Development of a Short Form for the MMPI-2 Based on Scale Elevation Congruence

Robert E. McGrath, Ray Terranova, David L. Pogge and Celina Kravic

Assessment 2003; 10; 13
DOI: 10.1177/1073191102250333

The PDF of this article has been modified
from its original version.

The appendix material was removed.

Development of a Short Form for the MMPI-2 Based on Scale Elevation Congruence

Robert E. McGrath
Ray Terranova
Fairleigh Dickinson University

David L. Pogge
Fairleigh Dickinson University and Four Winds Hospital

Celina Kravic
Fairleigh Dickinson University

The length of the Minnesota Multiphasic Personality Inventory (MMPI) is often considered a barrier to its use, leading to the development of short forms. Two methods of abbreviating the revised MMPI have now been developed. One agrees poorly with the long form in terms of which scales are elevated. The second ensures perfect congruence in which scales are elevated but requires computer administration. This article describes the development of a short form representing a compromise approach. The short form was derived using 800 psychiatric inpatients and cross-validated with samples of 658 inpatients and 266 outpatients. It is briefer than the computerized short form but does not achieve perfect congruence with the full inventory. It is longer than earlier noncomputerized short forms but demonstrates greater scale elevation congruence with the full inventory and allows estimates of more scales. The short form offers a reasonable alternative when the full inventory is impractical.

Keywords: MMPI; short form; classification accuracy; clinical elevations

One of the more commonly noted barriers to the use of the Minnesota Multiphasic Personality Inventory (MMPI) and its recent revision, the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), is its length (e.g., Gallucci, 1986; Groth-Marnat, 1997). The cognitive and time demands of the task often make the inventory impractical for individuals suffering specifically from those types of impairments for which the MMPI is most useful.

Efforts to reduce administration time have a long history (MacDonald, 1952) and ultimately led to the development of a series of short forms for the original MMPI. Various strategies were used to select items for these shortened versions, including factor analysis, inclusion on the greatest number of scales, and simple position in the inventory (Faschingbauer, 1974; Kincannon, 1968; Overall & Gomez-Mont, 1974). Length varied as well, between 71 and 168 items.

Initial enthusiasm for the short forms waned as problems became evident. Butcher and Hostetler (1990) summarized a number of these problems. The short-form scales demonstrated lower reliability than the full-length scales. Butcher and Hostetler also questioned the validity of the short-form scales, although they noted that validity was usually evaluated through correlations with the full-length scales rather than with external criteria. One of the most serious issues was poor high-point code congruence when compared to the full inventory. Many studies found that high-point codes based on the short form agreed with

The authors gratefully acknowledge the administration of Four Winds Hospital, Katonah, New York, for its support in the collection of data for this article. The authors are solely responsible for its content. Correspondence concerning this article should be addressed to Robert E. McGrath, School of Psychology T110A, Fairleigh Dickinson University, Teaneck, NJ 07666; e-mail: mcgrath@alpha.fdu.edu.

Assessment, Volume 10, No. 1, March 2003 13-28
DOI: 10.1177/1073191102250333

the full version in fewer than 50% of cases. Given the poor classification congruence between short forms and the full inventory, Streiner and Miller (1986) concluded that researchers interested in developing shorter measures of psychopathology would be better served by developing entirely new inventories.

Even so, the very large empirical basis for the interpretation of the MMPI scales, as well as the desire for a common framework for assessing individuals, make the development of an acceptable shortened version attractive. Two strategies have been suggested so far for abbreviating the administration of the MMPI-2. Dahlstrom and Archer (2000) developed a short form using the first 180 items of the MMPI-2, which will be called the MMPI-180 here. Although their short form is easy to administer and score, congruence with the full inventory was weak even according to the statistics they provided. High point agreement was only 50%, whereas 2-point codes were congruent in only about one third of cases. Finally, as was true of short forms developed for the original version of the MMPI, derivation was limited to the traditional 13 scales. The many scales added during the development of the MMPI-2 were omitted.

The second strategy involves computerized adaptive testing, so that the number of items administered varies across individuals. Various models have been evaluated for determining how many and which items to administer, but the following strategy has become the standard (Ben-Porath, Slutske, & Butcher, 1989; Handel, Ben-Porath, & Watt, 1999; Roper, Ben-Porath, & Butcher, 1991). An initial set of 150 items is administered, including all items from the *K* scale and the least frequently endorsed items from 27 other scales. These include scales *L* and *F*, the 10 clinical scales, and the 15 content scales. The computer then determines whether it would be possible for the respondent to achieve an elevated score ($T \geq 65$) if all remaining items on the scale were endorsed in the keyed direction. If not, administration of items from that scale is terminated. If an elevated score is still possible, administration of items from that scale continues until elevation is ruled out or all items are administered. This means that all items are administered if the scale is elevated.

Adaptive testing has several advantages over the traditional short-form approach. Full administration of elevated scales means that both exact elevation and the high-point code are known. However, the savings in length are not necessarily substantial. The number of items administered has been reduced by 13% to 31% across samples (Ben-Porath et al., 1989; Handel et al., 1999; Roper et al., 1991). Even when the goal was limited to determining whether the scale was elevated, reductions varied between 21% and 31% across samples. This approach also required computerized administration, which can be impractical or impossible in certain settings or with certain patients.

The purpose of the present study was to derive a short form for the MMPI-2. The short form that is described here is unique in several ways. First, as with the adaptive strategy, the short form was intended to achieve congruence at the level of the individual scale. Achieving congruence in classification at the scale level ensures the clinical utility of the instrument. Second, unlike previous short forms, level of congruence was considered more important in the development of this instrument than brevity. Finally, other scales besides the traditional validity and clinical scales were included. This resulted in the development of a short form that is longer than earlier short forms, although the savings are greater than those associated with the adaptive testing model.

METHOD

Participants

Three samples of psychiatric patients were used in this study. The only exclusionary criterion used was omission of more than 30 items on the MMPI-2.

The derivation sample consisted of 800 inpatients from Four Winds Hospital, a private psychiatric facility in New York State, who completed the MMPI-2 upon admission between January 1993 and August 1995. On average, members of this sample were 35.5 years old ($SD = 13.5$), with 13.9 years of education ($SD = 2.9$). The patients were 44.6% male and 55.4% female. The majority was identified either as single (48.2%) or married (29.4%). Ethnicity data were not available for this sample, but the population at the facility where data were gathered is approximately 62% White, 18% Black, 14% Hispanic, and 6% Other. Four Winds Hospital serves a mix of urban, suburban, and rural populations.

The first cross-validation sample consisted of 658 subsequent admissions to the same facility who completed the MMPI-2 between August 1995 and February 1999. These patients were on average 34.0 years old ($SD = 13.0$), with 14.1 years of education ($SD = 3.0$). The sample was 48.5% male and 51.5% female, and most were single (50.1%) or married (29.8%). Two additional sources of psychometric data were available for this sample. The Symptom Checklist–90 (SCL-90) (Derogatis, 1983) was completed at the same time as the MMPI-2 by 321 of the patients. For 295 patients, the primary therapist completed the Hopkins Psychiatric Rating Scale (HPRS) (Derogatis, 1983) within 72 hours of admission. The HPRS was developed as a rating-based supplement to the SCL-90. The clinician provides severity ratings on 17 symptom dimensions as well as a global severity rating, using a 7-point scale with anchor points. The HPRS is similar in format and content to the more familiar Brief Psychiatric Rating Scale. No

interrater reliability data were available as the HPRS results were extracted from patient archives.

The second cross-validation sample consisted of 266 outpatients who completed the MMPI-2 between 1995 and 1999 at an outpatient mental health clinic located on the campus of Fairleigh Dickinson University. The clinic occasionally serves students from the university, but the majority of the caseload is drawn from the surrounding urban and suburban communities. The MMPI-2 was administered either as part of the standard intake procedure for therapy or as part of a psychiatric or forensic evaluation. Average age was 33.4 years ($SD = 12.1$), with a mean education level of 14.9 years ($SD = 2.7$). The sample was 37.2% male and 62.8% female. The majority of participants were either married (21.4%) or single (58.6%); most were White (77.4%), with Blacks (8.3%) representing the next largest ethnic group.

Procedure

The derivation of the short form of the inventory proceeded as follows. All items from the *K* scale were included. Because portions of the *K* score are added to the raw scores for 5 of the clinical scales, Moreland (1984) recommended this as the most cost-effective method of enhancing congruence for MMPI short forms. Short scales were developed for the 2 remaining standard validity scales (*L* and *F*); 2 supplementary validity scales, True Response Inconsistency (TRIN) and Variable Response Inconsistency (VRIN); the 10 standard clinical scales, and the 15 content scales.

The short scales were developed as follows. Using the derivation sample, items within each scale were ranked on the basis of their corrected item-total correlations. For VRIN and TRIN, items represented ratings of 0 or 1 depending on whether the criterion for a particular item pair was met. The initial short scale consisted of the 10 items with the highest item-total correlations. In cases where the full scale consists of less than 20 items, the initial short scale included 50% of the items. Separate versions of scale 5 were generated for men and women.

For each scale, raw scores for the full-length scale were then regressed onto raw scores for the short scale to generate a best estimate of full-length raw score for each short-scale score. The regression was conducted separately for men and women. The estimated and actual full-length raw scores were then transformed to *T* scores, using *K* correction where appropriate. *T* scores were dichotomized based on whether they exceeded 64, except in the case of three validity scales where alternative cut scores were considered more appropriate (Butcher et al., 1989). *F* was dichotomized based on whether the *T* score exceeded 89. VRIN was dichotomized depending on whether the raw score exceeded 12. Finally, patients were split into three groups on TRIN: raw score less than 6, 6 to 12, or greater than 12.

The criterion for an acceptable short scale was set at an agreement rate of .90 with the full scale concerning whether the scale was elevated ($T \geq 65$). In other words, if both the full-scale raw score and the best estimate of that raw score based on the short scale were associated with a *T* score above 64, the two scales were considered in agreement; the same was true if both scores were associated with a *T* score less than 65. The criterion for a shortened scale was met if agreement was reached in 90% of cases or more.

Most short scales did not meet the criterion at first. In subsequent iterations, additional items were added until the criterion was reached. The number of items added in each iteration was a subjective estimate of the number of additional items needed based on previous iterations, although the number never exceeded 10. At times, the extended short form would exceed the criterion to the point that it would have been possible to reduce the length without failing the criterion. However, the scales were not shortened to enhance the likelihood of successful cross-validation.

For five of the clinical scales (2, 3, 4, 6, and 9), the number of items needed to achieve the criterion represented a relatively large proportion of the items on the full scale. We noted that these represent the five scales for which Weiner and Harmon (Weiner, 1948) identified obvious and subtle subscales. Given prior evidence suggesting that the subtle subscales seem to be measuring different domains than the obvious subscales (see Hollrah, Schlottmann, Scott, & Brunetti, 1995, for a review), we hypothesized that developing separate short forms for the subtle and obvious subscales that were then combined to produce the full-scale estimate could reduce the number of items necessary to achieve the criterion. This proved to be the case for four of the five scales. The number of items needed to achieve the criterion for Scale 4 was smaller when the short scale was based on the full scale rather than on the subtle and obvious subscales. This pattern replicated in both cross-validation samples. That is, in each case, the combination of subscale estimates was associated with a higher agreement rate than the full-scale estimate except in the case of Scale 4. Based on these findings, it was decided that the subtle and obvious subscales should be estimated separately for Scales 2, 3, 6, and 9.

One final design issue involved overlapping items. In many cases, items were included in the short forms of some scales on which they are scored but were omitted from others. Because these items would be administered anyway, the potential existed for increasing congruence further without increasing the number of items adminis-

tered by adding these items to scales on which they were omitted. Exploratory analyses with several scales suggested that the extra items added little to the agreement rates, however. Because this practice would have increased item overlap between the short scales, we decided against it.

It was found that 216 items were needed to achieve .90 agreement on each of the traditional 13 clinical and validity scales (including all 30 items from the K scale). An additional 81 items were needed to achieve agreement on the additional validity scales and the content scales. The version used to estimate the traditional 13 scales will be referred to as the MMPI-216 in the following, whereas the version used to estimate all 29 scales will be called the MMPI-297. For example, references to the MMPI-216 in the analyses indicate the analysis was restricted to the 13 traditional scales, whereas references to the MMPI-297 mean that at least some of the additional validity and content scales were also evaluated.

RESULTS

Additional Analyses With the Derivation Sample

Although the short form was designed to demonstrate congruence with the full-length inventory only in terms of scale elevation, four sets of analyses were conducted to explore for other forms of consistency. These criteria included the average difficulty of items, the prediction of low scores, high-point congruence, and high-point code congruence.

Item difficulty. The MMPI-2 manual (Butcher et al., 1989, Appendix I) provides descriptive data for each item in the inventory. Four of the variables are related to item difficulty. The first indicates whether an item is stated in positive or negative terms. The second variable is a rating of sentence structure, where 1 represents a simple sentence, 2 represents a compound sentence, and 3 represents a complex sentence. Sentence length is simply the number of words in the sentence. Finally, a Lexile value was computed for each item. This value represents a quantitative rating of reading difficulty based on several sentence characteristics, including sentence complexity and word familiarity. Table 1 presents results from comparisons of the short form and the remaining items on these four variables. Differences between the MMPI-297 items and the remaining items on proportions, means, and standard deviations were consistently small, and none of the statistical comparisons were significant. Overall, the short form seems to be consistent with the full inventory in terms of item difficulty.

TABLE 1
Item Difficulty Ratings for the Short Form and Full Inventory

	Negation	*Sentence Structure*	*Sentence Length*	*Lexile Score*
Short form	.13	2.0	11.3	480.5
		(1.0)	(5.1)	(296.7)
Remaining items	.12	2.0	11.2	506.6
		(1.0)	(4.7)	(297.2)

NOTE: Values for negation represent the proportion of items stated in the negative. All other columns contain means (and standard deviations). The difference in proportions was evaluated using $\chi^2(1, N = 567)$. Differences between column means were evaluated using *t* tests ($df = 565$). None were significant.

Prediction of low scores. Both clinical lore and research (e.g., Graham, Ben-Porath, & McNulty, 1997) suggest that relatively low scores on the clinical scales, although uncommon, can represent clinically useful information. The next set of analyses investigated whether the shortened scales could identify scores below the normal range. For each of the 10 traditional clinical scales and 15 content scales, profiles were dichotomized based on whether the *T* score for that scale was less than 40. Out of 25 comparisons, the agreement rate failed to exceed .90 only in the cases of Scale 5 (.89) and the Bizarre Ideation content scale (.88). It was therefore concluded that prediction of low scores could be evaluated as part of the cross-validation of the MMPI-297.

High-point congruence. The third analysis examined how well the MMPI-216 agreed with the long form in terms of which of the traditional clinical scales was the most elevated. This analysis excluded Scales 5 and 0. The short and full inventories were in agreement in only 52.6% of cases. It was therefore concluded that the MMPI-216 is not an acceptable basis for identifying the single most elevated clinical scale in the full inventory.

High-point code congruence. The high-point code is defined by the combination of scales that are most elevated and is often considered the key element in the interpretation of the clinical scales (e.g., Greene, 2000). The next set of analyses evaluated high-point code congruence between the MMPI-216 and the MMPI-2. No universal set of rules exists for high-point code classification (McGrath & Ingersoll, 1999a), but the following rules were chosen for their common use in previous research projects. First, the eight clinical scales excluding 5 and 0 were considered. Profiles in which all eight scales were less than 65 *T* were classified "Within Normal Limits." If only one scale exceeded 64 *T*, the profile was considered a spike profile. Finally, if two or more scales exceeded 64 *T*, it was classi-

fied according to the 2-point code. In cases of ties, lower numbered scales were given precedence.

Agreement rate was quite poor, only reaching .45. A series of additional analyses were conducted to evaluate whether this was a function of the code definition strategy. Two 3-point codes were added to the coding strategy, the 2-7-8 and 1-2-3 codes. For example, patients in the 2-7, 2-8, and 7-8 code groups were combined into a single group if 2, 7, and 8 were the three highest elevations above 64 *T*. This addition did not markedly affect the results, so the 3-point codes were eliminated.

The definition was then revised to allow for partial code matches. A partial match was based on two criteria. First, at least one scale had to be common to the short-form and full-length codes. Second, Lachar (1974) classified high-point codes depending on whether they are likely to reflect neurotic, character, or psychotic pathology, or are indeterminate. A partial match required consistency in terms of the class of pathology most likely represented by the code. Despite the more liberal standard, the agreement rate for partial matches was only .65.

The final modification was the restriction to well-defined codes. Graham, Timbrook, Ben-Porath, and Butcher (1991; see also Ben-Porath & Tellegen, 1995) have recommended only applying high-point code interpretive data when there is reason to believe the code is reliable. The standard they adopted for an acceptably reliable code was a *T* score for the less elevated scale in the code that is at least 5 points higher than the *T* score for the next highest clinical scale. For 800 of the patients in the derivation sample, 266 of the high-point codes defined previously proved to be well-defined. Among these cases, 56% demonstrated an exact code match, and 79% demonstrated a partial code match.

On the basis of these supplemental findings, we concluded that the failure of the MMPI-216 to predict the MMPI-2 high-point code is consistent across a variety of reasonable approaches to defining codes. Accordingly, use of the MMPI-216 to predict high-point codes cannot be recommended.

Cross-Validation

Table 2 provides statistics comparing the MMPI-297 scales with the MMPI-2 scales in the inpatient cross-validation sample. Table 3 provides the same information for the outpatient sample. Congruence concerning elevation of the scale is referred to as hit rate in the tables and may be considered the most important criterion for evaluating the validity of the shortened scales. However, additional statistics are presented to provide a more complete picture of the short form's effectiveness. Positive predictive power refers to the proportion of cases with an elevated *T* score on the short scale that had an elevated *T* score on the long form; negative predictive power refers to the proportion of cases with an unelevated *T* score on the short scale that were also not elevated on the long form of the scale.

The first statistic provided is the correlation between the full and short scale. Except in three cases (VRIN, TRIN, and Scale 5), these were all .80 or higher. Most were above .90.

The correlations are followed by a series of statistics associated with agreement on whether a scale is elevated. Overall hit rates were consistently very high. In neither sample did the hit rate for any scale fall below 88.6, and most values exceeded .90. Not surprisingly, positive and negative predictive power and kappa were more variable, although in most cases, the results were still acceptable. Kappa was poorest for those scales where elevated scores were particularly uncommon: TRIN, VRIN, and Scale 5.

The next set of statistics address the issue of congruence for low scores on the clinical and content scales. Results were more variable than for elevated scores. This finding could be expected given the relative infrequency of low scores. Even so, overall hit rates were generally quite high, with most reaching .90 or greater.

The next two columns in Tables 2 and 3 provide coefficient alphas for all scales except VRIN and TRIN. Reliability estimates for the full-length scales were similar to those reported for the clinical and content scales in the MMPI-2 manual (Butcher et al., 1989, Appendix D). Although the reliabilities for the MMPI-2 scales were consistently larger, the differences between MMPI-2 and MMPI-297 reliabilities were in general quite small. In both samples, the reliabilities for the shortened versions of Scale 4 and the male version of Scale 5 were actually higher than for the full-length scales.

The final column indicates *T* score agreement rates for the scales. These analyses were conducted to evaluate the proportion of cases in which the MMPI-297 and MMPI-2 scales generated consistent *T* scores. Based on the well-defined high-point code literature (Ben-Porath & Tellegen, 1995; Graham et al., 1991), a difference of 5 *T* points or less was considered a consistent outcome. A difference greater than 5 *T* was considered inconsistent. In general, consistency occurred in only about 70% of cases and was particularly low for Scales 5 and 8.

The results of the congruence analyses indicated that the shortened scales were highly correlated with the full-length scales. Hit rates were generally high either for the prediction of elevated scores or low scores. Although reliability tended to be lower for the shortened scales, the differences tended to be small. The MMPI-297 scales should

TABLE 2
Statistics for the Inpatient Cross-Validation Sample

		Elevated				*Low*				*Alpha*		T *Score*
	r	*HR*	*PPP*	*NPP*	κ	*HR*	*PPP*	*NPP*	κ	*Full*	*Short*	*Agreement*
Validity scales												
L	.89	93.0	94.2	92.9	.65					.61	.53	85.1
F	.84	90.1	76.7	94.1	.72					.85	.71	30.8
VRIN	.55	95.1	70.0	95.5	.29							
TRIN	.55	97.0	80.0	97.1	.28							
Clinical scales												
1	.93	90.7	91.6	89.9	.81	98.6	81.8	98.9	.66	.88	.83	69.5
2	.94	91.6	92.0	90.7	.80	98.6	100.0	98.6	.30	.82	.70	66.9
3	.94	89.8	87.5	92.6	.80	97.1	20.0	97.7	.09	.70	.65	75.1
4	.92	91.0	93.2	86.4	.80	99.8		99.8		.73	.85	70.8
5	.60	88.8	55.6	89.2	.10	89.4	46.2	90.2	.12	.49, .44	.50, .32	50.9
6	.95	90.6	91.1	89.7	.80	98.2	60.0	98.8	.49	.66	.57	70.7
7	.94	90.9	92.7	86.9	.79	98.9	33.3	99.5	.36	.93	.90	69.8
8	.91	88.6	89.9	85.6	.74	99.1		99.2		.93	.91	52.0
9	.94	90.3	92.1	89.7	.76	96.2	79.3	97.0	.63	.70	.68	69.9
0	.90	88.6	78.3	93.0	.73	94.2		94.2		.88	.87	70.2
Content scales												
ANX	.94	91.8	92.3	90.7	.82	98.8		98.8		.88	.84	81.5
FRS	.91	92.2	85.3	93.7	.75	93.5	81.8	93.9	.43	.80	.74	66.6
OBS	.92	90.0	86.2	92.5	.79	93.6		93.6		.83	.80	71.3
DEP	.94	93.5	94.7	90.2	.84	97.9		97.9		.93	.89	75.8
HEA	.91	89.7	87.8	90.9	.79	97.0		97.0		.87	.84	63.8
BIZ	.93	92.4	93.5	92.0	.82	87.5		87.5		.86	.78	60.5
ANG	.93	92.2	84.4	95.3	.81	94.7	56.7	96.5	.47	.81	.77	73.9
CYN	.94	92.2	81.7	94.9	.76	96.8	89.5	97.3	.75	.87	.82	83.4
ASP	.92	91.8	89.4	92.3	.75	95.7	70.8	97.7	.69	.82	.75	70.7
TPA	.92	93.0	84.3	94.6	.75	93.5	97.6	93.2	.62	.79	.77	77.1
LSE	.95	90.4	84.7	96.1	.81	95.6		95.6		.89	.85	78.3
SOD	.95	93.3	87.9	95.3	.83	90.9		90.9		.89	.85	80.7
FAM	.94	89.8	85.3	94.1	.80	96.7	75.0	97.6	.64	.85	.80	74.0
WRK	.93	91.6	91.5	91.9	.83	94.7		94.7		.90	.88	69.0
TRT	.93	89.4	90.6	88.3	.79	94.2		94.2		.88	.85	69.0

NOTE: *r* = correlation between the full-length and short-form scores. *T* score agreement refers to the percentage of cases in which short-form and full-length *T* scores differed by 5 points or less. All other columns are based on dichotomized values. Empty cells indicate cases where no short form *T* scores were less than 40. HR = hit rate; PPP = positive predictive power; NPP = negative predictive power; κ = kappa coefficient; VRIN = Variable Response Inconsistency; TRIN = True Response Inconsistency; ANX = Anxiety; FRS = Fears; OBS = Obsessiveness; DEP = Depression; HEA = Health Concerns; BIZ = Bizarre Mentation; ANG = Anger; CYN = Cynicism; ASP = Antisocial Practices; TPA = Type A; LSE = Low Self-Esteem; SOD = Social Discomfort; FAM = Family Problems; WRK = Work Interference; TRT = Negative Treatment Indicators.

not be used to estimate the exact value of the MMPI-2 scales, because more than 30% of estimates differed from true *T* scores by 5 points or more.

Concurrent Validity

The final set of analyses compared the concurrent validity of the MMPI-297 and MMPI-2 scales using the HPRS and SCL-90 data available for the inpatient cross-validation sample. Potentially invalid protocols were excluded from this set of analyses. Protocols on which the short-form *L* or *K* score was greater than or equal to 65 *T*, or on which the *F* score was greater than or equal to 90, were considered potentially invalid. The first author selected SCL-90 scales and HPRS items that could be considered conceptually related to each of the MMPI clinical and content scales. There were 75 relationships that were identified. Zero-order correlations were then computed between these criterion scales and the MMPI-2 scales. In those cases where this correlation significantly differed from zero at $p < .05$, the correlation was also computed for the MMPI-297 scale. This produced 46 pairs of correlations. Table 4 provides the results from those comparisons.

Despite a strategy that was biased toward the MMPI-2 scales, because those correlations had to be significant before the relationship was considered for the analysis, only

TABLE 3
Statistics for the Outpatient Cross-Validation Sample

		Elevated				*Low*				*Alpha*		T *Score*
	r	*HR*	*PPP*	*NPP*	κ	*HR*	*PPP*	*NPP*	κ	*Full*	*Short*	*Agreement*
Validity scales												
L	.90	91.4	100.0	90.0	.71					.67	.60	84.2
F	.79	94.7	42.1	98.8	.51					.83	.63	25.6
VRIN	.52	98.1	66.7	98.5	.44							
TRIN	.57	97.4	50.0	97.7	.22							
Clinical scales												
1	.93	91.4	78.8	96.8	.79	97.0	100.0	96.9	.62	.85	.83	68.8
2	.94	93.6	85.7	97.7	.86	98.5	100.0	98.5	.66	.82	.70	69.5
3	.93	91.7	80.8	96.3	.79	93.2	33.3	95.3	.22	.68	.63	74.8
4	.94	92.9	93.5	92.5	.85	96.6	03.4	96.6		.80	.87	75.6
5	.60	91.4	100.0	91.2	.28	86.1	33.3	88.6	.12	.65, .36	.68, .35	46.6
6	.93	90.2	79.0	95.1	.76	94.7	86.7	95.2	.62	.56	.48	69.2
7	.95	90.6	80.7	95.5	.78	95.5	76.9	96.4	.60	.92	.90	74.4
8	.92	89.5	81.0	95.0	.78	95.5	50.0	95.8	.13	.92	.90	57.5
9	.92	92.9	80.0	94.2	.64	92.5	76.5	93.6	.53	.64	.58	72.2
0	.91	94.4	76.7	97.8	.78	85.3		85.3		.88	.88	71.1
Content scales												
ANX	.92	91.0	84.3	94.0	.79	93.6		93.6		.89	.84	71.4
FRS	.91	95.5	86.2	96.6	.78	94.4	86.7	94.8	.61	.79	.73	72.2
OBS	.93	92.5	80.0	95.4	.75	87.2		87.2		.84	.80	72.2
DEP	.93	89.5	72.8	98.3	.75	94.4		94.4		.92	.87	64.7
HEA	.90	91.7	75.4	97.0	.76	97.7		97.7		.82	.84	69.5
BIZ	.88	92.9	80.0	94.2	.64	78.2		78.2		.80	.66	54.9
ANG	.92	93.6	64.3	99.1	.73	92.1	80.0	92.8	.50	.78	.75	77.1
CYN	.95	96.2	79.5	99.1	.84	91.0	81.8	91.8	.55	.88	.84	85.7
ASP	.91	93.2	76.7	95.3	.68	95.5	86.2	96.6	.78	.79	.73	77.1
TPA	.91	94.7	70.6	96.4	.60	90.2	79.2	91.3	.54	.72	.72	86.5
LSE	.96	95.5	85.7	98.5	.87	86.5		86.5		.90	.85	82.7
SOD	.94	95.1	88.6	96.1	.80	84.6		84.6		.88	.84	81.6
FAM	.95	93.6	84.6	97.3	.84	94.4	76.2	95.9	.65	.85	.81	83.5
WRK	.95	93.6	82.4	97.5	.83	86.8		86.8		.91	.88	75.6
TRT	.94	94.0	92.0	94.4	.81	80.1		80.1		.88	.85	72.9

NOTE: *r* = correlation between the full-length and short-form scores. *T* score agreement refers to the percentage of cases in which short-form and full-length *T* scores differed by 5 points or less. All other columns are based on dichotomized values. Empty cells indicate cases where no short form *T* scores were less than 40. HR = hit rate; PPP = positive predictive power; NPP = negative predictive power; κ = kappa coefficient; VRIN = Variable Response Inconsistency; TRIN = True Response Inconsistency; ANX = Anxiety; FRS = Fears; OBS = Obsessiveness; DEP = Depression; HEA = Health Concerns; BIZ = Bizarre Mentation; ANG = Anger; CYN = Cynicism; ASP = Antisocial Practices; TPA = Type A; LSE = Low Self-Esteem; SOD = Social Discomfort; FAM = Family Problems; WRK = Work Interference; TRT = Negative Treatment Indicators.

eight pairs of correlations differed significantly. Of those, six favored the MMPI-2 scale, whereas in two cases the correlation with the MMPI-297 scale was higher. It is noteworthy that these two both involved the Scale 4 short form. Instances where the MMPI-2 scale proved superior to the MMPI-297 scale tended to involve the SCL-90 Somatization, Paranoid Ideation, and Phobic Anxiety Scales. Only one of the significant differences involves clinician ratings, and only two involved the content scales. The differences between the mean and median correlations were small; both differences in the proportion of overlapping variance were less than .025.

Comparison With the Dahlstrom-Archer Short Form

Dahlstrom and Archer (2000) suggested their short form can be used as a "benchmark" for evaluating the effectiveness of short forms. For purposes of benchmarking, the MMPI-180 was scored in both cross-validation samples, and three of the key criteria from Tables 2 and 3 were used to evaluate congruence: correlation, hit rate for elevated scores, and *T* score agreement rate. The results of these analyses may be found in Table 5. Correlations were generally very similar; in most cases, the correlation for the MMPI-180 differed from that for the MMPI-216 by no

TABLE 4
Criterion-Related Validity

	Full-Scale r	*Short-Form* r
Scale 1		
SCL-90 Somatization	.612	.578[a]
Scale 2		
SCL-90 Depression	.708	.685
HPRS Depression	.327	.307
HPRS Sleep Disturbance	.181	.225
HPRS Psychomotor Retardation	.214	.247
HPRS Abjection-Disinterest	.282	.276
Scale 3		
SCL-90 Somatization	.541	.506[a]
Scale 4		
SCL-90 Interpersonal Sensitivity	.327	.465[b]
SCL-90 Hostility	.293	.359[b]
Scale 6		
SCL-90 Interpersonal Sensitivity	.450	.434
SCL-90 Paranoid Ideation	.471	.423[a]
HPRS Interpersonal Sensitivity	.204	.238
Scale 7		
SCL-90 Obsessive-Compulsive	.585	.554
SCL-90 Anxiety	.546	.518
SCL-90 Phobic Anxiety	.446	.392[a]
Scale 8		
SCL-90 Paranoid Ideation	.428	.350[a]
SCL-90 Psychoticism	.513	.493
Scale 9		
HPRS Excitement	.286	.260
HPRS Euphoria	.265	.199[a]
Scale 0		
SCL-90 Interpersonal Sensitivity	.517	.477
HPRS Interpersonal Sensitivity	.284	.265
ANX		
SCL-90 Obsessive-Compulsive	.550	.568
SCL-90 Anxiety	.607	.553[a]
SCL-90 Phobic Anxiety	.489	.455
HPRS Anxiety	.189	.168
HPRS Phobic Anxiety	.160	.130
FRS		
SCL-90 Obsessive-Compulsive	.305	.246
SCL-90 Anxiety	.368	.334
SCL-90 Phobic Anxiety	.425	.396
OBS		
SCL-90 Obsessive-Compulsive	.522	.517
SCL-90 Anxiety	.444	.456
SCL-90 Phobic Anxiety	.371	.373
DEP		
SCL-90 Depression	.713	.682
HPRS Depression	.271	.318
HPRS Abjection-Disinterest	.180	.203
HEA		
SCL-90 Somatization	.672	.644
BIZ		
SCL-90 Paranoid Ideation	.590	.575
SCL-90 Psychoticism	.561	.577
ANG		
SCL-90 Interpersonal Sensitivity	.391	.421
SCL-90 Hostility	.683	.694
HPRS Interpersonal Sensitivity	.161	.180
HPRS Hostility	.185	.181
TPA		
SCL-90 Hostility	.431	.464
HPRS Hostility	.168	.172
SOD		
SCL-90 Interpersonal Sensitivity	.446	.383[a]
HPRS Interpersonal Sensitivity	.239	.263
Median *r*	.426	.399
Mean *r*	.400	.391

NOTE: Significance tests were *t* tests for dependent correlations tested at $p < .05$. SCL-90 = Symptom Checklist–90; HPRS = Hopkins Psychiatric Rating Scale; ANX = Anxiety; FRS = Fears; OBS = Obsessiveness; DEP = Depression; HEA = Health Concerns; BIZ = Bizarre Mentation; ANG = Anger; TPA = Type A; SOD = Social Discomfort.
a. Full version is significantly superior to short form.
b. Short form is significantly superior to full version.

TABLE 5
Congruence Data for the Dahlstrom-Archer Short Form

	Inpatients			*Outpatients*		
Scale	r	*HR*	T *Score Agreement*	r	*HR*	T *Score Agreement*
L	.93	93.8	89.2	.94	94.3	90.2
F	.93	89.2	46.0	.94	95.5	59.8
K	.91	100.0	68.7	.92	100.0	71.1
1	.96	93.9	85.0	.96	96.2	86.5
2	.96	91.8	64.6	.96	95.5	55.6
3	.96	90.1	57.6	.94	92.1	43.6
4	.91	88.9	60.5	.93	90.1	66.5
5	.66	81.6	40.7	.75	83.8	47.0
6	.88	84.7	41.3	.84	85.7	48.1
7	.89	86.8	45.4	.89	86.8	56.0
8	.91	85.4	29.8	.91	88.3	44.7
9	.87	81.1	45.4	.85	77.1	51.9
0	.86	84.2	57.1	.85	87.6	54.5

NOTE: *r* = correlation between the full-length and short-form scores. HR = hit rate for dichotomized values. *T* score agreement refers to the percentage of cases in which short-form and full-length *T* scores differed by 5 points or less.

more than a few hundredths. However, hit rate statistics supported the superiority of the MMPI-216. Congruence rates were slightly higher for scales *L*, 1, 2, and 3 on the MMPI-180; scales that happen to be represented by more items on the MMPI-180 than on the MMPI-216. In all other cases, congruence was at least as good for the MMPI-216 as it was for the MMPI-180 regardless of which short form included more items. This pattern replicated for *T* score agreement, where the new short form was consistently superior to the Dahlstrom-Archer short form except for scales *L*, *F*, and 1.

DISCUSSION

The MMPI-297 represents a middle road between the MMPI-180 and the adaptive testing approach. It differs from the former in at least three ways. First, the MMPI-297 demonstrates reasonable congruence with the MMPI-2 on individual scale elevation, a key element of profile interpretation. This was achieved by increasing the number of items beyond that found in previous paper-and-pencil short forms. Second, it achieves this goal with only 36 more items than the MMPI-180 and with a minimum of item overlap. There are 176 instances of duplicate item scoring on the MMPI-180, whereas there are only 83 such instances on the MMPI-216. Third, content and newer validity scales can also be estimated. Given the relatively large number of items included in the MMPI-216 and MMPI-297, future research may identify other scales that can be effectively estimated from the short forms.

At the same time, the efficiency of the MMPI-297 is greater than that for the recommended adaptive testing strategy. The number of items administered is reduced by 47.6%, compared with a maximum average reduction of 31% in studies using adaptive testing. The MMPI-297 achieves this reduction without requiring computer administration, although the scoring of the short form is somewhat cumbersome and would be enhanced by the use of computer algorithms for the task. It is also important to remember that the recommended adaptive testing model results in perfect congruence for scale elevation and high-point code. The MMPI-216 is not recommended as a predictor of high-point codes.

The MMPI-216 and MMPI-297 are recommended as reasonable estimators of scale elevations when computer-based administration is not feasible. Preliminary data also suggest the shortened scales are similar to MMPI-2 scales in their ability to predict criterion variables. Results were particularly positive for Scale 4, where the shortened scale demonstrated higher reliability and validity coefficients than the original scale. Results were also encouraging for the content scales. Not surprisingly, given the greater homogeneity of the items in these scales, relatively fewer items were needed to achieve congruence.

One way to reduce length even further would have involved eliminating subtle items completely, because previous research has raised questions about their validity (Weed, Ben-Porath, & Butcher, 1990). However, because the goal of the development strategy was to predict elevation accurately, and the subtle items influence the elevation of the clinical scales, eliminating these items as a group would have made it more difficult to achieve congruence between full-length scales and short scales.

There are several practical objections that may be raised to the new short form. One is the complexity involved in scoring the instrument. To address this issue, a Windows-based scoring program has been developed by the first author that allows hand entry of the items and generates best estimates of full-inventory raw scores.[1] The program accepts input from items numbered consecutively (if a question set is created that only includes the short form items) or according to the standard item numbering (if the items are administered using standard materials, with the omitted items blocked out).

Another issue is the low rate of congruence with full-scale high-point codes. Because the high-point code is often seen as the starting point for MMPI interpretation (e.g., Greene, 2000), the inability to assume high-point congruence may be seen as a serious obstacle and raises questions whether the short form is superior to other instruments of similar length, such as the Personality Assessment Inventory (Morey, 1991). There are several reasons for using the short form presented here. Research suggests that the increment in validity offered by high-point codes over individual scale elevations is quite small (McGrath & Ingersoll, 1999b), and most of the research on the validity of the MMPI (including research on the criterion-related validity of the validity scales) focuses on elevations of individual scales rather than on the high-point code. Even without relying on the high-point code literature, the user of the MMPI-216 or MMPI-297 still has a larger body of validational research available as the basis of interpretation than may be said of any alternative inventory of similar length.

Smith, McCarthy, and Anderson (2000) have recently provided guidelines for identifying well-developed short forms. These may be used to evaluate the development strategy employed in the present study:

1. The parent measure should be of sufficient validity to justify developing a short form. This is clearly true of the MMPI.
2. Content coverage of the full measure should be preserved. As with most of the MMPI short forms, this was achieved by developing shortened versions for individual scales rather than with the inventory as a whole. The scale set is actually larger than that of any other existing MMPI short form.
3. The reliability of each short-form scale should be demonstrated. These data may be found in Tables 2 and 3.
4. Overlap between the short and full measures should be demonstrated with independent administrations of the two forms. This was not possible in the present study.
5. The factor structure of the full measure should be maintained. Correlations between the shortened and full-length scales were high enough to suggest a reasonable similarity in factor structure.
6. Content coverage within scales should be preserved. This occurred in those heterogeneous

scales where subtle and obvious subscales were estimated separately.

7. Validity of the short-form scales should be demonstrated in a sample not administered the full measure. Again, this was not possible in the present study.
8. Consistency in classification rates should be demonstrated. The MMPI-297 is particularly noteworthy for this criterion, as classification congruence was the primary justification for its development. However, it should be noted that no attempt was made to demonstrate consistency in classification on criterion variables such as diagnosis.
9. A meaningful savings in time or resources should be demonstrated. Greene (2000) estimated that more than 90% of respondents complete the MMPI-2 in 60 to 90 minutes. A reduction in length of more than 40% would therefore translate into a time savings of about 30 minutes per administration, with minimal loss in the classification accuracy or concurrent validity for individual scales. Savings can also be based on comparison with major alternatives to the MMPI-2. The MMPI-297 requires more items than the Millon Clinical Multiaxial Inventory, but fewer items than the Personality Assessment Inventory.

The failure to administer the short and full forms separately represents the only significant methodological failure in the development-validation model employed here. This issue should be addressed in future research on the MMPI-297.

In summary, the MMPI-297 offers what we consider to be a reasonable substitute for the MMPI-2 when the latter is not practical. It allows the clinician to estimate elevations on individual scales, so that MMPI research on the interpretation of elevated scale scores can be employed. However, subsequent research should focus on the effectiveness of the MMPI-297 when administered independently from the full inventory.

NOTE

1. The program is available by contacting the first author.

REFERENCES

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

Ben-Porath, Y. S., & Tellegen, A. (1995). How (not) to evaluate the comparability of MMPI and MMPI-2 profile configurations: A reply to Humphrey and Dahlstrom. *Journal of Personality Assessment*, *65*, 52-58.

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

Butcher, J. N., & Hostetler, K. (1990). Abbreviating MMPI item administration: What can be learned from the MMPI for the MMPI-2? *Psychological Assessment*, *2*, 12-21.

Dahlstrom, W. G., & Archer, R. P. (2000). A shortened version of the MMPI-2. *Assessment*, *7*, 131-141.

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

Faschingbauer, T. R. (1974). A 166-item short form for the group MMPI: The FAM. *Journal of Consulting and Clinical Psychology*, *42*, 645-655.

Gallucci, N. T. (1986). General and specific objections to the MMPI. *Educational and Psychological Measurement*, *46*, 985-988.

Graham, J. R., Ben-Porath, Y. S., & McNulty, J. L. (1997). Empirical correlates of low scores on the MMPI-2 scales in an outpatient mental health setting. *Psychological Assessment*, *9*, 386-391.

Graham, J. R., Timbrook, R. E., Ben-Porath, Y. S., & Butcher, J. N. (1991). Code-type congruence between MMPI and MMPI-2: Separating fact from artifact. *Journal of Personality Assessment*, *57*, 205-215.

Greene, R. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Groth-Marnat, G. (1997). *Handbook of psychological assessment* (3rd ed.). New York: John Wiley.

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.

Hollrah, J. L., Schlottmann, R. S., Scott, A. B., & Brunetti, D. G. (1995). Validity of the MMPI subtle items. *Journal of Personality Assessment*, *65*, 278-299.

Kincannon, J. C. (1968). Prediction of the standard MMPI scale scores from 71 items: The Mini-Mult. *Journal of Consulting and Clinical Psychology*, *32*, 319-325.

Lachar, D. (1974). *The MMPI: Clinical assessment and automated interpretation*. Los Angeles: Western Psychological Services.

MacDonald, G. L. (1952). A study of the shortened group and individual forms of the MMPI. *Journal of Clinical Psychology*, *8*, 308-309.

McGrath, R. E., & Ingersoll, J. (1999a). Writing a good cookbook. I: A review of MMPI high-point code studies. *Journal of Personality Assessment*, *73*, 149-178.

McGrath, R. E., & Ingersoll, J. (1999b). Writing a good cookbook. II: A synthesis of MMPI high-point code study effect sizes. *Journal of Personality Assessment*, *73*, 179-198.

Moreland, K. L. (1984). A cost-effective means of improving the statistical validity of an MMPI short form. *Journal of Clinical Psychology*, *40*, 134-136.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Overall, J. E., & Gomez-Mont, F. (1974). The MMPI-168 for psychiatric screening. *Educational and Psychological Measurement*, *34*, 315-319.

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.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*, 102-111.

Streiner, D. L., & Miller, H. R. (1986). Can a good short form of the MMPI ever be developed? *Journal of Clinical Psychology*, *42*, 109-113.

Weed, N. C., Ben-Porath, Y. S., & Butcher, J. N. (1990). Failure of Wiener and Harmon Minnesota Multiphasic Personality Inventory (MMPI) subtle scales as personality descriptors and as validity indicators. *Psychological Assessment*, *2*, 281-285.

Weiner, D. N. (1948). Subtle and obvious keys for the MMPI. *Journal of Consulting Psychology*, *12*, 164-170.

Robert E. McGrath is a professor of psychology at Fairleigh Dickinson University. He received his Ph.D. in clinical psychology at Auburn University in 1984. He has published numerous articles in the area of psychological assessment, particularly with the Minnesota Multiphasic Personality Inventory (MMPI), and is currently associate editor for the *Journal of Personality Assessment*.

Ray Terranova is currently completing his predoctoral clinical internship at Trenton Psychiatric Hospital in New Jersey. An active member of the Society of Personality Assessment, he plans on continuing research and clinical work in the area of assessment.

David L. Pogge received his bachelor's degree in psychology from Creighton University in 1979. He received his master's degree in psychology in 1983 and a Ph.D. in clinical psychology in 1986 from the University of New Mexico. He completed a 2-year internship and postdoctoral fellowship at the Payne Whitney Clinic of the New York Hospital-Cornell Medical Center. He has been director of psychology at Four Winds Hospital, an affiliate of the Albert Einstein College of Medicine, for the past 14 years and has been a clinical lecturer at Fairleigh Dickinson University for the past 11 years.

Celina Kravic is a 3rd-year student in Fairleigh Dickinson University's Ph.D. program in clinical psychology. Her primary research and clinical interest is in the area of cognitive and neuropsychological assessment of children diagnosed with attention deficit hyperactivity disorder.