

Validation of a Computerized Adaptive Version of the Schedule for Nonadaptive and Adaptive Personality (SNAP)

Leonard J. Simms and Lee Anna Clark
University of Iowa

This is a validation study of a computerized adaptive (CAT) version of the Schedule for Nonadaptive and Adaptive Personality (SNAP) conducted with 413 undergraduates who completed the SNAP twice, 1 week apart. Participants were assigned randomly to 1 of 4 retest groups: (a) paper-and-pencil (P&P) SNAP, (b) CAT, (c) P&P/CAT, and (d) CAT/P&P. With number of items held constant, computerized administration had little effect on descriptive statistics, rank ordering of scores, reliability, and concurrent validity, but was preferred over P&P administration by most participants. CAT administration yielded somewhat lower precision and validity than P&P administration, but required 36% to 37% fewer items and 58% to 60% less time to complete. These results confirm not only key findings from previous CAT simulation studies of personality measures but extend them for the 1st time to a live assessment setting.

Over the past 3 decades, computers increasingly have been used to automate the administration, scoring, and interpretation of results from a wide variety of psychological measures, including assessment of ability and academic achievement (e.g., Mills, 1999; Zenisky & Sireci, 2002), neuropsychological status (e.g., Russell, 2000), vocational interests (e.g., Hansen, Neuman, Haverkamp, & Lubinski, 1997), and personality (e.g., Roper, Ben-Porath, & Butcher, 1991, 1995; Vispoel, 2000; Vispoel, Boo, & Bleiler, 2001; Waller & Reise, 1989). Computers provide an objective, efficient, and reliable means for delivering assessment services to clients and research participants (e.g., Butcher, 1987). In the personality domain, the earliest computerized applications (e.g., Lushene, O'Neil, & Dunn, 1974) simply administered and provided scores for traditional paper-and-pencil (P&P) tests such as the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1951); others, however, went further by providing computerized interpretive reports for each examinee (e.g., Butcher, Perry, & Atlis, 2000; Snyder, 2000). Today, of course, many psychological tests have computerized versions.

A concern in both research and clinical settings is the length of many personality measures. For instance, an hour or longer often is required to complete such measures as the 567-item MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), the 344-item Personality Assessment Inventory (Morey, 1991), the 240-item NEO Personality Inventory—Revised (NEO-PI-R; Costa & McCrae, 1992), and the 375-item Schedule for Nonadap-

tive and Adaptive Personality (SNAP; Clark, 1993). The time required for such assessments is difficult to accommodate in many applied and research settings. Managed care companies have limited the types of assessments for which they will reimburse practitioners to those that require less time and effort to administer, score, and interpret. Research time also is scarce and costly. Moreover, long measures can lead to fatigue and drifting attention for many test takers, which ultimately compromise the validity of the test profile and complicate test interpretation. To address these concerns, many researchers have developed abbreviated versions, or short forms, but these suffer from a number of limitations (e.g., Smith, McCarthy, & Anderson, 2000). The reliability of results from short forms often is significantly lower than that of the original measures. In addition, short forms often are created for use in very specific populations (e.g., adolescents, psychiatric patients, undergraduates), which makes them perform less optimally when used in populations for which the short form was not initially intended. Finally, scores obtained from short forms often are, without empirical basis, ascribed the same validity status as the longer measures from which they were derived.

Computerized Adaptive Testing

A number of personality and measurement researchers (e.g., Reise & Henson, 2000; Roper et al., 1991, 1995; Waller & Reise, 1989) have discussed ways to use the flexibility afforded by personal computers to shorten measures in a way that maximizes comparability with the original construct and minimizes loss of measurement precision. This set of techniques, which originated in the ability and achievement testing literature, collectively is known as computerized adaptive testing (CAT). In the most basic sense, CAT permits the selection and administration of items that are individually tailored to the trait level of the examinee, with the potential of substantial item and time savings (Sands, Waters, & McBride, 1997; Wainer, 2000; Weiss, 1985). A typical CAT selects and administers only those items that provide the most psychometric information (i.e., yield the lowest standard errors of measurement) at a given trait level, eliminating the need to ad-

Leonard J. Simms and Lee Anna Clark, Department of Psychology, University of Iowa.

This work was supported by a grant from the University of Minnesota Press. We thank Eva Klohnen, James Marchman, Gregg Oden, Walter Vispoel, and David Watson for their helpful comments regarding the design of this study, as well as the undergraduate research participants who kindly volunteered to participate.

Correspondence concerning this article should be addressed to Leonard J. Simms, who is now at the Department of Psychology, Park Hall 218, University at Buffalo, State University of New York, Buffalo, NY 14260. E-mail: ljsimms@buffalo.edu

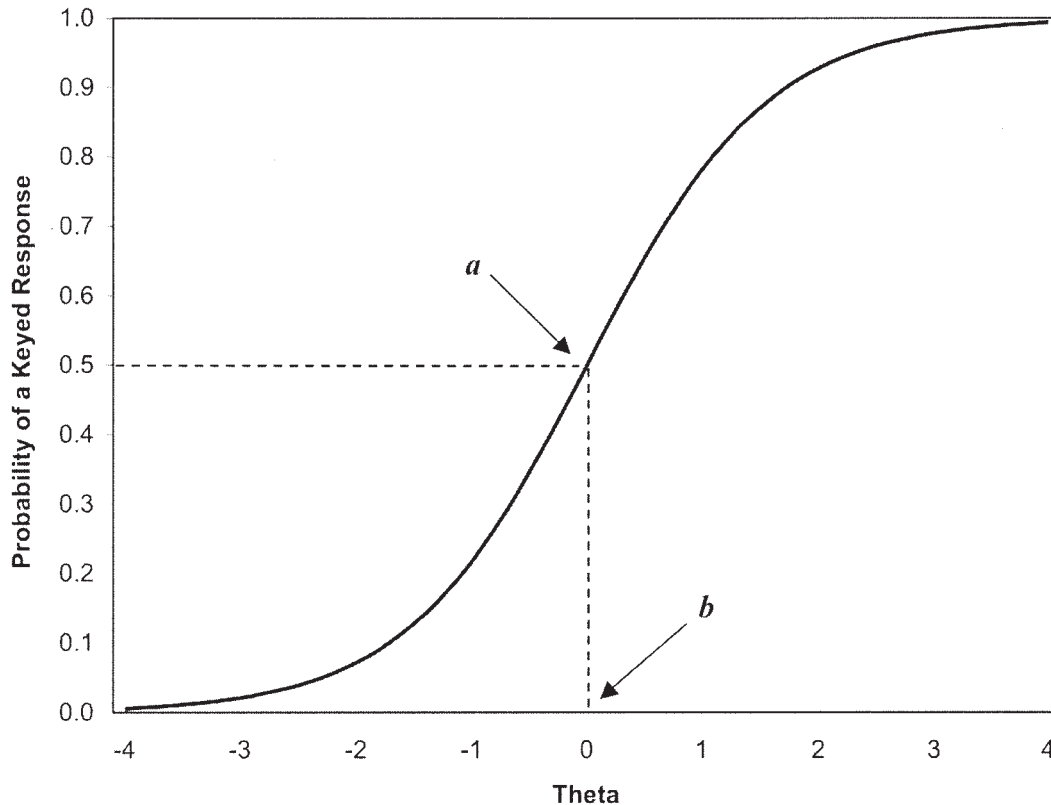


Figure 1. Parameters of a typical item response function.

minister items that have very low or very high endorsement probabilities given a particular examinee's trait level.

To illustrate, consider a hypothetical 40-item scale to assess mathematical ability, and assume that this ability constitutes a continuous dimension ranging from very low to very high ability. If, after several items are administered, it becomes clear that a given examinee is capable of answering trigonometry items correctly, then why ask the examinee to respond to simple multiplication items (which clearly are easier than trigonometry items)? In contrast to traditional testing (in which all 40 items must be administered in order to obtain a score), a CAT would administer only those mathematics items that are highly informative for this examinee. Thus, if the computer estimates that 20 items are too easy for this examinee, then only the remaining 20 hard items are administered. This greater efficiency, represented by item savings of 50% in this example, is the primary advantage of CAT over other forms of traditional and computerized assessment. Extrapolating to the personality domain, consider a 40-item measure designed to tap trait aggression. Similar to the previous example, if it becomes clear after several item responses that a given examinee readily endorses items such as "I have been in many fist fights in my lifetime," then presenting items such as "I sometimes hit things" is unnecessary because the response to the latter item may be inferred from the former item's response.

Item Response Theory

How does the computer know how difficult an item is or how much psychometric information it provides at different levels of

the trait? Clearly, the ability of a CAT application to work efficiently rests on its capacity to calibrate items properly along these dimensions. To do this, CATs typically are built on a foundation laid by item response theory (IRT). IRT includes a variety of related psychometric models that characterize test items by one or more item parameters (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985; Lord, 1980). These item parameters define an item response function (IRF; see Figure 1 for a prototypic IRF) that is unique to each item. An IRF describes the regression of the probability of a particular item response on an underlying trait. In IRT models, this underlying trait dimension is referred to as theta (θ). In most personality applications (e.g., Kamakura & Balasubramanian, 1989; Reise, 1999; Reise & Waller, 1990; Waller, 1999; Waller & Reise, 1989), the IRF is defined by two item parameters: item discrimination and item difficulty (also referred to as the *a* and *b* parameters, respectively) as depicted in Figure 1.¹ In two-parameter IRT models, item difficulty refers to the point along the trait continuum that is associated with a 50% probability that an examinee at that theta level will respond to the item in the keyed direction. High values of item difficulty are associated with items that have low endorsement probabilities (i.e., that reflect higher levels of the trait or ability). Item discrimination reflects the slope of the IRF at the difficulty level for the item. Steeper slopes reflect greater discrim-

¹ A third parameter—the pseudoguessing parameter (*c*)—is typically modeled in ability testing applications in which chance correct responding is an important phenomenon to model.

inatory power and psychometric information. The formula for the two-parameter logistic model (Birnbaum, 1968) is provided in Appendix A.

A powerful aspect of IRT for CAT applications is the concept of *information*. The parameters of an IRF can be combined into a single index that describes how precisely an item measures a trait at various points along the trait continuum (Hambleton et al., 1991; Hambleton & Swaminathan, 1985; 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 2 for several hypothetical IICs). An IIC has its peak at the difficulty level of the item, and the relative height of its peak (compared with other items in the same pool) is related to the item's discrimination parameter. CAT applications use item information to select for administration only those items that provide maximum psychometric information (i.e., are most discriminating) throughout a given test administration (Weiss, 1985).

Figure 2 exemplifies the information provided by four hypothetical test items across the trait continuum. Consider a CAT in which the current trait estimate is $\theta = 1.0$. Assuming the trait estimate does not change markedly after each item response, the computer would present the items in the following order: 1, 4, 2, 3. If, however, the trait estimate was located at $\theta = -1.0$, the items would be presented in this order: 2, 1, 3, 4. Another important concept in CAT is the *termination rule*, which indicates that item administration should stop when one or more prespecified conditions are satisfied. One such condition might be to stop presenting new items when reasonably informative items no longer exist in

the pool; other types of termination rules will be discussed below. Taking this example a step further, Items 3 and 4 might not have been administered in the first and second examples, respectively, because they offer very little psychometric information at those levels of theta. Thus, termination rules limit the number of items administered, and for CAT applications in which the item pool for a particular scale is sufficiently broad and large, marked efficiency gains can be achieved.

Most IRT models assume scale unidimensionality, which requires that items within a scale cohere and measure a common latent trait. This requirement has precluded the use of IRT for the analysis of relatively heterogeneous scales such as those from the MMPI and MMPI-2. Several attempts have been made to administer the MMPI adaptively with non-IRT methodology (Ben-Porath, Slutske, & Butcher, 1989; Handel, Ben-Porath, & Watt, 1999; Roper et al., 1991, 1995), but these methods are less precise and efficient than IRT-based CATs and generally result in substantial loss of information. Personality measures designed to measure relatively homogeneous traits (e.g., those developed with factor analytic methods) are better candidates for IRT calibration.

CAT, IRT, and Personality Assessment

Relatively few applications of IRT-based CAT appear in the personality literature, and all are computerized simulations. In the earliest study, Waller and Reise (1989) applied the two-parameter logistic model to the Absorption scale of the Multidimensional Personality Questionnaire (Tellegen, 1982). Waller and Reise used real-data simulations, based on responses from 1,000 participants

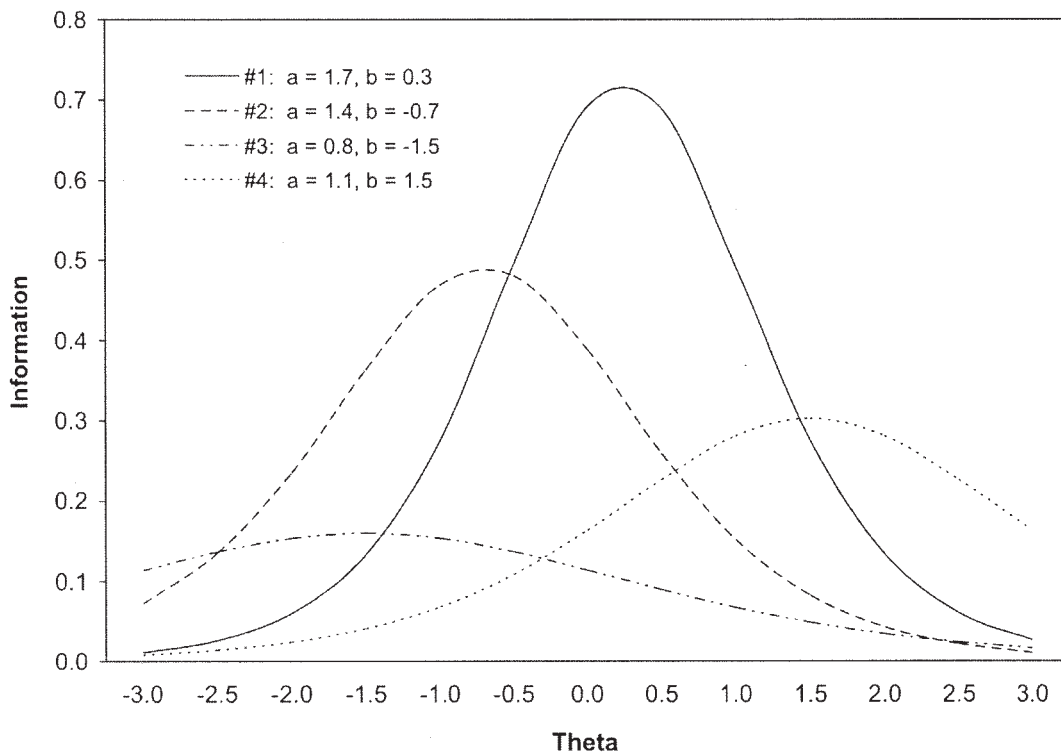


Figure 2. Several typical item information curves.

who previously had completed the Absorption scale in the traditional P&P format, to illustrate two possible adaptive testing strategies. Using a fixed-test-length strategy, in which the computer administered a fixed number of items, Waller and Reise achieved 50% item savings with little loss of measurement precision. Using a clinical-decision testing strategy, in which items were administered until the confidence interval surrounding trait estimates no longer included the cutoff value used to classify the subjects, they identified with perfect accuracy, with, on average, only 25% of the available items, individuals who were extreme on the Absorption trait (on the traditional P&P version).

In a similar demonstration, Kamakura and Balasubramanian (1989) conducted real-data CAT simulations on a relatively unidimensional subset of items from the Socialization scale of the California Psychological Inventory (Gough, 1975). They tested three possible adaptive strategies: (a) proceed until a maximum of 15 items were asked, (b) continue testing until the standard error around each examinee's trait estimate dropped below a prespecified value (e.g., 0.40), and (c) combine a and b—administer at least 10 items, and continue testing until a minimum standard error (of 0.40) was attained. These three CAT algorithms yielded item savings of 66%, 61%, and 60%, respectively. However, item savings were obtained at the expense of measurement error. The authors reported some loss in measurement precision for scores estimated from adaptive testing, but they concluded that these small increases in error were overshadowed by the substantial savings that were realized.

Waller (1999) completed a similar analysis using item responses from the MMPI. To circumvent the problem of multidimensionality described above, Waller (1999) factor analyzed MMPI item responses to yield 16 unidimensional factors within the MMPI item pool. He chose one such factor, 51 items labeled Denial of Somatic Complaints, for an IRT-based real-data CAT simulation. In this algorithm, item administration continued until two termination criteria were satisfied: (a) administer at least 20 items and (b) administer items until the item information associated with any remaining items fell below a threshold of 0.10 (i.e., until reasonably informative items no longer existed for a given examinee). Using this rule, the computer terminated item administration after only 20 items (representing item savings of 61%) for over half of those in the data set.

Finally, Reise and Henson (2000) conducted real-data simulations on the 30 facet scales of the NEO Personality Inventory—Revised (NEO-PI-R; Costa & McCrae, 1992). Using a polytomous IRT model (to account for the Likert scale used to rate NEO-PI-R items) and a maximum-information item selection adaptive testing algorithm, they achieved item savings of 50%, on average, across the 30 facet scales, with little loss in measurement precision. The authors also found that their CAT algorithm resulted in surprisingly little variability in the items administered within each facet. For 23 of the 30 facet scales, the first four CAT-administered items within each facet did not differ across simulated administrations, suggesting that the CAT algorithm may not have been necessary to achieve the results they reported. Reise and Henson concluded that constructing short forms of the facet scales by choosing the four “best” items (i.e., those that provided the most psychometric information) may yield equivalent results to those obtained with an adaptive algorithm. The authors also noted that, although their results and those of previous simulation studies

have been impressive, it is an open question whether the same findings would be obtained with live participants.

Recent data provide evidence that traditional personality measures administered by computer do not differ markedly from standard P&P administrations with respect to descriptive statistics (e.g., Finger & Ones, 1999). However, CATs include features not typically present in standard computerized tests, such as differences in item selection and presentation order across participants. Thus, the equivalence of CATs to their traditional P&P counterparts cannot be assumed and needs to be established empirically. Moreover, studies are needed to explore the feasibility and comparability of adaptive personality measures in which live participants complete IRT-based CATs.

The Present Study

The primary objective of this study was to extend the literature by collecting live testing data from participants completing a CAT version of an established personality measure to provide ecologically valid evidence of psychometric efficiency, comparability of scores across modes of administration, temporal stability, and convergent and discriminant validity. As discussed above, a primary assumption of IRT is scale unidimensionality. Thus, we conducted the present study using a CAT prototype designed from the item pool of the SNAP. The SNAP was developed with a combination of content analytic and factor analytic methods, which makes it a good candidate for an IRT-based conversion to CAT.

The SNAP is a 375-item, dichotomously scored self-report questionnaire that measures 15 relatively distinct personality trait dimensions relevant to personality pathology. The SNAP 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 labeled Negative Affectivity, Positive Affectivity, and Disinhibition Versus Constraint. In addition, the SNAP includes several validity and diagnostic scales that were not adaptively administered in the present study.

Method

Preliminary Studies and SNAP-CAT Construction

We developed a CAT prototype of the SNAP (SNAP-CAT) through a series of studies designed (a) to estimate IRT item parameters for all items within each SNAP scale, (b) to examine the appropriateness of the SNAP for CAT administration and establish an appropriate termination rule, and (c) to build the SNAP-CAT prototype for use in live testing. The full results of these studies are reported elsewhere (Simms, 2004), but we briefly review them here.

Item Parameter Estimation and Unidimensionality Assessment

We calibrated IRT item parameters on a sample of 3,995 individuals (59.2% female; 86.3% Caucasian) composed of 809 community-dwelling adults; 1,886 undergraduates; and 1,300 psychiatric patients who completed the standard P&P SNAP over the past decade. The size of this

calibration sample was more than adequate for this IRT calibration (Reise, 1999; Zickar, 2001). We estimated IRT item parameters for each of the SNAP's 15 trait and temperament scales using BILOG 3.1 (Mislevy & Bock, 1990), with program defaults. Consistent with prior IRT work in the personality literature, we chose to estimate the two-parameter logistic model. Summary statistics for item parameters are presented in Appendix B. BILOG raised some questions regarding model fit for some items. Most scales included very few poorly fitting items, but one scale—the 16-item Self-harm scale—was a clear outlier with 14 items classified as poorly fitting by BILOG. However, we calibrated all SNAP items and scales because (a) we wanted to maintain comparability with the traditional P&P SNAP and (b) BILOG's goodness-of-fit indices are not especially reliable for shorter scales such as those on the SNAP (Mislevy & Bock, 1990).

To assess the assumption of scale unidimensionality underlying most IRT models, we conducted a series of nonlinear factor analyses of the items for each SNAP scale.² Traditional linear factor analysis (i.e., the type implemented by most statistical programs) can result in spurious factors when applied to dichotomous data (Waller, Tellegen, McDonald, & Lykken, 1996). Thus, a number of factor analytic methods have been proposed to deal with the nonlinear item-trait regressions that can result from dichotomously scored items (Mislevy, 1986). Many of these methods rely on factor analyses of tetrachoric correlations, which represent estimates of the correlations that would have been observed had the variables been measured on a continuous metric (e.g., Waller, 1995, 2002). Thus, in the present study, we assessed scale unidimensionality by fitting a one-factor model to the items within each scale using MicroFACT (Waller, 1995, 2002)—software that conducts factor analyses on matrices of tetrachoric correlations—and then examining two indices of model fit calculated by the program: (a) the root-mean-square residual (RMSR) and (b) the goodness-of-fit index. The mean RMSR value was .086 (range = .058–.122), with lower values indicative of better model-to-data fit. The mean goodness-of-fit index value was .963 (range = .942–.986), with all but two scales in the excellent range (i.e., above .950; Hu & Bentler, 1999). These data support the unidimensionality of SNAP scales and provide evidence of their appropriateness for IRT modeling.

Termination Rule Selection

A key decision in CAT is when to stop the test. Commonly used options include the following: (a) after administering a prespecified number of items, (b) after the standard error of the trait estimate falls below a prespecified limit, (c) after reasonably informative items no longer exist for a given examinee, or (d) after some combination of these rules has been satisfied. To aid in the selection of a reasonable termination rule for the SNAP-CAT, we conducted computerized CAT simulations on all SNAP scales to examine the various possible termination rules just described. After multiple trials, we selected a two-part termination rule. First, for each scale, we specified a minimum number of items to administer to all examinees. The minimum test lengths, which appear in Appendix B, were chosen to cap potential information loss at approximately 33% based on results from the simulation studies. Second, we determined that items should be administered adaptively until either (a) the standard error of the trait estimate drops below 0.40 or (b) items with conditional information estimates (i.e., item information keyed to the current trait estimate) greater than 0.10 no longer exist in the pool. This combined termination rule yielded mean simulated item savings of 37% with psychometric information loss of 18%.

SNAP-CAT Construction

These item parameters and termination criteria were used to construct the SNAP-CAT prototype with MicroCAT (Assessment Systems Corporation, 1996), a DOS-based software package that allows the user to program and administer customized CATs.³ The SNAP-CAT was pro-

grammed as 15 independent adaptive tests (i.e., one test for each SNAP scale) that were linked together. To vary the presentation order, participants were randomized to one of the three starting positions. Within each scale, the SNAP-CAT was programmed to administer all items adaptively using a maximum-information item selection strategy, which means that items are selected for administration that provide the most psychometric information given the examinee's current trait estimate (which is recalculated after each item response).

Item presentation algorithm. Each test started with the administration of an item of median difficulty. After the participant responded, the program used the response to estimate theta and then searched the remaining scale items for the single item that provided the most psychometric information at that current trait estimate. The identified item was then administered, followed by theta estimation and assessment of the termination rule. This cycle of item selection, theta estimation, and termination rule assessment repeated until the termination rule described above was satisfied; once met, the adaptive theta estimate, standard error, test completion time, and the number of items administered were recorded. The computer then resumed presenting the remaining items from each scale adaptively until all had been administered, so that computerized full-scale scores could be calculated to provide a validity comparison that controlled for computerized administration. After all items within a given scale were administered, the full-scale theta, standard error, raw score, and test completion time were recorded, and the SNAP-CAT proceeded to the next scale.

User interface. The SNAP-CAT interface was designed to be very simple. Item presentation screens had black backgrounds with yellow text. Items were presented one-per-screen and were centered horizontally in the upper half of the screen. Below each item, three options—"TRUE," "FALSE," and a blank space option—were presented for participants to toggle through by hitting the space bar or arrow keys and to select by hitting the *Return* key. To minimize rapid or random responding, no responses were highlighted when a new item was presented. The blank space option was designed to be the computerized equivalent of leaving an item blank; thus, selecting this option resulted in scoring in the nonkeyed direction, which is typical for the SNAP and most traditional P&P tests. Item omits were tracked and recorded for later decision making regarding the validity of the protocols. After hitting the *Return* key, participants were presented with the next adaptively selected item and were no longer permitted to change their responses to the previously administered item.

Validation Study

Participants

A sample of 491 participants was recruited from undergraduate psychology courses at the University of Iowa. Course credit was offered as compensation for participation. Of these participants, 423 (86.2%) elected to return to the second session 1 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

² We also used other methods for establishing unidimensionality, such as exploratory factor analysis and inspection of alpha coefficients and average interitem correlations, and the results were consistent with those of the nonlinear factor analyses presented here.

³ Memory limitations associated with MicroCAT created a significant challenge to our original design. Our plan was to program the SNAP-CAT to intersperse all items throughout the test, but this proved impossible. Instead, we grouped items within each scale. Notably, this deviation from plan did not lead to markedly different internal consistency coefficients across modes or adversely affect the comparability of the SNAP-CAT to the paper-and-pencil version of the SNAP (SNAP-P&P).

1 and 2, respectively. All subsequent analyses were restricted to those who completed both assessments ($N = 413$). This reduced sample was 68.5% female and 88.6% Caucasian, and the mean age was 19.2 ($SD = 1.3$). These demographic characteristics were essentially the same as those of the full sample.

Testing Procedures

As participants arrived for the first session, they were randomly assigned to one of two blocks: traditional P&P version of the SNAP (SNAP-P&P) or the SNAP-CAT. Participants were invited back for a second session 1 week later during which they again completed either the SNAP-P&P or SNAP-CAT, with half randomly assigned to complete the same version of the SNAP and half to complete the other version. Participants were blind to which version they would be asked to complete until they arrived for the second session.

This assignment protocol yielded the following four groups: (a) P&P retest participants ($n = 106$) who completed the SNAP-P&P twice, (b) computerized retest participants ($n = 100$) who completed the SNAP-CAT twice, (c) P&P/CAT participants ($n = 105$) who completed the SNAP-P&P first and the SNAP-CAT second, and (d) CAT/P&P participants ($n = 102$) who completed the SNAP-CAT first and the SNAP-P&P second. These groups did not differ significantly with regard to sex, $\chi^2(3) = 1.14$, *ns*; ethnicity, $\chi^2(15) = 16.12$, *ns*; or age, $F(3, 409) = 0.60$, *ns*. Also, as described briefly above, the prototype SNAP-CAT used in this study was constructed such that all participants provided both computerized adaptive scores (CA; calculated after the termination rule was satisfied) as well as computerized full-scale scores (CF; calculated after all remaining items were administered) to provide a within-mode comparison between the adaptive and full-test versions.

SNAP-CAT and SNAP-P&P completion times were recorded for all participants. Timing for both modes began after Leonard J. Simms, who led all sessions, read instructions to participants and instructed them to begin, and timing ended when the participants were finished completing the SNAP. The computer recorded SNAP-CAT completion times, whereas the experimenter recorded SNAP-P&P completion times using a stopwatch: Participants completing the SNAP-P&P were asked to raise their hands immediately upon finishing the measure so that the experimenter could (a) promptly record their completion times and (b) hand out the supplemental questionnaire packets.

During the first session, participants also completed a general demographic information sheet and the 44-item version of the Big Five Inventory (BFI; John & Srivastava, 1999). At the second session, returning participants also completed the Eysenck Personality Questionnaire—Revised (EPQ-R; Eysenck & Eysenck, 1991). Participants also completed the Positive and Negative Affect Schedule—Expanded Form (Watson & Clark, 1994) at both sessions, but those data are not analyzed here. All measures other than the SNAP-CAT were completed with a P&P format, and participants recorded their responses on scannable answer sheets. Finally, after the second session, participants in the cross-mode groups were asked, with an open-ended format, which testing mode they preferred and why.

Measures

SNAP-P&P. To permit direct comparisons between the SNAP-CAT and the traditional version, a special P&P version of the SNAP (SNAP-P&P) was constructed that included only those items that were in the SNAP-CAT item pool. Thus, items that score only on scales that were not assessed by the SNAP-CAT (e.g., the diagnostic scales) were excluded. The resultant measure contained 297 items presented in standard-booklet order and printed in a 12-point Times New Roman typeface. The booklet included approximately 30 items per page, pre-

sented in two columns, and participants were instructed to record their answers on a separate scannable answer sheet produced by National Computer Systems (NCS; Sheet MP74593). This answer sheet was light gray in color with green printing, and was generically designed by NCS to accommodate up to 400 true–false responses as well as basic identifying information.

In their traditional format, SNAP scales have been shown to be internally consistent (median alphas range = .80–.85 across college students, community adults, and psychiatric patients) and temporally stable ($Mdn r = .87$ across intervals ranging in length from 7 days to 4 months; Clark, 1993; Clark, Simms, Wu, & Casillas, in press). In addition, the SNAP has demonstrated good convergent and discriminant validity in relation to the five-factor model of personality (Clark, 1993; Clark et al., in press; Clark, Vorhies, & McEwen, 1994; Reynolds & Clark, 2001), the Big Three personality traits (Clark, 1993), state and trait mood measures (Clark, 1993), other measures of personality disorder (Clark, Livesley, Schroeder, & Irish, 1996), and the MMPI-2 (Clark, 1993; Vittengl, Clark, Owen-Salter, & Gatchel, 1999).

BFI. The BFI 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 alpha coefficients of .84, .88, .82, .79, and .81, respectively, for the traits listed above in a sample of 711 English-speaking participants. They also reported good convergence with two other measures of the five-factor model.

EPQ-R. The EPQ-R is a 100-item measure that provides scores on three broad factors of personality: Neuroticism, Extraversion, and Psychoticism. The items are answered with a yes–no response format. Published internal consistency reliabilities range from .78 to .90 (Eysenck & Eysenck, 1991).

Data Analyses and Results

We conducted analyses (a) to assess test characteristics such as item and time savings associated with CAT administration, (b) to examine the psychometric comparability of scores derived from the SNAP-P&P and SNAP-CAT, and (c) to investigate which administration mode participants favored. For most of the analyses, participants in the two computerized and two P&P groups were collapsed into single groups, separately for Times 1 and 2, to test directly the effect of administration mode (i.e., P&P vs. computerized). Furthermore, as described above, within the computerized group we looked both at (a) full-scale scores (CF) based on complete administration of each scale by the computer and at (b) adaptive scores (CA) calculated when each participant met the termination criteria described above. Using both CF and CA scores permitted us to conduct comparisons that controlled for mode of administration.

We conducted all analyses on scores estimated on the IRT-based theta metric in order to eliminate the potential confound that can occur when scores are compared across different metrics (e.g., raw scores compared with thetas). Thus, for the P&P and CF scores, we estimated thetas for each participant on the basis of complete administration of each scale, whereas we estimated CA thetas using only the item responses that were collected prior to satisfaction of the termination criteria. All thetas were estimated with

Table 1
Item Savings, Completion Time Savings, and Loss of Psychometric Information in the Computerized Groups

Scale (items)	Mean no. of items					Mean completion time (min)						Mean psychometric information ^a					
	Time 1			Time 2		Time 1			Time 2			Time 1			Time 2		
	CF	CA	%sav	CA	%sav	CF	CA	%sav	CF	CA	%sav	CF	CA	%loss	CF	CA	%loss
Negative Temperament	28	14.4	49	14.5	48	1.70	0.86	49	1.42	0.72	50	13.6	9.8	28	13.1	9.5	27
Mistrust	19	11.8	38	12.1	36	1.23	0.75	39	1.02	0.64	37	7.7	6.0	22	7.4	5.9	21
Manipulativeness	20	13.0	35	12.5	38	1.53	0.91	40	1.19	0.69	43	6.2	5.3	14	6.1	5.2	15
Aggression	20	12.7	37	12.6	37	1.06	0.68	36	0.91	0.58	37	7.1	5.6	21	6.6	5.1	23
Self-Harm	16	12.1	24	12.4	23	0.87	0.63	28	0.73	0.54	26	7.7	6.6	15	7.6	6.5	15
Eccentric Perceptions	15	10.0	33	10.0	33	1.19	0.81	32	0.92	0.62	33	6.0	5.2	14	5.7	4.9	14
Dependency	18	11.2	38	11.3	37	1.14	0.71	38	0.95	0.60	36	6.3	5.4	15	6.1	5.2	15
Positive Temperament	26	12.8	51	12.5	52	1.54	0.75	51	1.29	0.62	52	8.7	6.5	25	8.2	6.2	25
Exhibitionism	16	10.3	36	10.6	34	0.92	0.59	36	0.76	0.50	34	6.9	5.9	15	6.6	5.6	14
Entitlement	16	10.4	35	10.4	35	0.94	0.55	42	0.74	0.43	42	6.3	5.7	9	6.1	5.6	9
Detachment	18	9.8	46	9.5	47	1.14	0.61	47	0.95	0.49	48	6.8	5.3	21	6.6	5.3	20
Disinhibition	35	17.9	49	17.6	50	2.51	1.20	52	2.02	0.94	54	8.4	6.2	26	8.4	6.2	26
Impulsivity	19	12.6	34	12.7	33	1.33	0.86	36	1.09	0.69	37	5.9	5.1	13	5.9	5.1	14
Propriety	20	18.3	9	17.2	14	1.46	1.35	8	1.16	1.00	14	5.3	5.2	2	4.9	4.7	4
Workaholism	18	12.3	32	12.0	33	1.19	0.83	30	0.99	0.67	32	6.3	5.6	11	6.0	5.3	11
Overall			36		37			38			38			17			17

Note. *ns* = 205 and 202 at Times 1 and 2, respectively. CF = computerized full scale; CA = computerized adaptive; %sav = percent savings; %loss = percent loss of information.

^a Mean psychometric information was calculated by averaging the inverse squares of participants' standard errors of measurement at the termination point.

expected a posteriori (Bock & Mislevy, 1982) methods;⁴ we calculated P&P thetas using BILOG, whereas we calculated CF and CA thetas using MicroCAT.

Efficiency Analyses

Analyses revealed significant overall completion 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 ($M = 25.7$ min, $SD = 6.7$) to Time 2 ($M = 21.5$ min, $SD = 6.0$), $t(412) = 22.5$ (for all *ts*, $p < .01$). At Time 1, mean completion times for the P&P, CF, and CA administrations were 30.7 ($SD = 5.0$), 20.6 ($SD = 3.6$), and 12.9 ($SD = 2.5$) min, respectively; all differences were significant, $t(411) = 23.27$, $t(411) = 45.13$, and $t(204) = 67.04$, for the P&P-CF, P&P-CA, and CF-CA comparisons, respectively. At Time 2, mean completion times for the P&P, CF, and CA administrations were 26.1 ($SD = 4.4$), 16.8 ($SD = 3.2$), and 10.4 ($SD = 2.2$) min, respectively; again, all differences were significant, $t(411) = 24.10$, $t(411) = 45.17$, and $t(204) = 66.14$, for the P&P-CF, P&P-CA, and CF-CA comparisons, respectively. Thus, the SNAP-CAT yielded overall time savings of 58.0% and 60.2% at Times 1 and 2, respectively, when compared with the SNAP-P&P, and 37.4% and 38.1% at Times 1 and 2, respectively, when compared with the full SNAP administered on the computer.

Savings and information loss were quantified at the scale level in the computerized groups. Results of these analyses appear in Table 1. Mean scale-level item savings were 36% and 37% at Times 1 and 2, respectively; in a similar manner, mean time savings were both 38%. As one might predict, the longest scales—Positive Temperament, Negative Temperament, and Disinhibition—yielded the greatest savings (around 50%). Propriety, how-

ever, was a clear outlier and consistently yielded the lowest savings (approximately 10%). Table 1 also includes estimates of mean psychometric information—calculated by averaging the inverse squares of participants' standard errors of measurement at the adaptive termination point (for the CA scoring) and at the end of each scale (for CF scoring)—as well as the proportion of information loss associated with CA administration. Mean information loss was 17% both at Time 1 and Time 2 (range = 2%–28%), suggesting that the adaptive algorithm yielded scale scores that were not as precise as their full-scale counterparts.

To assess the costs and benefits of adaptive administration more thoroughly, we conducted additional analyses in which we (a) calculated the efficiency of the CA and CF administration modes, at the scale level, by dividing psychometric information by the number of administered items (for item-keyed efficiency estimates) and per unit time (for time-keyed efficiency estimates) and (b) calculated the ratio of CA efficiency to CF efficiency to arrive at an estimate of relative efficiency (Hambleton et al., 1991). These results appear in Table 2 and show that CA administration, on average, yielded greater psychometric information per administered item (mean relative efficiency = 1.32 and 1.33 at Times 1 and 2, respectively) and per unit time (mean relative efficiency = 1.36 and 1.37 at Times 1 and 2, respectively) than did CF administration. Thus, although CA administration resulted in information

⁴ 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 certainly are possible with personality data, we chose to ameliorate this problem by using a Bayesian method known as expected a posteriori (EAP) theta estimation (Assessment Systems Corporation, 1996; Bock & Mislevy, 1982; Mislevy & Bock, 1990).

Table 2
Efficiency of Adaptive Versus Full-Scale Administration in the Computerized Groups

Scale	Mean psychometric information per item ^a						Mean psychometric information per minute ^a					
	Time 1			Time 2			Time 1			Time 2		
	E _{CF}	E _{CA}	RE _{CA/CF}	E _{CF}	E _{CA}	RE _{CA/CF}	E _{CF}	E _{CA}	RE _{CA/CF}	E _{CF}	E _{CA}	RE _{CA/CF}
Negative Temperament	0.49	0.68	1.39	0.47	0.66	1.40	8.02	11.39	1.42	9.25	13.31	1.44
Mistrust	0.41	0.51	1.24	0.39	0.49	1.26	6.29	8.04	1.28	7.33	9.17	1.25
Manipulativeness	0.31	0.41	1.32	0.30	0.42	1.40	4.06	5.85	1.44	5.11	7.58	1.48
Aggression	0.36	0.44	1.22	0.33	0.40	1.21	6.72	8.33	1.24	7.29	8.87	1.22
Self-Harm	0.48	0.54	1.13	0.47	0.52	1.11	8.82	10.48	1.19	10.43	11.93	1.14
Eccentric Perceptions	0.40	0.52	1.30	0.38	0.49	1.29	5.06	6.38	1.26	6.24	7.97	1.28
Dependency	0.35	0.48	1.37	0.34	0.46	1.35	5.55	7.59	1.37	6.45	8.66	1.34
Positive Temperament	0.34	0.51	1.50	0.32	0.49	1.53	5.69	8.73	1.53	6.41	10.01	1.56
Exhibitionism	0.43	0.57	1.33	0.41	0.53	1.29	7.51	10.03	1.34	8.62	11.20	1.30
Entitlement	0.39	0.55	1.41	0.38	0.54	1.42	6.69	10.48	1.57	8.25	13.12	1.59
Detachment	0.38	0.54	1.42	0.37	0.56	1.51	5.91	8.78	1.49	6.97	10.86	1.56
Disinhibition	0.24	0.35	1.46	0.24	0.35	1.46	3.35	5.22	1.56	4.17	6.61	1.59
Impulsivity	0.31	0.41	1.32	0.31	0.40	1.29	4.44	6.00	1.35	5.39	7.36	1.37
Propriety	0.26	0.28	1.08	0.25	0.27	1.08	3.63	3.85	1.06	4.23	4.71	1.11
Workaholism	0.35	0.46	1.31	0.33	0.44	1.33	5.28	6.74	1.28	6.05	7.94	1.31
<i>M</i>	0.37	0.48	1.32	0.35	0.47	1.33	5.80	7.86	1.36	6.81	9.29	1.37

Note. *ns* = 205 and 202 at Times 1 and 2, respectively. E_{CF} = computerized full-scale efficiency; E_{CA} = computerized adaptive efficiency; RE_{CA/CF} = relative efficiency.

^a Efficiency was calculated at the scale level by dividing psychometric information by the number of administered items (for item-keyed efficiency estimates) and per unit time (for time-keyed efficiency estimates). Relative efficiency was calculated as the ratio of CA efficiency to CF efficiency.

loss in an absolute sense, it is approximately 30%–35% more efficient than CF administration in terms of information gained per unit time or per administered item.

Item Presentation Analyses

We conducted analyses to determine the extent to which items in the original SNAP item pool were selected for administration by the adaptive algorithm. Space limitations do not permit a thorough description of these data; however, we present a brief overview of the results to demonstrate several themes culled from the analyses. For all SNAP-CAT items, we correlated item discrimination, item difficulty, the proportion of times in which the item was selected for administration, and the serial position of the item in the adaptive administration. Adaptive selection of items clearly led to some items being selected for administration more often than others: Both item discrimination ($r_s = .55$ and $.54$ at Times 1 and 2, respectively, $p < .001$) and item difficulty ($r_s = -.21$ and $-.20$ at Times 1 and 2, respectively, $p < .001$) significantly predicted how often items were administered adaptively. Thus, highly discriminating items and those lower in difficulty were selected more often than others. Item discrimination also predicted the serial position of item presentation ($r_s = -.62$ and $-.59$ at Times 1 and 2, respectively, $p < .001$); that is, items with higher discrimination values were more likely to be administered earlier in the test than those that discriminated less well, which is consistent with the maximum-information item selection strategy we used. Nevertheless, it also is clear that there was variability in the items administered across participants. Inspection of the item administration data revealed that the average SNAP-CAT item was administered adaptively to 62% ($SD = 31%$) of the participants. Furthermore,

fewer than 10% of SNAP-CAT items were administered adaptively to all participants, and only 2.5% of the items were never administered with the adaptive algorithm.

Psychometric Comparability

The most common way to define test-form equivalence operationally is through the concept of *psychometric equivalence*. As outlined by a number of authors (e.g., Ghiselli, Campbell, & Zedeck, 1981; Honaker, 1988) and in the recently revised *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], 1999), two tests can be considered psychometrically equivalent if they yield equal descriptive statistics and ranking of scores and if they correlate to the same degree with scores on other variables. If two forms of the same test meet these criteria, one can argue that both may share the same normative and validity data for the purposes of score interpretation (AERA, 1999; Honaker, 1988). However, strictly defined psychometric equivalence is very difficult to establish for short forms and adaptive tests designed to shorten a traditional P&P test with only a subset of the original scale items. The term *comparability* is more commonly used in such applications and, therefore, is adopted for the remainder of this article.

Mean-Level Comparability

To assess the effect of testing mode (i.e., SNAP-P&P vs. SNAP-CAT) on scores, we computed between- and within-mode effect sizes and confidence intervals, separately for Times 1 and 2. We calculated P&P-CF and P&P-CA effect sizes (i.e., cross-mode effect sizes) with the common methods described by Cohen

(1988), using the pooled standard deviation within each comparison to standardize the differences between group means; we calculated 99% confidence intervals for these effect sizes using methods described by Hedges and Olkin (1985). For the CF-CA comparisons (i.e., within-mode effect sizes), repeated-measures methods were used to account for the higher statistical power associated with such designs. We calculated CF-CA effect sizes and confidence intervals using formulas based on the *t* statistic resulting from paired *t* tests comparing the scale means within the computerized groups (see Dunlap, Cortina, Vaslow, & Burke, 1996, for a thorough discussion of repeated measures effect sizes and for derivation of the formulas we used).

Effect sizes and confidence intervals for all cross-mode and within-mode comparisons appear in Table 3. The cross-mode analyses revealed that only one scale—Self-harm—yielded mean theta differences whose 99% confidence intervals failed consistently (i.e., across Times 1 and 2, and for both P&P-CF and P&P-CA comparisons) to include zero, with CF and CA Self-harm thetas greater than P&P thetas at Times 1 and 2. The remaining cross-mode differences did not replicate across sessions, indicating that most likely they were due to random factors. The mean cross-mode effect sizes suggested slightly higher scores in the computerized groups on average, but these differences were well within 99% confidence intervals that included zero. Within-mode effect sizes comparing CF and CA administrations were very small, suggesting good comparability between full-scale and adaptively scored thetas while controlling for mode of administration. The 99% confidence intervals around the within-mode effect sizes were much tighter than those for the cross-mode comparisons—owing to the greater power associated with repeated-measures

designs—and thus several minor CF-CA differences were identified that likely are not practically significant (i.e., the largest absolute within-mode effect size was 0.07, which represents a *T* score difference of less than one point).

Rank-Order Comparability

We conducted test–retest and fidelity analyses both within and across administration modes; these results appear in Table 4. Test–retest correlations generally were high across administration modes. Mean retest correlations were .87 (range = .75–.93), .87 (range = .77–.90), and .84 (range = .73–.89) in the P&P, CF, and CA conditions, respectively, suggesting a slight retest decline for the adaptive testing condition. Mean cross-mode retest coefficients were .84 (range = .76–.90) and .83 (range = .77–.89) for the CF-P&P and CA-P&P comparisons, respectively. The cross-mode coefficients were nearly as high as the within-mode retest reliabilities; indeed, the mean disattenuated cross-mode coefficients (calculated with the within-mode retest reliabilities in the denominator) both were .97 for the CF-P&P and CA-P&P comparisons, suggesting good rank-order score comparability across administration modes. Finally, we calculated fidelity correlations at Times 1 and 2 within the computerized groups to compare adaptive with full-scale thetas. These coefficients, which essentially represent part-whole correlations, were uniformly high. Mean fidelity correlations were .98 (range = .94–.99) at both Time 1 and Time 2, suggesting good score recovery with the adaptive algorithm. Thus, to simplify presentation of the remaining analyses, only P&P and CA scores will be compared from this point forward.

Table 3
Effect Sizes and Confidence Intervals Comparing Mean-Level Differences Across Administration Modes

Scale	Paper-and-pencil minus computerized full scale				Paper-and-pencil minus computerized adaptive				Computerized full scale minus computerized adaptive			
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
	<i>d</i>	CI	<i>d</i>	CI	<i>d</i>	CI	<i>d</i>	CI	<i>d</i>	CI	<i>d</i>	CI
Negative Temperament	-0.06	±.25	-0.03	±.25	-0.07	±.25	-0.09	±.25	-0.02	±.04	-0.06	±.04
Mistrust	0.02	±.25	-0.15	±.25	0.00	±.25	-0.18	±.25	-0.02	±.04	-0.03	±.03
Manipulativeness	0.06	±.25	0.04	±.25	0.05	±.25	0.04	±.25	-0.01	±.04	0.00	±.04
Aggression	-0.21	±.25	-0.03	±.25	-0.20	±.25	0.02	±.25	0.01	±.04	0.05	±.06
Self-Harm	-0.72	±.26	-0.98	±.26	-0.75	±.27	-1.01	±.27	-0.02	±.05	0.01	±.05
Eccentric Perceptions	0.00	±.25	-0.25	±.25	-0.01	±.25	-0.24	±.25	-0.02	±.06	0.01	±.04
Dependency	-0.24	±.25	-0.12	±.25	-0.27	±.25	-0.10	±.25	-0.03	±.04	0.03	±.04
Positive Temperament	-0.05	±.25	-0.07	±.25	-0.05	±.25	-0.05	±.25	0.00	±.04	0.02	±.04
Exhibitionism	-0.03	±.25	0.03	±.25	-0.03	±.25	0.03	±.25	0.00	±.04	-0.01	±.04
Entitlement	0.01	±.25	0.09	±.25	-0.06	±.25	0.06	±.25	-0.05	±.03	-0.02	±.04
Detachment	0.02	±.25	-0.04	±.25	0.04	±.25	-0.09	±.25	0.01	±.05	-0.05	±.04
Disinhibition	0.06	±.25	0.00	±.25	0.11	±.25	0.06	±.25	0.06	±.05	0.07	±.05
Impulsivity	0.03	±.25	0.02	±.25	0.02	±.26	0.02	±.26	-0.01	±.04	0.00	±.04
Propriety	0.03	±.25	-0.30	±.25	0.03	±.25	-0.30	±.25	0.00	±.03	0.00	±.03
Workaholism	-0.27	±.25	-0.02	±.25	-0.30	±.25	-0.05	±.25	-0.03	±.03	-0.03	±.03
<i>M</i>	-0.09	±.25	-0.12	±.25	-0.10	±.25	-0.12	±.25	-0.01	±.04	0.00	±.04

Note. Paper-and-pencil *ns* = 208 and 211 at Times 1 and 2, respectively; computerized group *ns* = 205 and 202 at Times 1 and 2, respectively. Effect sizes were calculated with scores on the theta metric and reflect standardized differences between the two administration modes within each comparison. Positive effect sizes represent higher means in the first administration mode within each comparison, whereas negative effect sizes reflect higher means in the second administration mode. Effect sizes with 99% confidence intervals that do not include zero are presented in boldface. *d* = effect size; CI = 99% confidence interval.

Table 4
Test–Retest and Fidelity Correlations Across and Within Modes

Scale	Within-mode retest correlations			Cross-mode retest correlations		CF-CA fidelity correlations	
	PP ^a	CF ^b	CA ^b	CF-PP ^c	CA-PP ^c	Time 1 ^d	Time 2 ^e
Negative Temperament	.93	.87	.87	.81	.79	.98	.98
Mistrust	.84	.89	.83	.88	.87	.98	.98
Manipulativeness	.91	.86	.84	.86	.85	.98	.98
Aggression	.87	.77	.73	.86	.85	.96	.94
Self-Harm	.76	.86	.77	.83	.77	.96	.98
Eccentric Perceptions	.86	.78	.75	.82	.80	.94	.97
Dependency	.87	.87	.87	.76	.75	.99	.99
Positive Temperament	.86	.89	.89	.85	.83	.97	.98
Exhibitionism	.91	.90	.87	.87	.84	.99	.98
Entitlement	.75	.84	.81	.81	.81	.98	.98
Detachment	.84	.90	.88	.87	.86	.96	.97
Disinhibition	.91	.90	.88	.90	.89	.96	.95
Impulsivity	.88	.87	.86	.85	.84	.98	.98
Propriety	.87	.86	.86	.82	.82	.99	.99
Workaholism	.87	.86	.85	.83	.81	.99	.99
<i>M</i>	.87	.87	.84	.84	.83	.98	.98

Note. All correlations are significant ($p < .001$) and were calculated with scores on the theta metric. PP = paper and pencil; CF = computerized full scale; CA = computerized adaptive.

^a $n = 106$; ^b $n = 100$; ^c $n = 207$; ^d $n = 205$; ^e $n = 202$.

Analyses of Internal Structure

To assess internal structure, we conducted scale-level principal factor analyses, with varimax rotation, across testing modes. The scree plot, consistent with 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).⁵ The factor loadings, presented in Table 5, were reasonably comparable with those identified elsewhere (e.g., Clark, 1993; Clark et al., in press). Across administration modes and time, the SNAP scales formed three general factors that were labeled Negative Affectivity (NA), Positive Affectivity (PA), and Disinhibition Versus Constraint (DvC). Of the 180 factor loadings (3 factors \times 4 administrations \times 15 scales), only 22 (12.2%) “split,” loading greater than .30 on a second factor in addition to the one expected and, of these, in only 8 cases (4.4%) was the “off” loading higher than the expected loading (three for Manipulativeness, two each for Detachment and Workaholism, and one for Aggression). Moreover, in only 3 cases (1.7%; two for Dependency and one for Workaholism) did a scale fail to load on a factor as expected.

To quantify the structural similarity of these matrices, we computed congruence coefficients (Tucker, 1951). 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. In the present study, congruence coefficients were uniformly high. Cross-mode coefficients were .96, .97, and .93 at Time 1 and .95, .94, and .93 at Time 2 for the NA, PA, and DvC factors, respectively. In addition, congruence coefficients were computed between Time 1 and Time 2 factor loadings, yielding uniformly good cross-session congruence (all coefficients were greater than .97) for all factor matrices. Thus, the SNAP’s internal covariance structure appeared to replicate well across testing modes and sessions.

Convergent and Discriminant Validity

To assess the similarity of the convergent and discriminant validity patterns across modes, we computed correlations between the SNAP scales and two established measures of personality—the BFI and EPQ–R—which have been shown previously to correlate meaningfully with the SNAP (e.g., Clark, 1993; Clark et al., in press; Clark et al., 1994; Reynolds & Clark, 2001). Correlations with the BFI and EPQ–R appear in Tables 6 and 7, respectively. SNAP scales correlated similarly with BFI scales across testing modes, and the correlations were orderly and interpretable in the context of previous SNAP studies. As expected, the BFI Neuroticism scale correlated significantly with five of the seven scales comprising the SNAP NA factor, including, of course, Negative Temperament, but also Mistrust, Aggression, Self-harm, and Dependency, as well as Positive Temperament (negatively). BFI Extraversion correlated strongly with three of the four PA scales, specifically Positive Temperament, Exhibitionism, and Detachment (negatively). BFI Agreeableness correlated most consistently with Aggression, Mistrust, Manipulativeness, and Detachment (all negatively). Finally, BFI Conscientiousness was related to all four scales comprising the SNAP DvC factor, plus Manipulativeness and Self-harm (negatively). Correlations between the SNAP and

⁵ An important difference between the present factor analyses and those typically conducted with SNAP scales is that the version of Disinhibition used here contained overlapping items with several other 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 in the present study because we calculated thetas on subsets of Disinhibition items that varied across individuals, making it impossible to extract the overlapping variance. Despite this difference, the factor solutions reported here did not differ markedly from those reported previously (e.g., Clark, 1993; Clark et al., in press).

Table 5
Varimax Factor Loadings of Schedule for Nonadaptive and Adaptive Personality Scales on Three Principal Factors

Scale	Negative Affectivity				Positive Affectivity				Disinhibition vs. Constraint			
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA
Negative Temperament	.71	.70	.65	.70	-.06	.02	-.15	-.06	-.07	.04	-.05	-.08
Mistrust	.77	.61	.69	.79	-.08	-.08	-.03	-.06	.03	.19	.01	.07
Manipulativeness	.46	.32	.63	.43	.30	.08	.39	.12	.58	.67	.37	.62
Aggression	.62	.35	.48	.55	.03	.00	.08	-.07	.21	.37	.15	.25
Self-Harm	.56	.51	.53	.50	-.31	-.27	-.29	-.30	.23	.26	.22	.23
Eccentric Perceptions	.43	.37	.34	.47	.13	.17	.13	.14	.09	.02	.02	.02
Dependency	.24	.36	.39	.22	-.04	-.06	-.13	-.03	.00	-.01	-.03	.03
Positive Temperament	-.31	-.34	-.31	-.19	.72	.69	.74	.74	-.15	-.09	-.21	-.10
Exhibitionism	.00	-.12	.06	-.09	.63	.55	.62	.60	.23	.29	.09	.34
Entitlement	.10	-.01	.13	-.04	.60	.42	.54	.59	-.13	.04	-.29	.02
Detachment	.44	.49	.46	.55	-.50	-.39	-.52	-.38	-.09	-.03	-.11	-.07
Disinhibition	.27	.12	.45	.25	.15	.04	.27	.07	.91	.88	.79	.87
Impulsivity	.17	.03	.24	.05	.16	.05	.15	.15	.78	.69	.80	.76
Propriety	.12	.14	.07	.07	.25	.31	.21	.28	-.59	-.43	-.61	-.46
Workaholism	.06	.13	.08	.22	.34	.36	.24	.39	-.53	-.15	-.61	-.37

Note. Factor loadings above ±.30 are presented in boldface. Analyses are based on scores calculated with the theta metric. Disinhibition includes overlapping variance with several other scales. PP = paper and pencil (*ns* = 208 and 211 at Times 1 and 2, respectively); CA = computerized adaptive (*ns* = 205 and 202 at Times 1 and 2, respectively).

EPQ-R were orderly and meaningful as well. For example, EPQ-R Neuroticism correlated significantly with four of the SNAP NA factor scales: Negative Temperament, Mistrust, Aggression, and Self-harm. EPQ-R Extraversion correlated with three of the four SNAP PA factor scales, including Positive Temperament, Exhibitionism, and Detachment (negatively). Finally, EPQ-R Psychoticism was related to three of the four DvC factor scales (Disinhibition, Impulsivity, and Propriety, negatively) as

well as Manipulativeness and, to a lesser extent, Mistrust, Aggression, and Self-harm.

It is important to note that, although the convergent and discriminant pattern appears to be reasonably comparable across administration modes, the adaptive algorithm appears, on average, to have attenuated the convergent correlations. To examine the degree of validity loss, we (a) identified all variable combinations in which at least one mode yielded a correlation greater than or

Table 6
Correlations Between Schedule for Nonadaptive and Adaptive Personality Scales and Big Five Inventory Scales Across Administration Modes

Scale	Neuroticism				Extraversion				Openness				Agreeableness				Conscientiousness			
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA
Negative Temperament	.83	.71	.79	.74	-.26	-.25	-.27	-.25	-.09	-.11	-.20	-.10	-.37	-.24	-.33	-.28	-.22	-.13	-.21	-.11
Mistrust	.45	.37	.36	.42	-.31	-.25	-.20	-.30	.03	-.15	-.05	-.04	-.49	-.39	-.47	-.47	-.22	-.22	-.22	-.24
Manipulativeness	.09	.17	.14	.09	.08	-.09	-.03	.02	.00	-.09	-.03	.00	-.45	-.48	-.50	-.46	-.43	-.39	-.42	-.40
Aggression	.43	.38	.37	.45	-.11	-.12	-.05	-.15	-.03	-.11	-.08	.00	-.66	-.59	-.61	-.55	-.24	-.14	-.08	-.33
Self-Harm	.42	.37	.38	.35	-.36	-.29	-.27	-.31	-.02	-.07	-.17	.09	-.35	-.28	-.28	-.29	-.45	-.35	-.37	-.34
Eccentric Perceptions	.14	.10	.05	.11	-.04	-.04	-.08	-.14	.28	.18	.31	.18	-.21	.12	.00	-.21	-.20	-.10	-.17	-.08
Dependency	.34	.34	.33	.33	-.10	-.19	-.25	-.04	-.29	-.31	-.29	-.29	.04	.12	-.03	.18	-.22	-.13	-.20	-.12
Positive Temperament	-.37	-.44	-.39	-.38	.68	.66	.62	.65	.35	.37	.42	.22	.23	.35	.30	.29	.36	.27	.30	.29
Exhibitionism	-.10	-.11	-.17	-.11	.53	.48	.47	.55	.17	.21	.19	.19	-.15	-.08	-.10	-.06	-.10	-.02	-.11	-.03
Entitlement	-.07	-.25	-.13	-.28	.27	.28	.23	.33	.17	.25	.19	.15	-.04	-.04	-.04	-.01	.23	.14	.19	.22
Detachment	.26	.31	.34	.32	-.71	-.68	-.73	-.69	-.04	-.04	-.21	.05	-.33	-.43	-.40	-.48	-.16	-.13	-.07	-.16
Disinhibition	.00	.05	.04	-.01	.10	.07	.05	.09	.02	-.03	-.04	.00	-.35	-.27	-.40	-.25	-.67	-.54	-.61	-.55
Impulsivity	.04	.03	.01	-.02	.19	.20	.11	.27	.06	.03	.03	.15	-.23	-.12	-.27	-.14	-.63	-.44	-.59	-.48
Propriety	.12	-.04	.04	-.05	.00	.09	.08	.07	-.04	-.09	-.03	-.10	.09	.27	.13	.14	.48	.30	.39	.38
Workaholism	.02	.03	.04	-.03	.08	.10	.07	.11	.18	.17	.08	.19	.06	-.10	.02	.01	.55	.45	.55	.49

Note. Correlations above ±.18 are significant (*p* < .01). Correlations above ±.30 are presented in boldface. Analyses are based on scores calculated with the theta metric. PP = paper and pencil (*ns* = 204 and 207 at Times 1 and 2, respectively); CA = computerized adaptive (*ns* = 202 and 199 at Times 1 and 2, respectively).

Table 7

Correlations Between Schedule for Nonadaptive and Adaptive Personality Scales and Eysenck Personality Questionnaire-Revised Scales Across Administration Modes

Scale	Neuroticism				Extraversion				Psychoticism			
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA	PP	CA
Negative Temperament	.82	.75	.86	.81	-.23	-.25	-.24	-.22	.14	.13	.21	.08
Mistrust	.54	.47	.42	.61	-.21	-.26	-.17	-.25	.31	.35	.43	.31
Manipulativeness	.26	.20	.27	.25	.18	.00	.12	.02	.52	.52	.55	.51
Aggression	.40	.23	.25	.44	-.07	-.09	-.02	-.11	.39	.32	.38	.31
Self-Harm	.41	.45	.41	.38	-.29	-.30	-.28	-.30	.35	.39	.38	.37
Eccentric Perceptions	.34	.30	.29	.32	.04	.01	.01	.00	.22	.19	.19	.25
Dependency	.42	.34	.44	.34	-.08	-.23	-.18	-.13	-.04	-.02	.03	-.09
Positive Temperament	-.28	-.31	-.31	-.28	.67	.72	.70	.71	-.17	-.17	-.24	-.13
Exhibitionism	-.06	-.10	-.06	-.08	.56	.49	.53	.58	.11	.15	.09	.19
Entitlement	.01	-.12	-.04	-.13	.34	.19	.26	.32	-.04	.15	-.01	.06
Detachment	.23	.27	.27	.32	-.64	-.64	-.70	-.61	.18	.27	.25	.22
Disinhibition	.15	.13	.13	.13	.24	.15	.23	.16	.68	.63	.73	.67
Impulsivity	.09	.05	.03	.05	.29	.22	.25	.28	.56	.52	.64	.53
Propriety	.09	.05	.14	.02	.07	.13	.08	.16	-.52	-.50	-.50	-.55
Workaholism	-.04	.10	.04	.05	.07	.09	.04	.10	-.26	-.03	-.25	-.11

Note. Correlations above $\pm .18$ are significant ($p < .01$). Correlations above $\pm .30$ are presented in boldface. Analyses are based on scores calculated with the theta metric. PP = paper and pencil ($n_s = 207$ and 208 at Times 1 and 2, respectively); CA = computerized adaptive ($n_s = 203$ and 202 at Times 1 and 2, respectively).

equal to .30 (i.e., those presented in boldface in Tables 6 and 7), (b) squared the correlations for those variable combinations (at Times 1 and 2) to determine the amounts of criterion variance explained by the SNAP scales, and (c) divided CA variance proportions by P&P variance proportions to arrive at an index of relative concurrent validity. On average, the results revealed that CA administration yielded small to moderate validity losses at Time 1, with the CA administration accounting for 24.6% and 10.9% less of the explained variance for scales of the BFI and EPQ-R, respectively, compared with the P&P mode. These differences in explained variance essentially disappeared at Time 2 with the CA administration accounting for 6.6% less and 10.3% more, respectively, of the variance explained by the P&P mode.

Mode Preferences

Participants in the cross-mode groups were asked at the end of the study which administration mode they preferred and why. Of the 207 participants in the cross-mode conditions, 87% preferred the computerized condition, citing speed (49%) and ease of use (24%) as the primary reasons. The remaining 13% of participants preferred the SNAP-P&P version, citing dislike of not being able to change answers on the computerized version (64%) as the primary reason.

Discussion

The primary objective of this study was to extend the computerized adaptive personality literature by introducing and validating an IRT-based computerized adaptive version of a multiscale personality measure—the SNAP-CAT—with live research participants. With number of items held constant, computerized administration had little effect on descriptive statistics, rank ordering of

scores, reliability, and concurrent validity. Adaptive administration yielded significant time and item savings, achieved with small losses of absolute reliability and concurrent validity. However, as would be expected with a properly calibrated CAT, the SNAP-CAT was relatively more efficient than the full-scale SNAP, yielding greater measurement precision per unit time and per item administered. Descriptive statistics, test-retest stability, internal factor structure, and the convergent and discriminant validity pattern largely were comparable across administration modes. Although some differences were identified, only one (on Self-harm) replicated across sessions. Overall, participants preferred the computerized version to the P&P version. Thus, the SNAP-CAT seems to have promise as an alternative to traditional SNAP administration. Moreover, the present data suggest that CAT, in general, is a powerful technological advance in testing that can be reliably and validly applied to personality assessment when measurement efficiency is an important consideration.

Efficiency Gains

The SNAP-CAT yielded significant savings compared with the P&P version and full-scale administration on the computer. CAT administration of the SNAP was roughly 60% faster than the traditional P&P version. When controlling for computerized administration, time and item savings still amounted to more than 36% on average, with greater savings of approximately 50% for the longest scales on the SNAP. These savings are consistent with those identified in simulation studies with the SNAP item pool (Simms, 2004) as well as previous IRT-based CAT simulations (Kamakura & Balasubramanian, 1989; Reise & Henson, 2000; Waller, 1999; Waller & Reise, 1989). Furthermore, these savings are greater than those typically found for non-IRT CAT applica-

tions such as those created with the MMPI item pool (Handel et al., 1999; Roper et al., 1991, 1995). Time and item savings were achieved concomitantly with a small but noteworthy reduction in absolute measurement precision. We discuss possible remedies for precision loss of this variety subsequently.

In Reise and Henson's (2000) simulation results with the facet scales of the NEO-PI-R, they found little variability in terms of the items presented within each adaptive facet scale, suggesting that their algorithm resulted in little more than computerized short forms of the facet scales. We also assessed this possibility with the SNAP-CAT. Although item characteristics such as discrimination and difficulty clearly predicted when and if items were adaptively presented—with items higher in discrimination and lower in difficulty being presented more often than others—there was significant variability in the items presented across participants, in contrast to Reise and Henson's findings.

A key difference between the SNAP and NEO-PI-R that may account for the discrepant findings is scale length. All facet scales of the NEO-PI-R include eight items, whereas SNAP scales are approximately two to four times longer ($M = 2.5$). Larger item pools provide greater flexibility for adaptive testing because they often include more variability in item difficulty and discrimination. Although SNAP scales clearly are not as long as would be ideal for an adaptive testing application (e.g., Sands et al., 1997; Weiss, 1985), they clearly were long enough to yield truly adaptive scales. Moreover, if the sample had included individuals with a broader range of personality pathology, it is likely that we would have found even more variability in item administration.

A clear outlier with respect to efficiency gains was the Propriety scale. Analyses of participant termination data revealed that 82% of participants terminated within three items of the maximum for the scale. Post hoc analyses suggested that the reason for this anomaly is that the standard error portion of the termination criterion (i.e., stop administering items after the standard error drops below 0.40) was never satisfied for Propriety. Instead, participants were continually administered items until minimally informative items no longer existed in the item pool (i.e., the other main part of the termination algorithm). In IRT, the standard error varies across the trait continuum and is inversely related to test information (Embretson, 1996; Hambleton et al., 1991). Thus, the inability of Propriety to reach the standard error threshold can be attributed to relatively poor psychometric information (i.e., lower discrimination values) afforded by the 20 items that constitute the scale. Future studies could be conducted to identify new items for Propriety that discriminate better or, at minimum, add incremental information that would translate into lower standard errors. Meanwhile, it may be best simply to administer the complete scale, as adaptive administration would save only about 30 s in most cases if Propriety behaved like other SNAP-CAT scales.

Psychometric Equivalence

Mean-level analyses revealed largely comparable descriptive statistics across modes of administration. Only one significant difference replicated across sessions: Computerized Self-harm scores—both adaptive and full-scale—were significantly higher than P&P. The mode- and scale-specific nature of these findings suggests a possible Scale \times Mode interaction with respect to theta estimation. Both the BILOG and MicroCAT manuals cite Bock and Mislevy (1982) as the basis for their EAP theta estimation

procedures, which suggests that the programs should yield highly comparable results; however, a possibility exists that the programs have slightly different algorithms for dealing with extreme scores. Indeed, post hoc analyses revealed data consistent with such a hypothesis for the Self-harm scale (but no other SNAP scales). On the raw score metric, we first observed that Self-harm is significantly more skewed than all other SNAP scales: At Times 1 and 2, respectively, skewness = 2.04 and 1.90 in the computerized group and 1.80 and 2.70 in the P&P group. In contrast, for the remaining SNAP scales, mean skewness = 0.35 (range = -0.83 – 1.12) and 0.31 (range = -0.97 – 1.38), respectively, in the computerized group, and 0.37 (range = -0.76 – 1.49) and 0.39 (range = -0.88 – 1.25), respectively, in the P&P group. Next, on the theta metric we found a clear floor difference for Self-harm scores across modes: 78 (38%) and 83 (41%) of CF thetas were clustered at a theta floor of -0.44 at Times 1 and 2, respectively, whereas 76 (37%) and 103 (49%) of P&P thetas were clustered at a theta floor of -1.09 at Times 1 and 2, respectively. These floor scores were modal in each subgroup, and those participants scoring at the theta floor were exactly those who received a score of zero on the raw score metric. Thus, it appears that differential handling of zero scores led directly to the mean-level differences observed with Self-harm, and this problem likely can be ameliorated in a revised SNAP-CAT by attending more carefully to the algorithm used for handling zero scores. If, however, problems associated with Self-harm's skewness and poor model fit cannot be resolved, the best solution may be to administer Self-harm nonadaptively.

The few remaining cross-mode mean score differences were 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) that suggested that impression management tends to be lower with computerized personality test administration. However, Dwight and Feigelson (2000) also reported that the magnitude of this effect has decreased markedly in more recent studies, which suggests that differences associated with computerized administration have waned over the years as participants have become more familiar and comfortable with computers.

Rank ordering of scores was quite stable across modes of administration. On average, the computerized adaptive retest correlations ($M = .84$) were only slightly lower than those obtained from the computerized full-scale ($M = .87$) and P&P ($M = .87$) administrations. Furthermore, the cross-mode retest coefficients ($M_s = .84$ and $.83$) were nearly equivalent to the retest reliabilities. Retest and cross-mode coefficients associated with IRT-based personality CATs 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 we found. Finally, results of the fidelity analyses in the computerized groups suggest that one can achieve good score recovery (mean $r = .98$) while administering significantly fewer items.

Internal structural analyses revealed comparable loadings for all scales across modes of presentation. The three-factor structure found in these data are highly consistent with that identified previously (e.g., Clark, 1993; Clark et al., 1996). In particular, even the two scales that split significantly across factors in these analyses—Manipulativeness and Detachment—have shown simi-

lar evidence of splitting elsewhere (Clark, 1993; Clark et al., in press). Thus, the three-factor structure of the SNAP appears to be robust to variations in item presentation caused by our adaptive algorithm.

In a similar manner, the convergent and discriminant validity patterns were reasonably comparable across administration modes and were consistent with relations identified in other studies of the SNAP (e.g., Clark, 1993; Clark et al., in press). However, the adaptive algorithm resulted in a small but notable reduction in the strength of the concurrent correlations with the BFI and EPQ-R, especially for the data collected at the first session. This validity reduction likely is due specifically to the adaptive administration, but also could be related to differences in the amount of shared method variance across groups (recall that the BFI and EPQ-R were completed with P&P). Regardless, this finding—as well as the information loss described earlier—represent potential costs associated with the SNAP-CAT, and perhaps adaptive personality testing more generally, that deserve consideration in future CAT projects of this variety.

In particular, adaptive conversions of traditional personality measures—in which the adaptive version essentially represents an individually tailored short form of the original item pool—are likely to result in some reduction in precision and validity. However, losses such as these likely could be reduced significantly, and perhaps completely, if adaptive personality tests were constructed in a manner similar to those in the ability testing domain. Specifically, adaptive ability tests typically include much larger pools of items than are available on traditional personality measures, and these pools more uniformly sample the full range of the traits under consideration. Future research is needed to develop and assess the impact of adaptive personality assessment derived from larger item pools that more fully sample and discriminate across the entire trait continuum.

A limitation of our convergent and discriminant validity data are that we included supplemental measures assessing only broad, higher order dimensions of personality. The SNAP includes scales representing at least two levels of analysis: (a) broad scales such as Negative Temperament, Positive Temperament, and Disinhibition and (b) lower order scales such as Aggression, Detachment, and Propriety. We did not include scales that tap the more specific variance associated with the lower order SNAP scales. If certain items within a scale are more likely to be adaptively administered than others, subtle content biases may result that alter the meaning of the resultant scores. Thus, data comparing concurrently the SNAP and SNAP-CAT with more specific measures relevant to personality pathology are needed before firm conclusions are drawn regarding comparability of the SNAP-CAT with the traditional SNAP.

Participants clearly had different experiences of and reactions to the two administration modes. 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 and easier to complete than the SNAP-P&P. That participants enjoyed the computerized form better is consistent with a number of recent studies (Vispoel, 2000; Vispoel et al., 2001) examining equivalence between P&P and computerized versions of some common measures of self-concept and self-esteem. In these studies, participants generally preferred the computerized forms and rated them as easier to read, more comfortable and enjoyable to complete, and less fatiguing than the

P&P versions. In the present study, as in those by Vispoel and colleagues, all participants were college students, who might be expected to be quite familiar and comfortable with computers. Thus, generalizability of these findings to other populations (e.g., psychiatric patients or older persons) needs to be established through future studies.

However, in both of Vispoel's studies (Vispoel, 2000; Vispoel et al., 2001), data indicated that their computerized versions took participants longer to complete than the P&P versions, which runs contrary to our findings. Characteristics of both studies likely led to this discrepancy. Participants in Vispoel's studies recorded P&P responses in the same booklet as the items and provided computerized responses using a multiscreen presentation format in which multiple items were presented per screen. In contrast, P&P participants in our study recorded their answers on separate machine-scorable answer sheets that had to be marked completely (which is very common in applied clinical assessment settings). Having to switch between a booklet and answer sheet is more time consuming than simply marking answers in the item booklet itself and likely contributed to the significant time advantage we identified for computerized assessment in general. Thus, our data comparing completion times across P&P and computerized forms should be interpreted with these caveats.

Conclusions and Future Directions

The results of the present study suggest that the prototype SNAP-CAT is a largely comparable form of the traditional paper-and-pencil SNAP. The SNAP-CAT, with the notable exception of Self-harm, yielded reasonably comparable descriptive statistics, rank ordering of scores, internal correlational structure, and convergent and discriminant validity. Although the adaptive algorithm resulted in some reduction in precision and validity, it resulted in shortened testing time, significant item savings, and greater relative efficiency than the full SNAP. In light of calls for more efficient measures of personality and psychopathology, the SNAP-CAT appears to be a viable alternative to traditional testing. Moreover, CAT methodology shows great promise for personality assessment in general, especially for those measures that meet the assumption of unidimensionality associated with most IRT models.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Assessment Systems Corporation. (1996). *User's manual for the Micro-CAT testing system* (3rd ed.). St. Paul, MN: Author.
- Benet-Martinez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait method analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75, 720–750.
- 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.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In L. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 358–472). Reading, MA: Addison Wesley.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability

- in a microcomputer environment. *Applied Psychological Measurement*, 6, 431–444.
- 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., Perry, J. N., & Atlis, M. M. (2000). Validity and utility of computer-based test interpretation. *Psychological Assessment*, 12, 6–18.
- 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. (in press). *Schedule for Nonadaptive and Adaptive Personality—Second Edition (SNAP-2): Manual for administration, scoring, and interpretation*.
- 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.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Costa, P. T., Jr., & McCrae, R. R. (1992). *The NEO-PI-R: The Revised NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.
- Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, 1, 170–177.
- 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.
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8, 341–349.
- 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.
- Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: Freeman.
- Gough, H. G. (1975). *California Psychological Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.
- Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Handel, R. W., Ben-Porath, Y. S., & Watt, M. (1999). Computerized adaptive assessment with the MMPI-2 in a clinical setting. *Psychological Assessment*, 11, 369–380.
- Hansen, J. C., Neuman, J. L., Haverkamp, B. E., & Lubinski, B. R. (1997). Comparison of user reaction to two methods of Strong Interest Inventory administration and report feedback. *Measurement & Evaluation in Counseling & Development*, 30, 115–127.
- Hathaway, S. R., & McKinley, J. C. (1951). *The Minnesota Multiphasic Personality Inventory manual* (rev.). New York: Psychological Corporation.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Honaker, L. M. (1988). The equivalency of computerized and conventional MMPI administration: A critical review. *Clinical Psychology Review*, 8, 561–577.
- 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.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York: Guilford.
- 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.
- 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.
- 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.
- Morey, L. C. (1991). *Personality Assessment Inventory*. Odessa, FL: Psychological Assessment Resources.
- 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* (pp. 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.
- 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 testing with the MMPI-2. *Journal of Personality Assessment*, 65, 358–371.
- Russell, E. W. (2000). The application of computerized scoring programs to neuropsychological assessment. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (pp. 483–515). Mahwah, NJ: Erlbaum.
- Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- Simms, L. J. (2004). *Simulated psychometric efficiency of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality*. Manuscript in preparation.
- Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, 12, 102–111.
- Snyder, D. K. (2000). Computer-assisted judgement: Defining strengths and liabilities. *Psychological Assessment*, 12, 52–60.
- Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.
- 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. (2000). Computerized versus paper-and-pencil assessment of self-concept: Score comparability and respondent preferences. *Measurement and Evaluation in Counseling and Development*, 33, 130–143.

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.

Vittengl, J. R., Clark, L. A., Owen-Salter, 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.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.

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

Waller, N. G. (1999). Searching for structure in the MMPI. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 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.

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

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

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, 15, 337–362.

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* (pp. 141–158). Washington, DC: American Psychological Association.

Appendix A

Formula for the Item Response Function

The formula for the two-parameter logistic model (Birnbaum, 1968) is 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.

Appendix B

Descriptive Statistics for SNAP-CAT Item Parameters by Scale

Scale	No. of items		Item discrimination (a)				Item difficulty (b)			
	Total	Min	M	SD	Min	Max	M	SD	Min	Max
Negative Temperament	28	14	0.98	0.22	0.49	1.42	-0.07	0.61	-1.36	1.14
Mistrust	19	8	0.90	0.30	0.55	1.53	0.57	0.49	-0.76	1.50
Manipulativeness	20	9	0.78	0.27	0.10	1.24	1.26	1.01	0.17	4.53
Aggression	20	10	0.93	0.23	0.57	1.36	1.10	0.45	0.14	1.93
Self-Harm	16	8	1.15	0.31	0.71	1.80	1.08	0.62	-0.06	2.28
Eccentric Perceptions	15	7	0.87	0.30	0.37	1.29	0.84	0.64	-0.40	1.76
Dependency	18	8	0.85	0.31	0.40	1.36	0.96	0.85	-0.48	2.56
Positive Temperament	26	10	0.84	0.33	0.39	1.64	-0.72	0.58	-2.06	0.33
Exhibitionism	16	7	0.88	0.30	0.32	1.40	0.17	0.91	-1.66	1.86
Entitlement	16	7	0.82	0.42	0.16	1.50	0.21	0.88	-1.00	2.22
Detachment	18	7	0.96	0.40	0.36	1.89	0.77	0.85	-0.74	2.45
Disinhibition	35	15	0.62	0.20	0.29	1.21	0.81	0.68	-0.60	2.34
Impulsivity	19	7	0.73	0.33	0.30	1.41	0.63	0.41	-0.09	1.40
Propriety	20	9	0.62	0.15	0.38	0.88	-0.33	0.98	-1.43	2.61
Workaholism	18	7	0.80	0.34	0.41	1.56	0.39	1.03	-1.54	2.42

Note. Item parameters are based on calibration sample ($N = 3,995$). Min = minimum; Max = maximum.

Received June 4, 2003
 Revision received August 30, 2004
 Accepted September 7, 2004 ■