	(5.0)
Impulsivity (19)	6.9	(4.3)	6.7	(4.3)	6.7	(3.6)
Propriety (20)	11.2	(4.1)	11.0	(3.6)	11.3	(2.9)
Workaholism (18)	6.4	(3.8)	7.3	(4.0)	7.6	(3.5)

Note. PP = paper-and-pencil (n = 208); Computerized group raw-scale means (n = 205) were calculated using both conventional (CC) and adaptive (CA) scoring.

Table C6: Time 2 Descriptive Statistics (Raw Score Metric) for the Trait and Temperament Scales.

	PP		CC		CA	
Scale (no. of items)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
Negative Temperament (28)	13.1	(7.3)	13.3	(7.3)	13.9	(6.9)
Mistrust (19)	5.8	(4.6)	6.4	(5.1)	6.6	(4.7)
Manipulativeness (20)	6.2	(4.4)	5.8	(4.1)	5.7	(3.5)
Aggression (20)	4.4	(4.7)	4.1	(4.0)	3.9	(3.5)
Self-harm (16)	1.5	(2.5)	2.0	(2.8)	2.7	(2.0)
Eccentric Perceptions (15)	4.4	(3.7)	5.2	(3.8)	5.2	(3.5)
Dependency (18)	5.4	(4.0)	5.8	(4.0)	5.5	(3.3)
Positive Temperament (26)	18.7	(5.9)	19.1	(5.7)	19.0	(5.1)
Exhibitionism (16)	8.7	(4.3)	8.6	(4.3)	8.5	(3.8)
Entitlement (16)	8.7	(3.7)	8.4	(3.6)	8.6	(3.1)
Detachment (18)	4.2	(3.9)	4.2	(4.0)	4.7	(3.7)
Disinhibition (35)	13.2	(6.8)	13.1	(6.6)	12.4	(5.3)
Impulsivity (19)	6.8	(4.5)	6.6	(4.2)	6.5	(3.6)
Propriety (20)	11.1	(4.4)	12.3	(4.1)	12.3	(3.3)
Workaholism (18)	7.1	(4.1)	7.1	(4.4)	7.3	(3.8)

Note. PP = paper-and-pencil (n = 211); Computerized group raw-scale means (n = 202) were calculated using both conventional (CC) and adaptive (CA) scoring.

Table C7: Time 1 Descriptive Statistics (Theta Metric) for the Trait and Temperament Scales.

	PP		CC		CA	
Scale	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
Negative Temperament	-.08	(.94)	-.03	(.87)	-.02	(.85)
Mistrust	-.05	(.82)	-.07	(.91)	-.05	(.91)
Manipulativeness	.19	(.86)	.14	(.83)	.15	(.83)
Aggression	-.14	(.91)	.05	(.86)	.04	(.85)
Self-harm	-.45	(.68)	-.03	(.47)	-.02	(.45)
Eccentric Perceptions	.17	(.83)	.17	(.84)	.18	(.89)
Dependency	.00	(.88)	.22	(.92)	.24	(.92)
Positive Temperament	.17	(.92)	.22	(.94)	.22	(.90)
Exhibitionism	.27	(.95)	.30	(.88)	.30	(.87)
Entitlement	.22	(.86)	.21	(.83)	.27	(.86)
Detachment	-.34	(.83)	-.36	(.87)	-.37	(.86)
Disinhibition	.19	(.90)	.14	(.76)	.10	(.73)
Impulsivity	-.01	(.90)	-.04	(.91)	-.03	(.88)
Propriety	-.07	(.84)	-.09	(.73)	-.09	(.71)
Workaholism	-.25	(.88)	-.01	(.90)	.02	(.91)

Note. PP = paper-and-pencil (n = 208); Computerized group theta-scale means (n = 205) were calculated using both conventional (CC) and adaptive (CA) scoring.

Table C8: Time 2 Descriptive Statistics (Theta Metric) for the Trait and Temperament Scales.

	PP		CC		CA	
Scale	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
Negative Temperament	-.18	(.97)	-.15	(.93)	-.10	(.91)
Mistrust	-.21	(.89)	-.07	(.98)	-.04	(.97)
Manipulativeness	.17	(.97)	.13	(.89)	.13	(.87)
Aggression	-.14	(.96)	-.11	(.87)	-.16	(.86)
Self-harm	-.62	(.67)	-.04	(.50)	-.04	(.46)
Eccentric Perceptions	-.08	(.94)	.16	(.95)	.15	(.99)
Dependency	-.01	(.94)	.10	(.90)	.08	(.88)
Positive Temperament	.29	(.94)	.36	(.93)	.34	(.89)
Exhibitionism	.35	(.98)	.32	(.99)	.32	(.99)
Entitlement	.34	(.93)	.26	(.90)	.28	(.92)
Detachment	-.40	(.86)	-.36	(.93)	-.32	(.92)
Disinhibition	.18	(.89)	.18	(.81)	.13	(.77)
Impulsivity	-.05	(.94)	-.07	(.90)	-.07	(.88)
Propriety	-.09	(.92)	.18	(.87)	.18	(.85)
Workaholism	-.12	(.95)	-.10	(1.04)	-.07	(1.02)

Note. PP = paper-and-pencil (n = 211); Computerized group theta-scale means (n = 202) were calculated using both conventional (CC) and adaptive (CA) scoring.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.) Washington, DC: Author.

Assessment Systems Corporation. (1996). *User's manual for the MicroCAT testing system (3rd Ed.)*. St. Paul, MN: Assessment Systems Corporation.

Assessment Systems Corporation. (1999). *User's manual for the POSTSIM adaptive test simulation program*. St. Paul, MN: Assessment Systems Corporation.

Benet-Martinez, V., and John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. Journal of Personality and Social Psychology, 75, 729-750.

Ben-Porath, Y. S., & Butcher, J. N. (1986). Computers in personality assessment: A brief past, an ebullient present, and an expanding future. *Computers in Human Behavior*, *2*, 167-182.

Ben-Porath, Y. S., Slutske, W. S., & Butcher, J. N. (1989). A real-data simulation of computerized adaptive administration of the MMPI. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*, 18-22.

Bentler, P. M., & Wu, E. J. C. (1995). *EQS for Macintosh user's guide*. Encino, CA: Multivariate Software.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In L. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 358-472). Reading, MA: Addison-Wesley.

Biskin, B. H., & Kolotkin, R. C. (1977). Effects of computerized administration on scores on the Minnesota Multiphasic Personality Inventory. *Applied Psychological Measurement, 1,* 543-549.

Bock, R. D., & Aiken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443-459.

Butcher, J. N. (1987). The use of computers in psychological assessment: An overview of practices and issues. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 292-324). New York: Basic Books.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Butcher, J. N., Keller, L. S., & Bacon, S. F. (1985). Current developments and future directions in computerized personality assessment. *Journal of Consulting and*

Clinical Psychology, 53, 803-815.

Butcher, J. N., Perry, J. N., & Atlis, M. M. (2000). Validity and utility of computer-based test interpretation. *Psychological Assessment, 12*, 6-18.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *Minnesota Multiphasic Personality Inventory-Adolescent (MMPI-A): Manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.

Casillas, A. & Clark, L. A. (2000, August). *Dependency, impulsivity, and self-harm: Traits hypothesized to underlie the association between personality and substance use disorders*. Poster presented at the 108th Annual Meeting of the American Psychological Association, Washington, DC.

Casillas, A., Clark, L. A., & Hall, J. A. (2001, August). *Personality predicts problem outcomes in substance-abusers*. Poster presented at the 109th Annual Meeting of the American Psychological Association, San Francisco, CA.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523-562.

Clark, L. A. (1993). *Schedule for nonadaptive and adaptive personality (SNAP). Manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.

Clark, L. A., Livesley, W. J., Schroeder, M. L., & Irish, S. (1996). The structure of maladaptive personality traits: Convergent validity between two systems. *Psychological Assessment, 8*, 294-303.

Clark, L. A., Simms, L. J., Wu, K. D., & Casillas, A. (2002). *Schedule for nonadaptive and adaptive personality (SNAP). Revised manual for administration, scoring, and interpretation*. Manuscript in preparation.

Clark, L. A., Vittengl, J. R., Jarrett, R., & Kraft, D. (2001). *Why personality measures don't predict depression treatment outcomes: Partialling state from trait variance in assessment*. Unpublished manuscript.

Clark, L. A., Vorhies, L., & McEwen, J. L. (1994). Personality disorder symptomatology from the five-factor model perspective. In P. T. Costa, Jr., & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 95-117). Washington, DC: American Psychological Association.

Conn, S., & Rieke, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare

Psychopathy Checklist—Revised. *Psychological Assessment, 9*, 3-14.

Cooke, D. J., Mishie, C., Hart, S. D., & Hare, R. D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist—Revised (PCL:SV): An item response theory analysis. *Psychological Assessment, 11*, 3-13.

Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1992). *The NEO PI-R: The Revised NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.

Dwight, S. A., & Feigelson, M. E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement, 60*, 340-360.

Ekselius, L., Tillfors, M., Furmark, T., & Fredrikson, M. (2001). Personality disorders in the general population: DSM-IV and ICD-10 defined prevalence as related to sociodemographic profile. *Personality & Individual Differences, 30*, 311-320.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495-515.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341-349.

Embretson, S. E., & Hershberger, S. L. (1999). Summary and future of psychometric methods in testing. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 243-254). Mahwah, NJ: Erlbaum.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. London: Hodder & Stoughton.

Eysenck, H. J., & Eysenck, S. B. G. (1991). *Manual of the Eysenck Personality Scales (EPS Adult)*. London: Hodder & Stoughton.

Finger, M. S., & Ones, D. S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment, 11*, 58-66.

Fitts, W. H. (1965). A manual for the Tennessee Self-Concept Scale. Nashville, TN: Counselor Recordings and Tests.

Forbey, J. D., Handel, R. W., & Ben-Porath, Y. S. (2000). A real data simulation of computerized adaptive administration of the MMPI-A. *Computers in Human Behavior, 16*, 83-96.

Fowler, R. D. (1985). Landmarks in computer-assisted psychological assessment. *Journal of Consulting and Clinical Psychology, 53*, 748-759.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences. San Francisco*: W. H. Freeman and Co.

Goldberg, L. R. (1997). A broad-bandwidth, public-domain, personality inventory measuring lower level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personaliity psychology in Europe, Vol. 7*. The Netherlands: Tilburg University Press.

Golden, C. J. (1987). Computers in neuropsychology. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 325-343). New York: Basic Books.

Gough, H. G. (1975). *California Psychological Inventory manual*. Palo Alto, CA: Consulting Psychologists Press, Inc.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Handel, R. W., Ben-Porath, Y. S., & Watt, M. (1999). Computerized adaptive assessment with the MMPI-2 in a clinical setting. *Psychological Assessment, 11*, 369-280.

Hansen, J. C. (1987). Computer-assisted interpretation of the Strong Interest Inventory. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 292-324). New York: Basic Books.

Hare, R. D. (1991). *The Hare Psychopathy Checklist—Revised*. Toronto, Ontario, Canada: Multi-Health Systems.

Harlan, E., & Clark, L. A. (1999). Short forms of the Schedule for Nonadaptive and Adaptive Personality (SNAP) for self- and collateral ratings: Development, reliability, and validity. *Assessment, 6*, 131-145.

Hart, R. R., & Goldstein, M. A. (1985). Computer-assisted psychological assessment. *Computers in Human Services, 1*, 69-75.

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare Psychopathy Checklist: Screening Version* (1st ed.). Toronto, Ontario, Canada: Multi-Health Systems.

Hathaway, S. R., & McKinley, J. C. (1951). *The Minnesota Multiphasic Personality Inventory Manual* (rev.). New York: Psychological Corporation.

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology, 53*, 826-838.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Honaker, L. M. (1988). The equivalency of computerized and conventional MMPI administration: A critical review. *Clinical Psychology Review*, *8*, 561-577.

Honaker, L. M., Harrell, T. H., & Buffaloe, J. D. (1988). Equivalency of Microtest computer MMPI administration for standard and special scales. *Computers in Human Behavior*, *4*, 323-337.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureño, G., & Villaseñor, V. S. (1988). Inventory of Personal Problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56*, 885-892.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Huang, C., Church, A., & Katigbak, M. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross Cultural Psychology, 28*, 192-218.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory—versions 4a and 54*. Technical Report, Institute of Personality and Social Research, University of California, Berkeley, CA.

Kamakura, W. A., & Balasubramanian, S. K. (1989). Tailored interviewing: An application of item response theory for personality measurement. *Journal of Personality Assessment*, *53*, 502-519.

Karson, S., & O'Dell, J. W. (1987). Computer-based interpretation of the 16PF: The Karson clinical report in contemporary practice. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 218-235). New York: Basic Books.

Kim, J., & Pilkonis, P. A. (1999). Selecting the most informative items in the IIP scales for personality disorders: An application of item response theory. *Journal of Personality Disorders*, *13*, 157-174.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*, 457-477.

Knowles, E. S. (1988). Item context effects in personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Lachar, D. (1987). Automated assessment of child and adolescent personality: The personality inventory for children (PIC). In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 261-291). New York: Basic Books.

Lambert, M. E., Andrews, R. H., Rylee, D., & Skinner, J. R. (1987). Equivalence of computer and traditional MMPI administration with substance abusers. *Computers in Human Behavior, 3*, 139-143.

Levine, M. V. (1984). *An introduction to multilinear formula score theory*. (Personnel and Training Research Programs, Office of Naval Research, Measurement Series No. 84-4). Arlington, VA: Personnel and Training Research Programs.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lushene, R., O'Neil, H., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment, 38*, 353-361.

Mackinnon, A., Jorm, A. F., Christensen, H., Scott, L. R., Henderson, A. S., & Korten, A. E. (1995). A latent trait analysis of the Eysenck Personality Questionnaire in an elderly community sample. *Personality and Individual Differences, 18*, 739-747.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press.

Mennin, D. S., & Heimberg, R. G. (2000). The impact of comorbid mood and personality disorders in the cognitive-behavioral treatment of panic disorder. *Clinical Psychology Review, 20*, 339-357.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examinations General Test. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117-135). Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Chicago: Scientific Software International.

Moreno, K. E., & Segall, O. D. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169-179). Washington, DC: American Psychological Association.

Orlando, M., & Thissen, D. (2000). Liklihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Pilkonis, P. A., Kim, Y., Proietti, j. M., & Barkham, M. (1996). Scales for personality disorders developed from the Inventory of Interpersonal Problems. *Journal of Personality Disorders, 10*, 355-369.

Pinsoneault, T. B. (1996). Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2. *Computers in Human Behavior, 12*, 291-300.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danske Paedagogike Institut.

Ready, R. E., Clark, L. A., Watson, D., & Westerhouse, K. (2000). Self- and peer-related personality: Agreement, trait ratability, and the "self-based heuristic". *Journal of Research in Personality, 34*, 208-224.

Ready, R. E., Stierman, L., & Paulsen, J. S. (2001). Ecological validity of neuropsychological and personality measures of executive functions. *Clinical Neuropsychologist, 15*, 314-323.

Ready, R.E. & Clark, L.A. (2002). Correspondence of psychiatric patient and informant ratings of personality traits, temperament, and interpersonal problems. *Psychological Assessment, 14*, 39-49.

Ready, R.E., Watson, D., & Clark, L.A. (in press). Psychiatric patient and informant reported personality: Predicting concurrent and future Behavior. *Assessment*.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (219-242). Mahwah, NJ: Erlbaum.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO-PI-R. *Assessment, 7*, 347-364.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Reynolds, S. K., & Clark, L. A. (2001). Predicting dimensions of personality disorder from domains and facets of the Five-Factor Model. *Journal of Personality, 69*, 199-222.

Richman, W. L., Kiesler, S., Weisband, S. & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84,* 754-775.

Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study of personality change in young adulthood. Journal of Personality, 69, 617-640.

Roper, B. L., Ben-Porath, Y. S., & Butcher, J. N. (1991). Comparability of computerized adaptive and conventional testing with the MMPI-2. *Journal of Personality Assessment, 57,* 278-290.

Roper, B. L., Ben-Porath, Y. S., & Butcher, J. N. (1995). Comparability and validity of computerized adaptive with the MMPI-2. *Journal of Personality Assessment, 65,* 358-371.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 Scales. *Journal of Personality Assessment, 72,* 282-307.

Rozensky, R. H., Honor, L. F., Rasinski, K., Tovian, S. M., & Herz, G. I. (1986). Paper-and-pencil versus computer-administered MMPIs: A comparison of patients' attitudes. *Computers in Human Behavior, 2,* 111-116.

Russell, G. K. G., Peace, K. A., & Mellsop, G. W. (1986). The reliability of a micro-computer administration of the MMPI. *Journal of Clinical Psychology, 42,* 120-122.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs, 34* (4, Pt. 2, Whole No. 17).

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

SAS Institute, Inc. (1990). SAS Procedures Guide, Version 6, Third Edition. SAS Instiute, Inc.: Cary, NC.

Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). *Comparability of paper-and-pencil and computer adaptive test scores on the GRE General Test* (Research Report No. RR-98-38). Princeton, NJ: Educational Testing Service.

Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). *Field test of a computer-based GRE General Test* (Research Report No. RR-93-07). Princeton, NJ: Educational Testing Service.

Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer-adative GRE General Test* (Research Report No. RR-95-20). Princeton, NJ: Educational Testing Service.

Schubert, D. S. (1975). Increase of personality response consistency by prior response. *Journal of Clinical Psychology*, *31*, 651-658

Schuldberg, D. (1988). The MMPI is less sensitive to the automated testing format than it is to repeated testing: Item and scale effects. *Computers in Human Behavior*, *4*, 285-298.

Schuldberg, D. (1990). Varieties of inconsistency across test occasions: Effects of computerized test administration and repeated testing. *Journal of Personality Assessment*, *55*, 168-182.

Segall, D. O., & Moreno, K. E. (1999). Development of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 35-65). Mahwah, NJ: Erlbaum.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology, 75*, 1350-1362.

Snyder, D. K. (2000). Computer-assisted judgement: Defining strengths and liabilities. *Psychological Assessment, 12*, 52-60.

Soldz, S.., Vaillant, G. E. (1999). The Big Five personality traits and the life course: A 45-year longitudinal study. *Journal of Research in Personality*, *33*, 208-232.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, *66*, 341-349.

Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.

Thomas, T. J. (1990). Item-presentation controls for multidimensional item pools in computerized adaptive testing. *Behavior Research Methods*, *22*, 247-252.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg Self-Esteem Scale: A Comparison of psychometric features and respondent preferences. *Educational and Psychological Measurement, 61*, 461-474.

Vispoel, W. P., Wang, T., & Bleiler, T. (1997). Computerized adaptive and fixed-item testing of music listening skill: A comparison of efficiency, precision, and concurrent validity. *Journal of Educational Measurement, 34*, 43-63.

Vittengl, J. R., Clark, L. A., Owen-Salters, E., & Gatchel, R. J. (1999). Diagnostic change and personality stability following functional restoration treatment in a chronic low back pain patient sample. *Assessment, 6*, 79-92.

Vittengl, J. R., Clark, L. A., Owen-Salters, E., & Gatchel, R. J. (1999). Diagnostic change and personality stability following functional restoration treatment in chronic low back pain patients. *Assessment, 6*, 79-91.

Waller, N. G. (1995). *MicroFACT 1.0* [computer program]. St. Paul, MN: Assessment Systems Corporation.

Waller, N. G. (1999). Searching for structure in the MMPI. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (185-218). Mahwah, NJ: Erlbaum.

Waller, N. G. (2002). *MicroFACT 2.0* [computer program]. St. Paul, MN: Assessment Systems Corporation.

Waller, N. G., & Reise, S. P., (1989). Computerized adaptive personality assessment: An illustration with the Absorption scale. *Journal of Personality and Social Psychology, 57*, 1051-1058.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality, 64*, 545-576.

Wang, T., Hanson, B. A., & Lau, C. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23*, 263-278.

Watson, C. G., Manifold, V., Klett, W. G., Brown, J., Thomas, D., & Anderson, D. (1990). Comparability of computer- and booklet-administered Minnesota Multiphasic Personality Inventories among primarily chemically dependent patients. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*, 276-280.

Watson, C. G., Thomas, D., & Anderson, P. E. D. (1992). Do computer administered Minnesota Multiphasic Personality Inventories underestimate booklet-based scores? *Journal of Clinical Psychology, 48*, 744-748.

Watson, D., & Clark, L. A. (1994). *Manual for the Positive and Negative Affect Schedule (Expanded Form)*. University of Iowa; available from the authors.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789.

Weiss, D. J., & Vale, C. D. (1987). Computerized adaptive testing for measuring abilities and other psychological variables. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 325-343). New York: Basic Books.

White, D. M., Clements, C. B., & Fowler, R. D. (1985). A comparison of computer administration with standard administration of the MMPI. *Computers in Human Behavior, 1*, 143-162.

Wilson, D., Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Chicago: Scientific Software International.

Wilson, F. R., Genco, K. T., & Yager, G. G. (1985). Assessing the equivalence of paper-and-pencil vs. computerized tests: demonstration of a promising methodology. *Computers in Human Behavior, 1*, 265-275.

Windle, C. (1954). Test-retest effect on personality questionnaires. *Educational & Psychological Measurement, 14*, 617-633.

Windle, C. (1955). Further studies of test-retest effect on personality questionnaires. *Educational & Psychological Measurement, 15*, 246-253.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 219-226.

Wu, K., & Clark, L. A. (2002). *Daily Behavior: Empirical Relations with Personality Traits*. Manuscript submitted for publication.

Zickar, M. J. (2001). Conquering the next frontier: Modeling personality data with item responses theory. In B. W. Roberts & R. Hogan (Eds.), *Personality Psychology in the Workplace*. Washington, DC: American Psychological Association.