Psychological Assessment
2007, Vol. 19, No. 1, 14–24

1040-3590/07/$12.00 DOI: 10.1037/1040-3590.19.1.14

Computerized Adaptive Personality Testing: A Review and Illustration With the MMPI–2 Computerized Adaptive Version

Johnathan D. Forbey
Ball State University

Yossef S. Ben-Porath
Kent State University

Computerized adaptive testing in personality assessment can improve efficiency by significantly reducing the number of items administered to answer an assessment question. Two approaches have been explored for adaptive testing in computerized personality assessment: item response theory and the countdown method. In this article, the authors review the literature on each and report the results of an investigation designed to explore the utility, in terms of item and time savings, and validity, in terms of correlations with external criterion measures, of an expanded countdown method-based research version of the Minnesota Multiphasic Personality Inventory—2 (MMPI–2), the MMPI–2 Computerized Adaptive Version (MMPI–2–CA). Participants were 433 undergraduate college students (170 men and 263 women). Results indicated considerable item savings and corresponding time savings for the adaptive testing modalities compared with a conventional computerized MMPI–2 administration. Furthermore, computerized adaptive administration yielded comparable results to computerized conventional administration of the MMPI–2 in terms of both test scores and their validity. Future directions for computerized adaptive personality testing are discussed.

Keywords: computerized adaptive testing, computerized test administration, MMPI–2, predictive validity

Computers have long been used in the administration, scoring, and interpretation of psychological tests (Ben-Porath & Butcher, 1986). In the area of personality and psychopathology assessment, computer technology has for the most part been used to administer self-report measures in their standard format, score them, and generate automated interpretations. These applications, although providing significant improvements in reducing error and the amount of time needed to administer, score, and interpret tests, do not take full advantage of the opportunities offered by computer technology. Computerized adaptive (CA) testing is a relatively recent innovation that opens up new possibilities in personality assessment.

In general, the CA administration of a test involves administering only those items needed to answer a referral question (Waller & Reise, 1989). Optimally, adaptive testing reduces the number of items and time required to administer a test without attenuating test validity. Such an approach to assessment serves to reduce the burden on the test taker in terms of time and tedium associated with extended testing, as well the costs associated with supervision of testing.

CA testing has been used extensively in ability and achievement testing (e.g., state licensure exams for nursing, the Graduate Record Examination); however, its use in personality assessment remains uncommon (Reise & Henson, 2003). This may reflect difficulties in adopting item response theory (IRT), the most common approach to CA ability and achievement testing, for use in personality assessment. The IRT approach to adaptive testing is designed to locate an individual's standing on a latent trait, theta (θ). Items are typically administered on the basis of an algorithm that considers its discrimination (α) and difficulty (β) parameters (Waller & Reise, 1989). For unidimensional constructs, such as ability (e.g., the SAT and Graduate Record Examination), IRT approaches have been applied widely (Reise & Henson, 2003).

IRT-based adaptive testing also has been explored in the area of health assessment. For example, in a simulation study, Ware, Gandek, Sinclair, and Bjorner (2005) found that IRT-based CA estimates were nearly equivalent to full-scale administration estimates of outcome measures in assessing outcome in physical rehabilitation. Lai, Cella, Chang, Bode, and Heinemann (2003) reported mixed results in applying IRT-based adaptive testing to identifying fatigue among cancer patients (82.6%) versus the general population (66.8%) with a nine-item self-report scale.

In contrast to assessment of ability, aptitude, and health, efforts to develop IRT-based approaches to CA personality assessment have met with limited success to date (Reise & Henson, 2000, 2003). Several studies have been conducted exploring the application of IRT to personality and psychopathology measures. Most have only investigated the feasibility of applying this method to personality testing without evaluating its impact on test score validity. For example, using IRT-based approaches, several researchers have explored the item pools of specific instruments such as the Beck Depression Inventory (BDI; Hammond, 1995; Santor, Ramsay, & Zuroff, 1994) and the Hare Psychopathology Checklist (Cooke & Michie, 1997) to examine how well the existing items of these measures "fit" an IRT model applied to the

Johnathan D. Forbey, Department of Psychological Science, Ball State University, and Yossef S. Ben-Porath, Department of Psychology, Kent State University.

This research, including development of the research version of the MMPI–2–CA, was supported by a grant from the University of Minnesota Press, publisher of the MMPI–2. Portions of this study were presented at the 2005 MMPI–2 Workshops and Symposia, Fort Lauderdale, FL.

Correspondence concerning this article should be addressed to Johnathan D. Forbey, Department of Psychological Science, North Quad 121, Ball State University, Muncie, IN 47304. E-mail: jdforbey@bsu.edu

latent traits or underlying constructs assessed by the measures. Fraley, Waller, and Brennan (2000) examined the psychometric properties of several measures of adult attachment and concluded that they could be improved through application of IRT models. These studies have focused on item fit for specific instruments or scales, rather than CA testing per se. In at least one such investigation, Chernyshenko, Stark, Chan, Drasgow, and Williams (2001) concluded that there were significant obstacles to applying IRT models to personality measures such as the 16PF and Goldberg's Big Five personality measure.

Validity Studies of IRT Models in CA Personality Assessment

Reise and Henson (2000) explored the scale score intercorrelations of simulated IRT-based administration of the NEO Personality Inventory—Revised (NEO-PI–R) in a sample of 1,059 college students. Relying on real-data simulations (in which responses to conventional paper-and-pencil test administration are used to simulate responses to an adaptive test administration), Reise and Henson reported that an IRT-based CA administration of the NEO-PI–R could accurately reproduce facet scores, although the variance of the latent trait scores was greatly reduced, as typically only four of the eight items of each facet scale were required to produce an accurate estimation of the original facet scale score. As with other IRT-based studies of personality assessment, no extratest correlates were reported to examine the impact of using IRT methodology on test score validity.

A recent study by Simms and Clark (2005) is the first published investigation of the impact of IRT-based adaptive personality assessment on test score validity. With 491 undergraduate college students, Simms and Clark applied an IRT procedure to the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993) and compared time and item savings as well as the validity of an IRT-based CA version and paper-and-pencil version of the SNAP. Extratest criterion measures included the Big Five Inventory (John & Srivastava, 1999) and the Eysenck Personality Questionnaire—Revised (Eysenck & Eysenck, 1991). Simms and Clark reported that significant amounts of item savings were achieved (36% to 37%), with a corresponding time savings of 58.0% to 60.2%, compared with the paper-and-pencil version. However, the corresponding loss of information was fairly high in three of four analyses they reported, as 6.6% to 24.6% of variance in the criterion measures was not accounted for by the adaptive administration compared with the paper-and-pencil administration, whereas in one analysis, the CA version accounted for 10.3% more of the variance in criterion measures.

CA Testing With the Minnesota Multiphasic Personality Inventory—2 (MMPI–2)

Several researchers have suggested that the IRT approach is inappropriate for use with the traditional scales and subscales of the MMPI–2 (Carter & Wilkinson, 1984; Panter, Swygert, & Dahlstrom, 1997).[1] Waller (1999) indicated that IRT could be applied to factor scales of the MMPI–2; however, these factor scales are not routinely scored, nor has research been conducted demonstrating their validity. An alternative method to an IRT approach to CA testing with personality measures such as the MMPI–2—the *countdown method*—was described by Butcher, Keller, and Bacon (1985) as a variant of the variable termination criterion approach to adaptive testing described by Weiss (1985).

The countdown method, as proposed by Butcher and colleagues (1985), classifies individuals into one of two groups (elevated or not elevated) on the basis of whether they exceed or do not exceed a cutoff criterion on a given scale. The cutoff criterion is typically the raw score that corresponds to a *clinical elevation* on a given scale. For example, if an MMPI–2 scale contains 20 items, all of which are keyed in the "true direction," and the scale requires that 10 items be endorsed in the keyed direction (i.e., *true*) to reach a clinically elevated score, then if 11 items are answered in the nonkeyed direction (i.e., *false*), the threshold score for clinical elevation (i.e., a raw score of 10 items endorsed *true*) is impossible to reach; therefore, no remaining items need to be administered for that particular scale because a clinically significant elevation has been ruled out. Two forms of applying the countdown method, the classification method and the full scores on elevated scales (FSES) method, have been suggested.

The first of these two countdown method approaches, the *classification* method, terminates scale administration once elevation is either ruled in or ruled out (i.e., the cutoff score is either reached or impossible to reach). This approach generates only an indication of whether the test taker produced an elevated score on a scale. Using the previous example of a 20-item scale that requires 10 items to be endorsed in the true direction to reach elevation, once either 10 items are endorsed in the true direction or 11 items are endorsed in the false direction, scale administration is terminated.

The second form of applying the countdown method, *FSES*, was first described by Ben-Porath, Slutske, and Butcher (1989). In the FSES approach, if the elevation cutoff score is reached, all remaining items on a scale are administered so that a full score, indicating the actual elevation, is generated for that scale. Using the previous example of a 20-item scale that requires 10 items to be endorsed in the true direction to reach elevation, if 11 items are endorsed in the false direction, then scale administration is terminated; however, if 10 items are endorsed true, then all remaining items on that scale will be administered.

A number of researchers have explored the countdown method with the MMPI–2. The first involved a real-data simulation using two personnel and two clinical samples (Ben-Porath et al., 1989). The two personnel samples included 470 participants applying for positions as airline pilots, and the two clinical samples included 232 psychiatric inpatients and 566 chemical dependency patients. Ben-Porath and colleagues (1989) explored the amount of item savings that could be achieved through the classification and FSES methods. Furthermore, several different item administration orderings were explored. These item orderings included (a) administering the least to most (L–M) frequently endorsed (by the MMPI–2 normative sample) items and (b) the most to least (M–L) frequently endorsed items with *K* scale items inserted after every 5th item, and only the first 150 items administered L–M or M–L, with the remaining items administered in booklet order. Results indicated that the L–M approach produced the most item savings

[1] It is important to note that some investigators reserve the term *adaptive testing* for procedures based on IRT. However, we adopt a definition that is less restrictive, one that would include other non-IRT methods.

compared with the other item strategies, with a range of 19.8% to 31.3% of items saved in the various samples. Furthermore, as expected, the clinical samples required more items to be administered than did the personnel samples, as the personnel samples were expected to have considerably fewer scale elevations compared with the clinical samples.

Roper, Ben-Porath, and Butcher (1991) conducted the first study of the countdown method in which the MMPI–2 items were administered following the adaptive algorithm. These authors explored the comparability of adaptive (using the L–M item ordering) and conventional testing with the MMPI–2 in a sample of male and female college students (n = 62 and 147, respectively). In this study, two validity scales (L and F), the 10 clinical scales, and the 15 content scales were administered adaptively (all of the K items were administered within the first 150 items), and test–retest correlations were calculated between the conventional and adaptive modalities. A comparison indicated strong similarity between adaptively and conventionally administered profiles. For the clinical scales, test–retest correlations were comparable to the normative sample test–retest correlations; however, for the content scales, adaptive–conventional test–retest correlations were higher than they were for the normative sample. Item savings ranged from 26.7% to 27.1% for the FSES procedure in this study.

A subsequent study by Roper, Ben-Porath, and Butcher (1995) was the first to explore the comparative validity of CA and computerized conventional (CC) testing, again with male and female college students (n = 237 and 334, respectively). Administration modalities included a booklet test–retest, a booklet and CA test–retest, and a CC–CA test–retest. Roper and colleagues found that correlations with criterion measures, including the BDI (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the State-Trait Personality Inventory (STPI; Spielberger, 1979), the Anger Expression Scale (Spielberger, 1986), and the Symptom Checklist—90—Revised (Derogatis, 1983), did not differ significantly across the three administration conditions (i.e., booklet, CC, or CA). Item savings ranged from 20% to 23% for the FSES procedure, with a corresponding time savings of 30% to 31%.

Handel, Ben-Porath, and Watt (1999) expanded on the results of Roper et al. (1995) by exploring the validity of the CA version of the MMPI–2 (MMPI–2–CA) in a clinical setting. Specifically, Handel and colleagues used a sample of 77 male Veterans Affairs inpatients and outpatients assessed at intake to a substance abuse treatment program, comparing their results on the MMPI–2–CA to the CC administration. Extratest measures included the BDI (Beck et al., 1961), the fifth edition of the Addiction Severity Index (McClellan et al., 1992), the Structured Clinical Interview for *DSM–III–R* Self-Report Personality Questionnaire (2nd ed.; Spitzer, Williams, Gibbon, & First, 1990), the NEO-PI–R (Costa & McCrae, 1992), and the Aggression Questionnaire (Buss & Perry, 1992). Handel and colleagues reported that test–retest correlations between the validity, clinical, and content scales revealed only one statistically significant difference between the CA and CC administrations, in which Clinical Scale 5 was significantly lower in the CA condition. With regard to extratest measures, Handel et al. reported that Clinical Scales 1 and 3 had somewhat attenuated correlations with extratest criteria (in this case, single items from the Addiction Severity Index) in the CA condition; however, no other differences existed between the CA and CC conditions, providing further evidence of the comparability of these two administration modalities. Finally, Handel et al. found that the CA administration resulted in a 31.4% item savings compared with the CC method.

Overall, the findings of previous research examining the CA administration of the MMPI–2 using the countdown method have indicated that substantial item and time savings can be achieved by CA administration. Furthermore, and of most noted importance, little or no loss of information in terms of test validity has been found in college and clinical samples. Since these initial studies, a new and expanded research version of the MMPI–2, the MMPI–2–CA, has been developed. Unlike the previous version, which was MS-DOS based and adaptively administered only select scales (i.e., select validity scales, and the 10 clinical and 15 content scales), the current version is based on a Windows graphical user interface and can adaptively administer all MMPI–2 scales that are part of the standard scoring materials for the test. These include the validity scales (VRIN, TRIN, F, F_B, F(p), L, K, and S), the 10 basic clinical scales, the 15 content scales, 15 supplementary scales (including the substance abuse scales), the Psychopathology–5 (PSY–5), and the restructured clinical (RC) scales. In addition, the new version allows the test administrator to select a priori which scales to administer and what cutoffs to use in the adaptive algorithm.

The current investigation was designed to expand on the results of previous studies by exploring the utility and validity of the newly developed research version of the MMPI–2–CA. Scores generated by MMPI–2–CA administration of all the standard scales of the instrument using both FSES and classification approaches were compared with those generated by the CC version of the MMPI–2. The amount of time and item savings that can be garnered by the MMPI–2–CA were explored as well. Finally, and of most noted importance, we explored CA versus CC validity using a variety of extratest measures selected to assess the constructs measured by the various MMPI–2 clinical, RC, content, and supplementary scales.

Method

Participants

A total of 517 participants were recruited for the current study, which is a part of a larger, ongoing data collection project.[2] Potential participants were undergraduate men (n = 214) and women (n = 303) who took part in exchange for credit in their undergraduate introductory psychology course. Their ages ranged from 18 to 53 years (M = 19.39, SD = 3.07) and years of education ranged from 12 to 16 (M = 12.68, SD = .88), with 1st- and 2nd-year students making up the majority (84.6%). They were primarily Caucasian (87%, n = 450), 7.4% were African American (n = 38), and 5.6% either had a different ethnicity or did not report their ethnicity (n = 29).

Potential participants were removed from the study if they produced an invalid MMPI–2 in either the CC or the CA condition

[2] Initially, 563 participants took part in this portion of data collection; however, 46 did not have complete data from a second testing session (conducted exactly 1 week after the first session) and, therefore, were not included in the current study. Reasons for the missing second session data included session cancellation due to unforeseen circumstances (e.g., university closure), data corruption or loss, or the participant failing to return to the second session.

or in both conditions. Invalid MMPI–2 was defined as having either a Cannot Say raw score (CNS) $\geq$ 30 or a True Response Inconsistency (TRIN), Lie (L), or Defensiveness (*K*) T score $>$ 80, or an *F*(p) T score $>$ 100. In addition, a variation of the Variable Response Inconsistency (VRIN) scale was developed specifically for the CA module (VRIN–CA); VRIN–CA T scores $>$ 80 were also used to identify invalid profiles. A total of 84 participants (16.2%) produced an invalid profile at Time 1, Time 2, or both testing sessions. No significant differences were found between valid and invalid groups in terms of age, $t(515) = -.046$, $p \leq .964$, or education, $t(440) = -.288$, $p \leq .774$. However, there was a significant difference between groups in terms of gender and ethnicity, with men producing more invalid profiles than women, $\chi^2(1, N = 517) = 4.992$, $p \leq .025$, and African Americans and those with unreported or other ethnicities producing more invalid profiles than Caucasians, $\chi^2(2, N = 517) = 31.626$, $p \leq .001$. The remaining sample included a total of 170 men and 263 women, with an overall mean age of 19.39 years (*SD* = 3.09), and mean education of 12.86 years (*SD* = .88). The final sample consisted predominantly of Caucasians (90.3%, *n* = 391), with African Americans (4.6%, *n* = 20) and those reporting other ethnicities (5.1%, *n* = 22) making up a smaller overall percentage.

Instruments

The MMPI–2 is a 567-item, true or false, self-report inventory that assesses an individual's characteristics across a number of domains (e.g., personality, psychopathology, social, and behavioral). In addition, the MMPI–2 has a number of scales designed to detect potential invalid styles of responding (e.g., under-, over-, or inconsistent responding). One-week test–retest reliability estimates ranged from .70 to .93 for men (*n* = 82) and from .54 to .92 for women (*n* = 111) for the clinical scales, from .77 to .91 for men and from .78 to .91 for women for the content scales (Butcher et al., 2001), and from .62 to .88 for men and women combined (*n* = 193) for the RC scales (Tellegen et al., 2003).

The MMPI–2–CA consists of 557 items, removing 10 items that are not scored on any of the existing standard MMPI–2 scales. Test-taker item responses can be entered using either a mouse or a keyboard. The software allows the test administrator to select from several administration options. These include a full conventional administration, an FSES adaptive administration of all scales, a classification adaptive administration of all scales, or an administrator's selection or creation of a set of scales (i.e., "modules") to administer either adaptively or conventionally to the test taker. The test administrator can also select the cutoffs for determining scale elevation. In the current investigation, we used the default cutoffs (described next), and we examined only the conventional, FSES, and classification administrations.

The program administers items from each of the selected scales and associated items from least to most (L–M) frequently endorsed (by the MMPI–2 normative sample) in the keyed direction, as this has been previously demonstrated to be the most effective approach to adaptive testing with the MMPI–2 (Ben-Porath et al., 1989). Items are administered sequentially from each of the selected scales. For example, if all 10 clinical scales are administered, an item is administered from Scale 1, followed by an item from Scale 2, and so forth. To establish the adaptive procedure's termination rules, a T-score cutoff of 65 was used for Validity Scale *K* and for all clinical, content, RC, and supplementary scales, A T-score cutoff of 70 was used for Validity Scales VRIN–CA, TRIN, and L, and a T-score cutoff of 80 was used for *F*(p).

Fourteen criterion measures, selected to reflect the constructs and content of several MMPI–2 scales, were included in the current study. See Table 1 for information regarding these criterion measures.

Table 1
Brief Descriptions and Internal Consistencies (Combined Gender) for Criterion Measures or Scales in the Current Study

Criterion measure	Author	*n* (items)	Symptom assessed	α[a]
Screener for Somatoform Disorders (SSD)	Janca et al., 1995	12	Physical complaints	.83
Beck Depression Inventory (BDI)	Beck et al., 1961	21	Depressive symptoms and attitudes	.89
Machiavellianism—IV (MACH-IV)	Christie & Geis, 1970	20	Negative beliefs about others (Negativism)	.65
Family Functioning Scale (FFS)	Tavitan, Lubiner, Green, Grebstein, & Velicer, 1987	40	Positive family functioning	.89
Drug Abuse Screening Test (DAST)	Skinner, 1982	20	Drug use/abuse	.86
Michigan Alcohol Screening Test (MAST)	Selzer, 1971	24	Alcohol use/abuse	.72
State Trait Personality Inventory (STPI)	Spielberger, 1979	36	Trait Anger	.86
			Trait Anxiety	.88
Obsessive Compulsive Scale (OCS)	Gibb, Bailey, Best, & Lambirth, 1983	20	Obsessiveness	.70
Magical Ideation Scale (MIS)	Eckblad & Chapman, 1983	30	Magical thinking	.81
Perceptual Aberration Scale (PAS)	Chapman, Chapman, & Raulin, 1978	35	Perceptual abnormalities	.88
Barratt Impulsivity Scale—10 (BIS-10)	Barratt, 1985	34	Motor impulsivity	.69
			General impulsivity	.80
Internal State Scale (ISS)	Bauer et al., 1991	17	Hypomanic activation	.73
			Depressive symptomatology	.74
Fears Questionnaire (FQ)	Marks & Mathews, 1979	15	Social phobia	.68
Behavioral Inhibition/Activation System (BIS/BAS)	Carver & White, 1994	20	Funseeking	.71
			Inhibition	.74

[a] The *n*s for internal consistencies range from 422 to 515 for the combined genders.

Procedure

All participants were randomly assigned to complete either a CC–CC (i.e., they completed the computerized conventional version of the test twice) or CC–CA (computerized conventional once, CA another time) administration of the MMPI–2 exactly 1 week apart. Administrations of the CC and CA were counterbalanced. Participants in the CC–CA condition were randomly assigned to complete either a FSES or classification administration of the MMPI–2. Furthermore, all participants completed one of two sets of criterion measures during each session. The measures in each criterion set were counterbalanced, as was the administration order of the criterion sets so that there was no association between MMPI–2 administration condition and criterion set. All participants completed all criterion measures by the end of the second testing session. Criterion measures were considered invalid if 10% or more of the items were not answered. All participants received credit in their introductory psychology course in exchange for participation.

Results

The first question examined involves the amount of item and time savings that can be accomplished by using the FSES and classification versions of the MMPI–2–CA compared with the CC version. Table 2 provides a breakdown of each by administration modality and administration time (i.e., Time 1 or Time 2). As indicated, the conventional modality administers 557 items (as 10 items are not scored on any of the standard MMPI–2 scales). For the classification administration, at Time 1, a mean of 442.17 items was administered (20.6% of items not administered); at Time 2, a mean of 436.91 items was administered (21.6% of items not administered). For the FSES administration, at Time 1, a mean of 458.18 items was administered (17.7% of items not administered); at Time 2, a mean of 459.55 items was administered (17.5% of items not administered). An effect size calculation for mean differences between the mean number of items administered for the classification ($n = 154$, $M = 439.54$, $SD = 23.83$) and FSES ($n = 145$, $M = 458.81$, $SD = 36.73$) modalities reveals a medium effect size: Cohen's $d = .63$, $t(297) = 12.84$, $p \leq .001$. A comparison of the mean number of items administered in the combined conventional administration ($n = 567$, $M = 557$, $SD = 0.00$) with the combined classification and FSES administrations yielded very large effect sizes, Cohen's $d = 10.69$, $t(719) = 61.17$, $p \leq .001$, and 5.94, $t(710) = 23.34$, $p \leq .001$, respectively, indicating that these two adaptive modalities result in substantial item savings compared with a conventional administration.

The administration time data indicate that, in all three administration modalities (i.e., conventional, classification, and FSES), the second administration of the test was considerably briefer than the initial administration, although the number of items administered did not vary much. For each modality, the mean administration time for the second administration was roughly 7 min shorter than for the first, suggesting that as test takers become more familiar with the administration format and interface (i.e., the computer), their time to validly complete the test shortens considerably. In terms of cross-modality comparisons, as expected, the classification procedure resulted in the lowest overall mean administration time (28.85 min), the FSES resulted in the next lower overall mean administration time (31.17 min), and the classification administration time was the longest (35.96 min). Effect size calculations for mean differences indicate a large effect for the classification administration, Cohen's $d = .82$, $t(719) = 9.31$, $p \leq .001$, and a somewhat smaller but still substantial difference for the FSES administration, Cohen's $d = .56$, $t(710) = 6.44$, $p \leq .001$, compared with the conventional administration. A comparison of the differences between the classification and FSES modalities indicated a relatively small effect size, Cohen's $d = .29$, $t(297) = 2.50$, $p \leq .05$.

In Table 3 test–retest zero-order correlations are reported within modalities and condition (CC–CC, CC–CA classification, and CC–CA FSES) by gender. For this analysis, and all subsequent correlational analyses, findings are reported by gender because previous research has indicated some gender differences in correlation patterns for individual scales (Butcher et al., 2001). Only the

Table 2
Number of Items Administered and Administration Time per Administration Modality by Administration Order

		n (items)					Time (min)				
MMPI-2 version	*n*	Minimum	Maximum	*M*	*SD*	Average not administered (%)	Minimum	Maximum	*M*	*SD*	Average savings (%)
Conventional											
Time 1	278	557	557	557	0.00	0.0	19	78	39.61	8.45	
Time 2	289	557	557	557	0.00	0.0	12	64	32.44	7.64	
Combined	567	557	557	557	0.00	0.0	12	78	35.96	8.80	
Classification											
Time 1	77	394	479	442.17	20.37	20.6	18	59	32.83	8.70	17.1
Time 2	77	358	492	436.91	26.72	21.6	12	40	24.87	5.53	23.4
Combined	154	358	492	439.54	23.83	21.1	12	59	28.85	8.29	19.8
FSES											
Time 1	78	393	527	458.18	34.54	17.7	20	61	34.15	7.40	13.8
Time 2	67	397	536	459.55	39.37	17.5	14	48	27.70	6.73	14.6
Combined	145	393	536	458.81	36.73	17.6	14	61	31.17	7.78	13.3

Note. All conventional administrations used, including those in the computerized conventional-computerized adaptive conditions, in reporting item and time savings on the basis of administering all 557 possible items of the Minnesota Multiphasic Personality Inventory—2 (MMPI-2).

Table 3
Test-Retest Correlations for CC-CC, CC-CA Classification, and CC-CA FSES for Men and Women for Scales of the Minnesota Multiphasic Personality Inventory-2

	CC-CC		CC-CA classification		CC-CA FSES			CC-CC		CC-CA classification		CC-CA FSES			CC-CC		CC-CA classification		CC-CA FSES	
Scale	Men	Women	Men	Women	Men	Women	Scale	Men	Women	Men	Women	Men	Women	Scale	Men	Women	Men	Women	Men	Women
Validity							Content							PSY-5						
VRIN-CA	.14	.32	.34	.45	.26	.39	ANX	.82	.89	.80	.82	.83	.88	AGGR	.78	.78	.59	.74	.78[f]	.78
TRIN	.13	.17	.07	.22	.46[f]	.39	FRS	.57	.86	.83[b,e]	.87	.61	.86	PSYC	.70	.77	.77	.77	.84	.84
F	.74[a]	.76	.49	.82	.79[f]	.79	OBS	.78	.85[a]	.78	.74	.83	.83	DISC	.85	.88[a]	.72	.77	.86[f]	.81
F_B	.50	.87	.78[b]	.81	.74[d]	.85	DEP	.88	.88	.79	.91	.83	.88	NEGE	.84	.86	.81	.86	.86	.85
F(p)	.46	.61	.60	.67	.74[d]	.70	HEA	.81	.94[a,c]	.67	.81	.83[f]	.79	INTR	.83	.87	.78	.80	.78	.86
L	.77[a]	.78	.54	.73	.58	.67	BIZ	.78	.76[a]	.74	.53	.77	.72[f]	RC						
K	.72	.86[a,c]	.60	.67	.77	.71	ANG	.82	.85	.78	.77	.82	.80	RCd	.84	.89	.77	.88	.86	.89
S	.83	.91[a,c]	.74	.77	.85	.84	CYN	.83	.85	.82	.78	.76	.76	RC1	.80	.92[a,c]	.68	.77	.77	.76
Clinical							ASP	.85	.87[c]	.80	.80	.77	.76	RC2	.83	.85[a]	.75	.74	.73	.82
1	.80	.95[a,c]	.75	.79	.77	.80	TPA	.64	.80[a]	.60	.65	.78	.77	RC3	.80	.86[a,c]	.71	.75	.72	.73
2	.70	.88[a]	.73	.79	.79	.86	LSE	.75	.82	.70	.84	.75	.87	RC4	.90	.91[a]	.82	.82	.88	.87
3	.67	.83[a,c]	.66	.70	.67	.69	SOD	.92[a]	.93[a]	.80	.87	.89	.91	RC6	.66	.67	.66	.71	.85[d,f]	.71
4	.77	.79	.63	.82	.80	.82	FAM	.78	.86	.69	.79	.82	.88[f]	RC7	.82	.84	.82	.82	.79	.85
5	.88	.86[a,c]	.76	.44	.76	.68[f]	WRK	.83	.87	.77	.84	.84	.90	RC8	.81	.83[a,c]	.72	.55	.72	.70
6	.76	.72	.55	.77	.72	.73	TRT	.69	.86	.74	.81	.78	.85	RC9	.81	.85[a]	.71	.74	.85[f]	.81
7	.87	.90	.79	.86	.90[f]	.87	Substance Abuse													
8	.87[a]	.90	.72	.84	.86[f]	.87	AAS	.79	.90[a]	.76	.81	.89[f]	.83							
9	.83	.79[a]	.77	.64	.86	.83[f]	APS	.76	.74	.62	.70	.80[f]	.71							
0	.92[a]	.92	.82	.86	.87	.88	MAC-R	.82	.76	.77	.74	.82	.73							

Note. CC-CC: men (n = 49), women (n = 85); CC-CA class: men (n = 64), women (n = 90); CC-CA FSES: men (n = 57), women (n = 88). All correlations significantly different at $p \leq .05$. CC = computerized conventional administration; CA = computerized adaptive administration; FSES = full scores on elevated scales method; Validity: VRIN = Variable Response Inconsistency, TRIN = True Response Inconsistency, *F* = Infrequency, L = Lie, *K* = Defensiveness; Clinical: 1 = Hypochondriasis, 2 = Depression, 3 = Hysteria, 4 = Psychopathic Deviate, 5 = Masculinity-Femininity, 6 = Paranoia, 7 = Psychasthenia, 8 = Schizophrenia, 9 = Hypomania, 0 = Social Introversion; Content: ANX = Anxiety, FRS = Fears, OBS = Obsessiveness, DEP = Depression, HEA = Health Concerns, BIZ = Bizarre Mentation, ANG = Anger, CYN = Cynicism, ASP = Antisocial Practices, TPA = Type A Behavior, LSE = Low Self-Esteem, SOD = Social Discomfort, FAM = Family Problems, WRK = Work Interference, TRT = Negative Treatment Indicators; Substance Abuse: AAS = Addiction Acknowledgement Scale, APS = Addiction Potential Scale, MAC-R = MacAndrew Alcoholism Scale-Revised; PSY-5 = Psychopathology-5; AGGR = Aggressiveness, PSYC = Psychoticism, DISC = Disconstraint, NEGE = Negative Emotionality/Neuroticism, INTR = Introversion/Low Positive Emotionality; RC = Restructured Clinical: RCd = Demoralization, RC1 = Somatic Complaints, RC2 = Low Positive Emotions, RC3 = Cynicism, RC4 = Antisocial Behavior, RC6 = Ideas of Persecution, RC7 = Dysfunctional Negative Emotions, RC8 = Aberrant Experiences, RC9 = Hypomanic Activation.
[a] CC-CC significantly higher than CC-CA classification. [b] CC-CA classification significantly higher than CC-CC. [c] CC-CC significantly higher than CC-CA FSES. [d] CC-CA FSES significantly higher than CC-CC. [e] CC-CA classification significantly higher than CC-CA FSES. [f] CC-CA FSES significantly higher than CC-CA classification.

validity, clinical, content, PSY–5, RC, and substance abuse scales are reported, as these were the ones included in subsequent correlational analyses with criterion measures. Fisher's *R* to *Z* transformations were conducted to allow for tests of whether there were significant differences between the test–retest correlations of the CC–CC versus CC–CA classification, CC–CC versus CC–CA FSES, and CC–CA classification versus CC–CA FSES administrations within gender. To reduce the likelihood of concluding that the modalities yield similar results when they do not, we did not apply a Bonferroni correction, and we set alpha at .05 for all comparisons. An effect size statistic *Q* was calculated to reflect the magnitude of differences between the transformed correlation values, with .10, .30, and .50 reflecting small, medium, and large effect sizes, respectively.

For men, a comparison of CC–CC versus CC–CA classification administrations indicates that five CC–CC T–*R* correlations were significantly higher than in the CC–CA classification condition (*F*, L, 8, 0, and SOD), with all effect sizes falling in the medium range. Two CC–CA classification T–*R* correlations were significantly higher than in the CC–CC condition (F_B and FRS), with a large effect size for FRS and medium effect size for F_B. Three CC–CA FSES T–*R* correlations were significantly higher than in the CC–CC condition (F_B, *F*(p), and RC6), all in the medium effect size range. Eleven CC–CA FSES T–*R* correlations were significantly higher than in the CC–CA classification condition (TRIN, *F*, 7, 8, HEA, AAS, APS, AGGR, DISC, RC6, and RC9), with all effect sizes falling in the medium range, with the exception of *F*. One CC–CA classification correlation (FRS) was significantly higher than in the CC–CA FSES condition, reaching a large effect size. Overall, for men, RC6 was higher in the CC–CA FSES administration than CC–CC or CC–CA classification, and FRS was higher in the CC–CA classification administration than in the CC–CC or CC–CA FSES administrations; otherwise, no other consistent differences were found across administration modalities.

For women, 20 CC–CC T–*R* correlations were significantly higher than in the CC–CA classification condition (*K*, *S*, 1, 2, 3, 5, 9, OBS, HEA, BIZ, TPA, SOD, AAS, DISC, RC1, RC2, RC3, RC4, RC8, and RC9), with six effect sizes falling in the large range (*S*, 1, 5, HEA, RC1, and RC8), and the remainder falling in the medium range. Ten CC–CC T–*R* correlations were significantly higher than in the CC–CA FSES condition (*K*, *S*, 1, 3, 5, HEA, ASP, RC1, RC3, and RC8), with three effect sizes falling in the

large range (1, HEA, and RC1), and the remainder falling in the medium range. Four CC–CA FSES T–*R* correlations were significantly higher than in the CC–CA classification condition (5, 9, BIZ, and FAM), all of which reached a medium effect size. For women, *K*, *S*, 1, 3, 5, HEA, RC1, RC3, and RC8 were significantly lower in both CA conditions compared with the CC condition. Overall, no MMPI–2 scales showed a consistent difference between administration modality across gender.

Table 4 reports the zero-order correlations between criterion measures and conceptually related MMPI–2 scales by administration modality and gender. Differences between correlations by administration modality and gender were tested using a *t* test comparison for dependent correlations, and significant *t* test values were converted to Cohen's *d* values to provide effect size levels (small = .20, medium = .50, and large =.80). Alpha was set to .05 for all comparisons.

For men, there was considerable similarity across modalities, with only 13 significant differences out of 120 scale-by-scale comparisons. One CC Time 1 correlation was significantly higher compared with CC Time 2 (4 with STPI Anger), with a medium effect size. Two CC Time 2 correlations were significantly higher compared with CC Time 1 (9 with ISS Activation; INTR with FQ Social Phobia), with medium and large effect sizes, respectively. Four CC correlations were significantly higher than in the CA classification modality (1, HEA and RC1 with SSD; RCd with BDI), all with medium-level effect sizes. Three CC correlations were significantly higher than in the CA FSES modality (ANG with STPI Anger; RC6 with PAS; DISC with BIS/BAS Funseeking), all with medium-level effect sizes. Three CA FSES correlations were significantly higher than in the CC condition (MAC-R with DAST; RCd with STPI Anxiety; PSYC with MIS), all with medium-level effect sizes. Of the scales that had significant differences in correlations with criteria between CC and CA administration modalities, three differed in the T–*R* correlations for men. These scales, HEA, DISC, and RC6, demonstrated significantly higher T–*R* correlation in the CA FSES condition than in the CA classification condition. However, each scale had a significantly lower correlation with one of the several criterion measures with which they were compared.

For women, again, the correlations were more similar than disparate across modalities (i.e., 23 differences out of 120 scale-by-scale comparisons). Two CC Time 1 correlations were significantly higher compared with CC Time 2 (OBS, RC7 with OCS), both with medium effect sizes. Six CC Time 2 correlations were significantly higher compared with CC Time 1 (DEP and RC2 with BDI; 4, RC4, AAS with MAST; RC4 with STPI Anger); effect sizes for the differences were all medium, with the exception of RC4 with STPI Anger, which was small. Four CC correlations were significantly higher than in the CA classification modality (2, DEP and RCd with BDI; RCd with STPI Anxiety), and all had small effect sizes, with the exception of DEP and BDI. Eleven CA FSES correlations were significantly higher than in the corresponding CC modality (1 with SSD; DEP, RC2 with BDI; 2, DEP with ISS Depression; RC4 with DAST; RC7 with STPI Anxiety; 6, BIZ, RC6 with MIS; and 9 with ISS Activation), with two having small effect sizes (1 and SSD; 6 and MIS), one having a large effect size (RC2 and BDI), and the remainder having medium effect sizes. Of the scales that demonstrated significant differences in correlations with criteria between CC and CA administration modalities, six differed in the T–*R* correlations. These scales, 1, 2, 9, BIZ, RC2, and RC4, demonstrated a higher T–*R* correlation for CC but had a higher correlation with one of the several criterion measures with which they were compared in the CA FSES condition, and one (2) had a lower correlation in the CA classification condition. Overall, no consistent differences between the conventional and adaptive modalities of the MMPI–2 were found in the correlation with criterion measures across gender. Differences in T–*R* correlations appeared to have minimal impact on correlations with criterion measures across CA modalities.

Discussion

Our analyses had two primary goals: to examine the amount of item and corresponding time savings that can be achieved through classification and FSES MMPI–2–CA administration, and the cost, in terms of predictive validity, associated with these savings. Overall, our findings indicate that adaptive administration of all MMPI–2 scales reduces the amount of time and number of items needed to obtain assessment information without a significant attenuation of test score validity in this sample.

Some interesting findings emerged in examining item and time savings. As expected, the MMPI–2–CA classification modality resulted in the most savings in number of items administered. Given that the MMPI–2 classification procedure simply rules in or rules out elevation, it should be expected that this administration modality would result in the most savings across the board. For the FSES adaptive administration, we found a large effect size for item savings compared with the CC administration. However, interestingly, our analyses indicated only a medium effect for the difference in the number of items administered in the FSES CA administration compared with the classification CA administration. Thus, only a modest amount of items savings is gained in choosing the classification procedure over the FSES approach to MMPI–2–CA administration. In terms of time savings, a somewhat larger effect was found for the classification CA administration (.82) compared with the FSES CA approach (.56); however, in absolute numbers, the difference between these two modalities was relatively small. On average, the classification administration saved only 2.32 min compared with FSES administration in this nonclinical sample. It is possible that in clinical settings, where more elevation is to be expected, the classification procedure would offer more of a time-savings advantage compared with the FSES approach. Nevertheless, our results indicate that in nonclinical settings, the FSES approach is likely to provide some additional data (regarding elevation) at minimal additional cost (in terms of administration time).

Another interesting finding related to time savings was that the second administration, regardless of modality, always resulted in a substantial reduction in administration time. Roughly 7 min were saved during the second administration of any version of the MMPI–2 (i.e., CA or CC), although a comparable number of items was administered both times and the resulting scores were comparably valid. This indicates that as individuals became familiar with computerized administration of the MMPI–2, their response times to items were reduced, but the validity of their responses was not.

In terms of *T–R* correlations, differences between the CC, CA classification, and CA FSES administrations were found in 56 of

Table 4
Correlations Between Criterion Measures and Conceptually Related Minnesota Multiphasic Personality Inventory—2 (MMPI-2) Scales by Administration Modality and Gender

	Men						Women					
	Classification				FSES		Classification				FSES	
Criterion measure/MMPI-2 scale	CC	CC	CC	CA	CC	CA	CC	CC	CC	CA	CC	CA
Screener for Somatoform Disorders (SSD)												
1	.56	.63	.71[c]	.57	.50	.49	.72	.75	.74	.79	.73	.82[f]
HEA	.55	.65	.60[c]	.35	.47	.52	.72	.74	.76	.80	.71	.77
RC1	.58	.68	.62[c]	.40	.35	.41	.76	.77	.71	.77	.67	.75
Beck Depression Inventory (BDI)												
2	.54	.48	.68	.56	.53	.42	.71	.69	.72[c]	.62	.63	.70
DEP	.79	.78	.80	.76	.67	.69	.77	.84[b]	.86[c]	.74	.69	.79[f]
NEGE	.69	.75	.64	.55	.58	.54	.55	.61	.66	.59	.65	.72
RCd	.79	.78	.80[c]	.69	.72	.65	.73	.83[b]	.81[c]	.74	.73	.77
RC2	.46	.51	.63	.50	.46	.42	.70	.73	.64	.56	.51	.69[f]
Internal State Scale (ISS) Depression												
2	.37	.17	.53	.45	.17	.10	.62	.61	.50	.47	.48	.60[f]
DEP	.63	.57	.57	.58	.28	.25	.67	.63	.67	.66	.52	.62[f]
NEGE	.42	.45	.37	.41	.16	.03	.47	.46	.40	.38	.41	.44
RCd	.56	.47	.52	.48	.25	.15	.73	.69	.62	.64	.54	.60
RC2	.11	.15	.47	.39	.14	.20	.67	.65	.45	.44	.51	.56
Machiavellianism—IV (MACH-IV) Negativism												
CYN	.54	.49	.47	.52	.50	.59	.58	.62	.39	.48	.62	.52
RC3	.57	.48	.49	.55	.58	.61	.58	.62	.32	.43	.58	.54
Family Functioning Scale (FFS)												
4	−.52	−.50	−.47	−.41	−.43	−.45	−.50	−.54	−.46	−.39	−.56	−.64
FAM	−.56	−.64	−.58	−.56	−.67	−.57	−.63	−.64	−.59	−.62	−.71	−.71
RC4	−.43	−.42	−.26	−.12	−.39	−.31	−.36	−.40	−.28	−.18	−.52	−.44
Drug Abuse Screening Test (DAST)												
4	.41	.25	.41	.31	.63	.61	.54	.54	.41	.38	.21	.28
RC4	.71	.69	.60	.53	.64	.62	.65	.65	.52	.52	.54	.66[f]
AAS	.69	.58	.65	.59	.71	.64	.66	.71	.50	.56	.64	.69
APS	.37	.39	.19	.19	.31	.43	.27	.28	.24	.25	.28	.24
MAC-R	.43	.37	.37	.41	.28	.51[f]	.48	.36	.48	.46	.39	.47
Michigan Alcohol Screening Test (MAST)												
4	.14	.17	.15	.25	.53	.49	.24	.41[b]	.27	.21	.15	.14
RC4	.24	.14	.25	.27	.59	.56	.44	.56[b]	.17	.16	.21	.29
AAS	.25	.19	.41	.28	.52	.52	.40	.53[b]	.24	.23	.36	.38
APS	.16	.09	.01	.07	.41	.32	.34	.27	.27	.25	.19	.17
MAC-R	.21	.26	.28	.20	.29	.40	.35	.28	.33	.33	.17	.13
State Trait Personality Inventory (STPI) Anger												
4	.31[a]	.11	.30	.30	.53	.54	.47	.36	.37	.43	.46	.51
ANG	.50	.55	.65	.60	.85[c]	.76	.74	.72	.69	.59	.63	.65
RC4	.23	.22	.12	.20	.57	.52	.36	.45[b]	.34	.38	.45	.46
State Trait Personality Inventory (STPI) Anxiety												
7	.72	.75	.75	.64	.77	.81	.80	.80	.79	.76	.76	.82
ANX	.66	.69	.71	.69	.76	.77	.76	.78	.75	.69	.79	.82
NEGE	.62	.65	.56	.59	.75	.71	.68	.71	.71	.72	.75	.79
RCd	.67	.69	.80	.74	.80	.88[f]	.81	.81	.84[c]	.78	.81	.85
RC7	.58	.55	.65	.57	.68	.72	.67	.65	.64	.64	.70	.80[f]
Obsessive Compulsive Scale (OCS)												
7	.34	.33	.26	.33	.04	.08	.56	.48	.40	.38	.36	.34
OBS	.35	.44	.38	.43	.07	.18	.53[a]	.40	.26	.14	.39	.43
RC7	.38	.41	.26	.32	.12	.12	.59[a]	.45	.40	.41	.39	.36
Magical Ideation Scale (MIS)												
6	.47	.47	.31	.40	.47	.58	.15	.14	.16	.10	.33	.48[f]
8	.46	.42	.53	.48	.52	.62	.41	.42	.59	.51	.54	.60
BIZ	.62	.61	.63	.53	.72	.79	.64	.67	.48	.45	.65	.79[f]
PSYC	.53	.51	.55	.46	.68	.79[f]	.57	.56	.54	.54	.63	.69
RC6	.51	.48	.47	.37	.64	.67	.30	.39	.35	.43	.46	.64[f]
RC8	.61	.59	.60	.53	.70	.72	.58	.65	.52	.43	.64	.69
Perceptual Aberration Scale (PAS)												
6	.47	.53	.39	.42	.61	.60	.39	.42	.14	.16	.28	.39
8	.49	.52	.53	.43	.66	.65	.59	.63	.39	.42	.45	.47

(*table continues*)

Table 4 (*continued*)

Criterion measure/MMPI-2 scale	Men						Women					
	Classification				FSES		Classification				FSES	
	CC	CC	CC	CA	CC	CA	CC	CC	CC	CA	CC	CA
Perceptual Aberration Scale (PAS) (*continued*)												
BIZ	.66	.74	.50	.41	.73	.63	.59	.63	.42	.40	.52	.55
PSYC	.60	.64	.41	.33	.66	.70	.49	.49	.47	.41	.44	.39
RC6	.60	.56	.38	.29	.80[c]	.71	.39	.42	.21	.23	.32	.34
RC8	.58	.69	.54	.39	.66	.67	.59	.68	.50	.46	.56	.58
Behavioral Inhibition/Activation System (BIS/BAS) Funseeking												
9	.30	.26	.22	.36	.46	.39	.35	.30	.28	.29	.32	.25
DISC	.23	.21	.41	.25	.46[e]	.32	.33	.33	.29	.22	.40	.31
RC9	.34	.34	.35	.33	.45	.33	.27	.29	.33	.30	.39	.31
Internal State Scale (ISS) Activation												
9	.24	.42[b]	.31	.27	.38	.41	.17	.22	.25	.26	.30	.45[f]
RC9	.42	.39	.22	.30	.32	.44	.20	.22	.33	.37	.23	.26
Fears Questionnaire (FQ) Social Phobia												
0	.42	.47	.48	.49	.47	.39	.65	.65	.62	.62	.52	.50
SOD	.36	.42	.34	.36	.28	.24	.52	.56	.49	.48	.41	.39
INTR	.07	.41[b]	.29	.15	.17	.27	.44	.42	.45	.38	.40	.35
Behavioral Inhibition/Activation System (BIS/BAS) Inhibition												
DISC	−.39	−.43	−.38	−.39	−.15	−.21	−.30	−.20	−.28	−.32	−.09	−.12

Note. All correlations significantly different at $p \leq .05$. Total possible *n*s range from 49 to 64 for men and from 85 to 90 for women; however, some criterion measures were missing or invalid, thus reducing the total *n* for each gender in some cases. FSES = full scores on elevated scales; CC = computerized conventional administration; CA = computerized adaptive administration. MMPI-2 subscales: 1 = Hypochondriasis (Clinical); HEA = Health Concerns (Content); RC1 = Somatic Complaints (Restructured Clinical); 2 = Depression (Clinical); DEP = Depression (Content); NEGE = Negative Emotionality/Neuroticism (Psychopathology-5; PSY-5); RCd = Demoralization (Restructured Clinical); RC2 = Low Positive Emotions (Restructured Clinical); CYN = Cynicism (Content); RC3 = Cynicism (Restructured Clinical); 4 = Psychopathic Deviate (Clinical); FAM = Family Problems (Content); RC4 = Antisocial Behavior (Restructured Clinical); AAS = Addiction Acknowledgement Scale (Substance Abuse); APS = Addiction Potential Scale (Substance Abuse); MAC-R = MacAndrew Alcoholism Scale-Revised (Substance Abuse); ANG = Anger (Content); 7 = Psychasthenia (Clinical); ANX = Anxiety (Content); RC7 = Dysfunctional Negative Emotions (Restructured Clinical); OBS = Obsessiveness; 6 = Paranoia (Clinical); 8 = Schizophrenia (Clinical); BIZ = Bizarre Mentation (Content); PSYC = Psychoticism (PSY-5); RC6 = Ideas of Persecution (Restructured Clinical); RC8 = Aberrant Experiences (Restructured Clinical); 9 = Hypomania (Clinical); DISC = Disconstraint (PSY-5); RC9 = Hypomanic Activation (Restructured Clinical); 0 = Social Introversion (Clinical); SOD = Social Discomfort (Content); INTR = Introversion/Low Positive Emotionality (PSY-5). [a] CC Time 1 significantly higher than CC Time 2. [b] CC Time 2 significantly higher than CC Time 1. [c] CC significantly higher than CA classification. [d] CA classification significantly higher than CC. [e] CC significantly higher than CA full scores on elevated scales (FSES) method. [f] CA FSES significantly higher than CC.

300 scale-by-scale comparisons. Of note, no test–retest correlations were consistently higher in the CC modality compared with the CA modalities across genders, indicating that there is no pattern of reduced comparability of conventional and adaptive administrations. Moreover, the most informative indicator of the comparability of results across modalities involves a comparison of the validity of scale scores generated by conventional versus adaptive administration of the MMPI–2. These results, reported in Table 4, indicate that for men, there was considerable similarity in the validities of conventionally and adaptively administered MMPI–2 results. In the CC versus CA classification procedure comparisons, only four CC correlations were significantly higher with criterion measures than their CA counterparts. Interestingly, three of these four correlations were on health concerns scales (1, HEA, and RC1), which is similar to the findings of Handel et al. (1999) in a sample of male Veterans Affairs patients tested at intake to a substance abuse treatment program. On the other hand, in the CC versus CA FSES comparison, there were no significant differences in the prediction of criteria with these health-related MMPI–2 scales. Three scales for men, ANG, DISC, and RC6, revealed a significantly higher correlation for the CC administration compared with the CA FSES administration, whereas three CA FSES correlations, PSYC, MAC-R, and RCd, were significantly higher in the CA FSES administration.

For women, the correlations between MMPI–2 scales and criteria were also quite similar across administration modalities, and in 11 cases, the CA FSES administration actually produced significantly higher validity coefficients. In 6 cases, the correlations for the CC Time 2 administration were significantly higher than the CC Time 1 administration, and in 4 cases the CC correlation was significantly higher than the CA classification correlation. Overall, there were very few differences in predictive validity for men and women between CC and CA MMPI–2 administrations, indicating that the savings in number of items and amount of administration time do not come at the cost of reduced validity of the resulting scale scores.

Simms and Clark (2005) posited that the MMPI–2–CA countdown method of adaptive testing "generally results in substantial loss of information" (p. 30). Our results indicate, on the contrary, that the MMPI–2–CA countdown method-based approach, especially using the FSES strategy, yields equally valid scores, and in some cases (i.e., for women in this study), the FSES approach

actually produced significantly higher correlations with extratest criterion measures than did conventional test administration. Indeed, our findings are in line with those reported in previous MMPI–2–CA studies (e.g., Handel et al., 1999; Roper et al., 1991, 1995) in demonstrating no significant loss of information with this approach.

Several shortcomings of the current study need to be considered in identifying future directions for research in this area. First, the use of college students limits the generalizability of findings of the current investigation. Replication in clinical and other settings is needed. In addition, not all of the traditionally scored MMPI–2 scales were subjected to analyses in the current study. Follow-up research exploring the comparability of MMPI–2–CA scores with other scales is needed.

A relatively low number of men were included in each of the administration modalities, with total possible sample sizes ranging from 49 to 64 for each correlational analysis, although, with missing data, these numbers were somewhat reduced. Whereas a possible lack of power may have adversely affected analyses, an examination of the correlations with criterion measures across administration modalities reveals a remarkable similarity among correlations: In some cases, the CA and CC administrations correlations are virtually identical, in others, one administration or the other is somewhat higher (although the difference did not reach statistical significance); neither modality reveals a distinct advantage in terms of a pattern of higher correlations.

Finally, the MMPI–2–CA software makes it possible to administer subsets of MMPI–2 scales rather than the entire set of measures, as was done in the present investigation. The impact of administering only some, but not all of the scales of the instrument, in terms of both time savings and the validity of the resulting scores needs to be investigated.

In contrast to the area of ability testing, CA personality testing has yet to be applied routinely in practice, in spite of the clear advantages this technology may offer in terms of time savings. As reviewed earlier, one reason why this may be so is that several authors have questioned the appropriateness of adopting the primary adaptive testing methodology used in ability testing, IRT, for personality assessment. Moreover, until Simms and Clark (2005) published their study of adaptive administration of the SNAP, no data were available on the impact of IRT-based adaptive testing on the validity of the resulting scale scores. Absent data examining the validity of IRT-based CA personality tests, users are unlikely to embrace this technology.

In contrast, the countdown method, as applied to the MMPI–2 (in both the FSES and classification modalities), has now been the subject of a number of investigations comparing its validity with conventional administration of the instrument. These studies have generally indicated that the MMPI–2–CA yields results that are as valid as those generated by conventional test administration, but with significant time savings. Moreover, scores generated by the countdown method are actual T scores, directly interchangeable with the results of conventional test administration, and the FSES procedure yields a full T score based on the administration of all of the items on any scale that exceeds a designated cutoff. As a result, it likely requires more items than would an IRT-based administration of the same scale, a trade-off that users may find acceptable in return for obtaining more readily interpretable (based on the existing literature) results.

As users become more familiar and comfortable with CA personality assessment, a number of new or expanded applications may be examined. For example, a subset of the scales of the MMPI–2–CA (targeting treatment goals) could be administered to individuals receiving psychotherapy on a regular basis to assess treatment progress or outcome. Use of a brief version of the instrument to screen for targeted forms of psychopathology (e.g., depression, anxiety) may also be possible. Such approaches would, of course, need to be empirically validated prior to being implemented clinically.

Adaptive personality and psychopathology testing is in its early stages of development. As computer technology evolves, additional developments, such as testing with hand-held devices, and improved audiovisual interfaces will likely be explored as well. As researchers take advantage of these developments, they should keep in mind the needs and expectations of applied personality assessors.

References

Barratt, E. S. (1985). Impulsiveness subtraits: Arousal and information processing. In J. T. Spence & C. E. Izard (Eds.), *Motivation, emotion, and personality* (Vol. 5, pp. 137–146). New York: North-Holland.

Bauer, M. S., Crits-Christoph, P., Ball, W. A., Dewees, E., McAllister, T., Alahi, P., et al. (1991). Independent assessment of manic and depressive symptoms by self-rating. *Archives of General Psychiatry, 48,* 807–812.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 12,* 57–62.

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

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

Buss, A. H., & Perry, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology, 63,* 452–459.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. K. (2001). *MMPI–2 (Minnesota Multiphasic Personality Inventory—2): Manual for administration, scoring, and interpretation.* Minneapolis: University of Minnesota Press.

Butcher, J. N., Keller, L. S., & Bacon, S. F. (1985). Current developments and future directions in computerized personality assessment. *Journal of Consulting and Clinical Psychology, 53,* 803–815.

Carter, J. E., & Wilkinson, L. (1984). A latent trait analysis of the MMPI. *Multivariate Behavioral Research, 19,* 385–407.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology, 67,* 319–333.

Chapman, L. J., Chapman, J. P., & Raulin, M. L. (1978). Body-image aberration in schizophrenia. *Journal of Abnormal Psychology, 87,* 399–407.

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

Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism.* New York: Academic Press.

Clark, L. A. (1993). *Schedule for Nonadaptive and Adaptive Personality (SNAP). Manual for administration, scoring, and interpretation.* Minneapolis: University of Minnesota Press.

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

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

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI–R) and NEO Five-Factor Inventory (NEO-FFI).* Odessa, FL: Psychological Assessment Resources.

Derogatis, L. J. (1983). *SCL–90–R administration, scoring and procedures manual: II.* Towson, MD: Clinical Psychometric Research.

Eckblad, M. L., & Chapman, L. J. (1983). Magical ideation as an indicator of schizotypy. *Journal of Consulting and Clinical Psychology, 51,* 215–225.

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

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78,* 350–365.

Gibb, G., Bailey, J., Best, R., & Lambirth, T. (1983). The measurement of obsessive compulsive personality. *Educational and Psychological Measurement, 43,* 1233–1237.

Hammond, S. M. (1995). An IRT investigation of the validity of nonpatient analogue research using the Beck Depression Inventory. *European Journal of Psychological Assessment, 11,* 14–20.

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.

Janca, A., Burke, J. D., Isaac, M., Burke, K. C., Costa e Silva, J. A., Acuda, S. W., et al. (1995). The World Health Organization Somatoform Disorders Schedule: A preliminary report on design and reliability. *European Psychiatry, 10,* 373–378.

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 Press.

Lai, J., Cella, D., Chang, C., Bode, R. K., & Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT–Fatigue scale. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 12,* 485–501.

Marks, I. M., & Mathews, A. M. (1979). Brief standard self-rating for phobic patients. *Behaviour Research and Therapy, 23,* 563–569.

McClellan, A. T., Kushner, H., Metzger, D., Peters, R., Smith, I., Grissom, G., Pettinati, H., & Argeriou, M. (1992). The fifth edition of the Addiction Severity Index. *Journal of Substance Abuse Treatment, 9,* 199–213.

Panter, A. T., Swygert, K. A., & Dahlstrom, W. G. (1997). Factor analytic approaches to personality item-level data. *Journal of Personality Assessment, 68,* 561–589.

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

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81,* 93–103.

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.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analysis of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6,* 255–270.

Selzer, M. (1971). The Michigan Alcoholism Screening Test: The quest for a new diagnostic instrument. *American Journal of Psychiatry, 127,* 1653–1658.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17,* 28–43.

Skinner, H. A. (1982). The Drug Abuse Screening Test. *Addictive Behaviors, 7,* 363–371.

Spielberger, C. D. (1979). *Preliminary manual for the State Trait Personality Inventory (STPI*). Tampa: University of South Florida.

Spielberger, C. D. (1986). *The Anger Expression (EX) Scale.* Tampa: University of South Florida.

Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1990). *Structured Clinical Interview for the DSM–III–R personality disorders (SCID–II).* Washington, DC: American Psychiatric Press.

Tavitian, M. L., Lubiner, J., Green, L., Grebstein, L. C., & Velicer, W. F. (1987). Dimensions of family functioning. *Journal of Social and Behavioral Personality, 2,* 191–204.

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The MMPI–2 Restructured Clinical Scales: Development, validation, and interpretation.* Minneapolis: University of Minnesota Press.

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

Ware, J. E., Jr., Gandek, B., Sinclair, S. J., & Bjorner, J. B. (2005). Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. *Rehabilitation Psychology, 50,* 71–78.

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

Received July 20, 2005
Revision received January 20, 2006
Accepted January 25, 2006 ■