JOURNAL OF PERSONALITY ASSESSMENT, *73*(2), 179–198

Writing a Good Cookbook: II. A Synthesis of MMPI High-Point Code System Study Effect Sizes

Robert E. McGrath and Joel Ingersoll

School of Psychology
Fairleigh Dickinson University

This article continues a review of the high-point code system studies of the MMPI. Using meta-analytic methods, we conducted an analysis of effect sizes associated with these studies. Effect sizes were on average small. The finding is inconsistent both with the image of the MMPI as a powerful clinical instrument and with at least some of the previous evidence on MMPI effect sizes. The finding is discussed in some detail, including potential sources of bias in the analysis, comparisons with previous statistical reviews, and reasons clinicians place a high degree of faith in interpretations based on the high-point code. We believe the existing high-point code system research may not adequately reflect the true validity of these codes.

The high-point code commonly serves as the starting point for the clinical interpretation of the Minnesota Multiphasic Personality Inventory (MMPI).[1] A large literature exists identifying clinical correlates of MMPI high-point codes, and this literature has reinforced the perception of the inventory as an empirically grounded clinical instrument. In particular, 10 published studies have been completed that address the interpretive significance of high-point codes in general, with implications for a general interpretive system for the MMPI. These high-point code system studies were inspired by Meehl's (1956) "Wanted—A Good Cookbook," in which the principles of actuarial personality description were first outlined.

The previous article (McGrath & Ingersoll, this issue) reviewed methodological characteristics of the 10 high-point code system studies. This article assumes familiarity with the issues raised in the previous article and extends the discussion

[1]As in McGrath and Ingersoll (this issue), the abbreviation *MMPI–1* is used for the original version (Hathaway & McKinley, 1983), *MMPI–2* for the revised version (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), and *MMPI* for the test in general.

with an investigation into the typical strength of relations between high-point codes and clinical variables as reported in those studies.

We have two reasons to suspect the mean effects would not be large despite the common image of the MMPI high-point codes as powerful clinical predictors. The first was methodological. In the previous article we indicated that the high-point code system studies were designed to maximize clinical usefulness, sometimes at a potential cost in power when compared to other actuarial studies. The second was empirical. In two high-point code system studies that incorporated cross-validation into the design, both noted the relatively low rate of replication even across large samples (Gynther, Altman, & Sletten, 1973a; Lewandowski & Graham, 1972). Other authors (Lachar, 1968; Williams & Butcher, 1989b) have expressed dissatisfaction with the extent to which their findings concurred with a priori expectations about the codes.

Using meta-analytic statistical procedures, we estimated and compared mean effect sizes from the 10 high-point code system studies.[2] After summarizing the results of these analyses, we discuss the implications of the findings for future research in this area and for clinical use of the MMPI.

METHOD

Studies

We based our analysis on statistical results from 10 high-point code system studies. McGrath and Ingersoll (this issue) described the common elements of the high-point code system studies and provided a detailed description of each of the 10 studies completed to date that meet the criteria. To summarize, the high-point code system studies are characterized by the development of an MMPI-based typology for a general psychiatric population using decision rules in which the relative elevation of clinical scales plays a central role and by the identification of clinical correlates for each class within the typology. The studies have included both adolescents and adults, in both inpatient and outpatient settings. Rules for high-point code classification have varied markedly across studies.

Procedure

The statistical procedures used in this study were developed in the context of meta-analysis. However, we limited the review to effects found in published studies because these studies have provided the basis for the general MMPI interpretive strategy based on high-point codes. We made no attempt to address

[2]The term *effect size* has multiple meanings in the literature. It is used here as a generic term for the class of statistics reflecting the strength of a relation.

the "file drawer problem" (Rosenthal, 1979), the inflation of mean effects when the analysis is limited to published studies; additionally, we made no attempt even to include studies that examined MMPI high-point codes but did not meet criteria for a high-point code system study. Considering the results of this analysis as an estimate of the "true" mean effect for high-point codes as predictors of clinical variables is, therefore, inappropriate. The purpose of this analysis is to provide a summary of data from an important series of studies as a basis for discussion of those studies.

The rest of this section is devoted to key issues in the conduct of the analyses. The Appendix provides information about the more technical aspects of the analyses and issues particular to each study.

The high-point code system studies are discovery oriented. The set of criterion variables typically included any clinical variables available to the researchers. For example, several of the studies examined the relation between IQ scores and high-point codes, even though high-point codes are not traditionally considered predictive of intellectual performance. If we included all the relations examined in the high-point code system studies in our analysis, regardless of whether a conceptual justification for the relationship existed, it would have drastically deflated mean effect sizes. The first step then was to identify those relations from each study that could be expected to exist on conceptual grounds.

We began this procedure with a review of two MMPI texts (Graham, 1987; Greene, 1991) that provide interpretive narratives for spike and 2-point codes. These narratives combine information drawn from the high-point code system studies, from other high-point code studies, and from clinical lore. We examined each combination of high-point code and criterion variable from each study, although we excluded some descriptors for reasons described in the Appendix. For each combination, we independently judged whether the relation would be expected to exist based either on the narrative statements for the code or on what each MMPI scale comprising the code is supposed to measure. In cases of 3-point or 4-point codes not discussed in the two texts, we reviewed the component 2-point codes. To be conservative, we both had to identify the relation as conceptually justified before it was included in the analyses.

We found the conceptual status of the relation sometimes difficult to judge. For example, one of the descriptors in the Missouri system was "Unrealistic hostility." Although this descriptor was not specifically noted in the clinical descriptions of the 6-8 code, Greene (1991) indicated that clients with this code are more likely to "evidence a thought disorder with paranoid features. … Any social relationship they do maintain will be tinged with resentment, suspiciousness, and hostility" (p. 280). Therefore, to consider it an expected correlate of the 6-8 was a reasonable extrapolation. The validity of many of these judgments could be questioned, particularly given a type of dementia we found sets in during the course of reviewing thousands of comparisons. Given the large number of judgments involved, com-

puting agreement or reliability statistics for the independent ratings was not practical.

A second problem involved negative relations, where a code might be expected to predict a lower intensity for or the absence of a symptom. To simplify computation, and because relations expected to be negative were relatively rare and more difficult to judge, we excluded them from the analysis. In the end, we identified 2,201 comparisons from the 10 studies as expected to result in a positive relation.

In every case, the coding variable was a dichotomous variable (code group vs. comparison group member). Criterion variables were either dichotomous or quantitative. We chose the correlation coefficient as the effect size statistic to be estimated because it is appropriate for either type of data.

A final problem that emerged was the incomplete statistics available from many of the studies. The only high-point code system study that provided the data needed to compute *r* for every relation of interest was Lachar (1968).[3] Marks and Seeman (1963) published much of their statistical data in two appendices. However, in cases where the criterion variable represented continuous data, they provided means without standard deviations. A large portion of the data is missing from a similar table in Gilberstadt and Duker (1965). Recent studies have been more consistent about reporting statistics, but only for significant outcomes.

To determine whether original data were still available, we contacted nine of the high-point code system study authors, including at least one author from eight of the studies (omitting Gilberstadt & Duker, 1965, and Lachar, 1968). We were provided with the complete set of means, standard deviations, and *t* values from both samples in the Lewandowski and Graham (1972) study. In all other cases the authors indicated the original data and statistics were not available, either because they had been destroyed or were not in a form suitable to our purposes.

Fortunately, in cases where a series of significance tests is conducted using the same sample, the proportion of significant outcomes provides an estimate of the mean power of the analyses. The mean effect size can then be estimated based on the sample size, alpha level, and estimated mean power (Hedges & Olkin, 1980). In those studies for which complete data were not available, using this procedure meant identifying which predicted relations were significant. Significant relations sometimes had to be inferred from a narrative provided for the code. The proportion of the expected relations that resulted in significant outcomes provided the estimate of the mean power for those tests.

Larger groupings of analyses were preferred to generate more reliable estimates of power. One implication of using this method is that some of the effect sizes used for our analyses represented estimated means for sets of analyses. Effect sizes were weighted by both the number of comparisons represented and by the sample

[3]Although Graham, Ben-Porath, and McNulty (1999) provided a complete set of descriptive statistics for their analyses, these tables were not available until this article was in press.

size when generating mean effect sizes across studies. Traditional methods for statistically evaluating the results of a meta-analysis, such as significance tests or multiple regression, were not available.

RESULTS

Mean Effects by Study

Table 1 provides the mean correlation coefficient from each study. With two exceptions, the means are less than .10. The larger mean effects in the Gilberstadt and Duker (1965) and Graham et al. (1999) studies probably reflect methodological differences. The former study excluded patients whose case histories were not "representative" of the code as a whole. Halbower also used this strategy in the very first cookbook study (Meehl, 1956), where the exclusion of cases was intended to reduce the impact of sampling error and faulty clinical judgment on the clinical description of code group members. The studies differed, however, in that Gilberstadt and Duker used significance tests to identify descriptors that were related to the code groups. Most likely the exclusion of unrepresentative cases would also reduce within-group variability, increasing standardized effect sizes and the power of significance tests.

Two factors could have been responsible for the much larger mean effect in the Graham et al. (1999) study. One was the restriction of code group membership to profiles with well-defined codes (code scales exceeding all other clinical scales by at least 5 *T* points). Second was the use of 90 patients with all clinical scale *T* scores below 65 rather than a general psychiatric sample as the comparison group.

Two sources of evidence suggest the latter was more important than the former. McNulty, Ben-Porath, and Graham (1998) have recently published some addi-

TABLE 1
Mean Effect Size by Study

Study	*Mean r*
Graham, Ben-Porath, and McNulty (1999)	.314
Gilberstadt and Duker (1965)	.172
Archer, Griffin, and Aiduk (1995)	.096
Lachar (1968)	.085
Kelley and King	.071
Marks, Seeman, and Haller (1974)	.066
Marks and Seeman (1963)	.065
The Missouri System	.054
Williams and Butcher (1989b)	.042
Lewandowski and Graham (1972)	.022

tional data using their sample. For each member of their original sample who produced a valid MMPI, a high-point code was determined based on the two highest clinical scales (excluding Scales *5* and *0*). Ties were decided based on numerical precedence, a common practice in high-point code system studies. For each code group two subgroups were identified. The well-defined group included those profiles where the code scales exceeded other scales by at least 5 *T* points and at least one code scale was greater than 65 *T*. These criteria are very similar to those used by Graham et al. (1999) for well-defined codes. The poorly defined group included the rest of the code group. For four high-point codes, validity coefficients were then generated separately for patients with well-defined and poorly defined codes.

Although well-defined codes tended to be more highly correlated with conceptually related criteria, the differences were on average small. Using data provided by McNulty et al. (1998; their Tables 3 and 4), the mean weighted validity coefficient for the well-defined code groups is estimated to be .31, versus .27 for poorly defined groups, for a subset of their criteria.

On the other hand, Archer, Griffin, and Aiduk (1995) provided data for patients who produced within-normal-limits profiles in their study. This group was associated with more significant outcomes than any other code group examined except one, and those two groups had substantially more significant outcomes than any other group. Furthermore, effects for the within-normal-limits group were consistently in the negative direction. Taken in combination these two sets of findings suggest the use of a within-normal-limits comparison group probably contributed more to the higher mean effect size than the use of well-defined codes.

Potential Moderators

We conducted a series of comparisons to evaluate various methodological factors as potential moderators of the mean effect size by computing weighted means for various combinations of the studies. These analyses must be viewed cautiously because the small number of studies makes isolating the causes for differences in mean effects difficult. For example, whenever results from the Gilberstadt and Duker (1965) or Graham et al. (1999) studies are included, the mean effect is inevitably increased regardless of whether the methodological issue under consideration impacts on effect size. Without these two, however, the number of studies is too small to support some important comparisons. As a compromise, the results are presented both without and with data from the Graham et al. study, which represents a different type of study than the others by virtue of its unique comparison group.

The first entry in Table 2 indicates the overall mean. Even with Graham et al. (1999) included, the mean correlation is only .071; without this study the mean drops to .059. We then compared studies in which the effect sizes were computed

TABLE 2
Aggregations Across Studies

	Without Graham et al. (1999)		*With Graham et al. (1999)*	
Variable	*N*	*Mean r*	*N*	*Mean r*
Overall	9	.059	10	.071
Computed versus estimated				
Computed	3	.043	3	.043
Estimated	6	.062	7	.076
Adolescents versus adults				
Adolescents	2	.059	2	.059
Adults	7	.059	8	.074
Outpatients versus inpatients[a]				
Outpatients	1	.071	2	.175
Inpatients	6	.058	6	.058
Type of data				
Staff or other ratings	5	.056	6	.070
Chart review	5	.061	5	.061
Self-report	1	.086	2	.216
MMPI–2 versus MMPI–1				
MMPI–2	1	.096	2	.224
MMPI–1	8	.058	8	.058
More versus less restrictive coding				
Restrictive	2	.098	3	.174
Nonrestrictive	7	.055	7	.055
Validity data used versus not used				
Used	3	.089	4	.154
Not used	6	.055	6	.055

Note. *N* = the number of studies with the design feature. The sum of the *N*s is greater than 10 for type of data because studies included multiple types of criteria.

[a]The two adolescent studies included mixed inpatient and outpatient samples and were excluded from this analysis.

from the original statistics to studies where they were estimated from significance test outcomes. We found little difference between the means. Even with the Graham et al. study factored in, the means differed by less than .04, or a difference in the proportion of shared variance of less than .004%. The results support the validity of the estimation procedure as an alternative to the computational method.

We conducted the next two comparisons to examine differences across populations. Whether the sample included adolescents or adults seems to have had a trivial impact on the mean correlation. Without the Graham et al. (1999) study, the mean correlations were the same to three decimal places. Despite reasons for believing the MMPI–1 was not appropriate for adolescents (see Butcher et al., 1992), the accuracy of code-based interpretations seemed no worse for adolescents than

for adults. One must remember, however, that both studies with adolescents used the Marks, Seeman, and Haller (1974) normative data for their *T* score transformations. The results cannot be generalized to adolescent profiles scored using the adult norms, as was common practice with the MMPI–1. The results indicate that within this set of studies data can be appropriately aggregated across age groups.

The third analysis compared outpatients to inpatients. Given that the MMPI was developed using inpatients, we were surprised to find slightly stronger mean effects for outpatients than inpatients. We must interpret the results cautiously. The Graham et al. (1999) study has a particularly strong influence on the outcome, and only one other research team (Kelley and King) restricted the sample to outpatients, although the mean correlation for that study was in the upper half of the distribution.

The fourth analysis compared categories of criterion variables. The results are fairly consistent when the data involve evaluation of the patient by others, either through direct ratings of the patient or chart review. However, self-report data was associated with larger effect sizes, particularly in the Graham et al. (1999) study.

The fifth analysis compared the MMPI–1 to the MMPI–2. As with the comparison of outpatients to inpatients, only one study besides Graham et al. (1999) has been published with the MMPI–2. The results indicate a slight increase in the mean effect for the MMPI–2 when compared with the MMPI–1. The two MMPI–2 studies happened to be the two studies that contributed data to the mean effect size for self-report data.

Because self-report data were associated with a larger mean effect than other categories of data, the comparison between the MMPI–1 and MMPI–2 was repeated without the self-report data. The relative outcome was the same, suggesting the difference between versions of the test is not solely attributable to the use of self-report criteria. One possible explanation for this finding suggests that eliminating weak items in the revised edition reduced error variance in code classification.

The last two analyses examined characteristics of the code definition strategy. The first compared studies using more restrictive coding methods (Gilberstadt & Duker, 1965; Graham et al., 1999; Marks & Seeman, 1963) to those with more liberal strategies. The second compared studies that consistently used validity data to exclude cases (Archer et al., 1995; Gilberstadt & Duker, 1965; Graham et al., 1999; Williams & Butcher, 1989a, 1989b) with those that did not. In both cases the more restrictive strategy was associated with a slight increase in the mean, even when Graham et al. was excluded. The Gilberstadt and Duker study also tilted the results in both cases. However, the study with the next highest mean correlation also used validity data. The results support the conclusion that using validity data and, to a lesser extent, a restrictive code definition strategy can produce a small improvement in fit. The latter is also consistent with the results of the McNulty et al. (1998) study. One should remember, however, that using the validity data also ren-

ders about 15% of profiles unclassifiable (McGrath & Ingersoll, this issue) and that prior research suggests restrictive strategies have a similar effect (see Marks et al., 1974, and Wiggins, 1972, for reviews). Using these strategies to improve fit has a cost for inclusiveness.

Comparison of High-Point Codes

The next set of analyses examined the mean correlations for various codes. We conducted the analysis using all 10 studies, then repeated without Graham et al. (1999), and again without Gilberstadt and Duker (1965) to examine the impact of those two studies. A code was only included in this analysis if it appeared in two or more studies. To increase the number of codes meeting this criterion, we combined all permutations of the code even if the original study treated them separately because Gynther, Altman, and Sletten (1973a) found that differences between permutations tend to be small. Even so, only 19 codes met the criterion even when we included all 10 studies.

Table 3 lists the seven codes associated with mean effects of .10 or greater. Not surprisingly, they include some of the better known MMPI codes, such as the 2-7-8 and 1-2-3. A tentative conclusion suggests that clinicians can feel more confident in the interpretation when the high-point code is included in this group. One might even hypothesize that in cases of profiles that meet criteria for more than one code, a preference for the codes included in Table 3 can lead to slightly larger mean effects. However, one should note that this conclusion, even if valid, applies only to the population of general psychiatric patients. For example, research has raised doubts about the generalization of conclusions concerning codes involving Scales

TABLE 3
Codes From Multiple Studies With the Highest Mean Effect Sizes

Code	*Mean r*
1-2-3[a]	.176
2-7-8[a]	.160
2-4-7	.141
2-7	.132
4-9	.130
2-4	.122
6-8[b]	.112

Note. The following codes were also evaluated: 1-2, 1-3, 2-3, 2-8, 3-4, 4-6, 4-7, 4-8, 6-9, 7-8, 8-9, and 2-4-8.

[a]Mean $r \geq .10$ even without the Graham et al. (1999) study. [b]Mean $r \geq .10$ even without the Graham et al. (1999) and Gilberstadt and Duker (1965) studies.

1 and *3* from this population to chronic pain patients (e.g., McGrath, Sweeney, O'Malley, & Carlton, 1998).

DISCUSSION

Given the image of the MMPI as a powerful clinical instrument, the results of these analyses were surprising and disappointing. Correlations of .059 to .071 are generally considered small (Cohen, 1988), accounting for less than 1% of the variance of criterion variables. The first issue that deserves consideration is whether the results could reflect errors in the analysis. The second is how it compares to previous summaries of MMPI effect sizes. The third is why clinicians place so much faith in the MMPI interpretive system based on high-point codes.

Possible Sources of Error

We identified two obvious potential sources of error in the analyses. First, perhaps we did a particularly bad job of identifying predictable relations, either because we were poor judges or because the interpretive narratives we used to guide our judgments did not accurately reflect the actuarial literature. Several findings argue against this explanation.

Williams and Butcher (1989b) identified those criterion variables they expected would be related to each of their codes, based on prior experience with the MMPI and earlier analyses with the same data set. They also indicated the proportion of those relations that proved to be significant. Rather than repeating the task of selecting predictable relations, we used their results for our analyses. For the Williams and Butcher study, our findings represent the outcome for a different set of judges. The estimated mean *r* proved to be .042, which was at the lower end of the distribution of mean effects and consistent with general findings.

McNulty et al. (1998) also identified expected correlates for each of four codes. They initially hypothesized that 19 scales from their Patient Description Form (PDF) would be related to one or more of the codes. In fact, none of the correlations were significant for 9 of those 19 scales, suggesting the mean power for their analyses of expected relations was probably less than .60. Such an estimate is also consistent with (in fact, lower than) the mean power estimate we derived for our analysis of their PDF data.

In both cases, our results were consistent with those from other judges who are recognized experts on the use of the MMPI. It is unlikely the present results can be wholly or even largely blamed on unreasonable errors in the relations selected for analysis.

A second possible source of error was the identification of significant outcomes from narratives provided in some of the source studies. This is also unlikely to be a major contributor to the outcome. The review of narratives was only necessary for a few studies where the mean correlation was generated using the estimation procedure. The mean effect size was actually larger for the studies in which we estimated effect sizes than for studies where they were directly computed. If any bias occurred in the interpretation of the narratives, it was most likely a positive one.

Comparison With Previous Reviews

Comparing our findings with those from previous reviews of MMPI effect sizes raises some interesting speculations. The meta-analyses by Atkinson (1986) and Parker, Hanson, and Hunsley (1988) are the best known of these reviews. Atkinson examined MMPI validation studies collected from every 5th year of *Psychological Abstracts* between 1960 and 1980. Parker et al. compared "unknown validity" (exploratory) and "convergent validity" (confirmatory) studies published in the *Journal of Clinical Psychology* and the *Journal of Personality Assessment* between 1970 and 1981. Garb, Florio, and Grove (1998) have recently reanalyzed the Parker et al. data.

In both cases the results were more consistent with general beliefs concerning the MMPI. Atkinson (1986) reported a mean correlation between MMPI scales and clinical correlates of .30 to .40 depending on the statistic used. Garb et al.'s (1998) reanalysis of the Parker et al. data indicates the mean correlation between MMPI scales and extra test correlates was .48 in confirmatory analyses and .11 in exploratory analyses.

On the other hand, Hedlund (1977) reviewed studies that examined correlates of individual MMPI scales. Except for the use of individual scales as predictors rather than dummy variables based on high-point codes, the studies he reviewed were very similar to the high-point code system studies in methodology. Hedlund examined relations between MMPI signs and a wide variety of free criteria on an exploratory basis. His findings were consistent with ours. One of his conclusions could be applied just as well to the current findings: "virtually all of the observed relationships … were of very low order, often accounting for less than 1% of the covariance and seldom accounting for as much as 10%" (pp. 749–750).

Several factors could contribute to these differences. The impact of methods on effect sizes has been addressed several times in this review (also see McGrath & Ingersoll, this issue). The sampling procedure used (cross-sectional vs. retrospective), the type of predictor variable (dichotomized vs. quantitative), and the degree of dissimilarity between the comparison and target groups all play a role. The high-point code system researchers as a group chose more clinically useful but less powerful options, and the studies reviewed by Hedlund (1977) were also cross-

sectional. The two previous meta-analyses combined results regardless of method, which should have produced larger mean effects.

Second, we hypothesize that the bias against the publication of negative findings (see Greenwald, 1975; Rosenthal, 1979) would result in greater overestimation of true effects in the Atkinson (1986) and Parker et al. (1988) reviews, which were limited to published studies. Hedlund (1977) included several unpublished data sets. Our review is also restricted to published data. However, it seems likely that studies with very large criterion sets should result in enough significant outcomes to satisfy editors' biases. We have already noted that our analyses included 2,201 relations from the 10 studies even after eliminating unexpected and potentially negative relations. In contrast, Atkinson found only 115 significance test results in 87 MMPI studies, and Parker et al. indicated an average of about 27 analyses in the 411 studies they reviewed.

The difference in the number of criteria could have impacted the mean correlations in another way. The smaller the set of criteria, the more likely the researcher will be able to ensure their reliability and validity. The validity of the criteria was never directly addressed in any of the high-point code system studies, but some information is available on the interrater reliability of the clinician ratings. Although this represents the largest single class of criterion variables used in the high-point code system studies, Marks and Seeman's (1963) study was the only one in which the reliability of these ratings was examined.[4] They found the mean reliability between raters was only about .50. A recent high-point code study by McGrath et al. (1998) reported similar results. In both cases these ratings were specifically collected for the study; we suspect clinicians may be even less reliable when the data are based on routine activities such as writing chart notes or checking off problem areas at intake. A lone intake worker's rating of hostility in the course of a standard initial interview is very questionable as the basis for judgments about the patient's level of hostility. However, worth noting is that even when the criteria were well-validated self-report measures such as the scales from the Symptom Checklist–90 the correlations were still not substantial, especially when one considers the probable impact of shared method variance on these relations.

Sources of Discrepancy From Clinician Perceptions

If the results of our analyses are accurate concerning mean effects in the high-point code system studies, another issue worth discussing is why clinicians who frequently use the MMPI put so much faith in the interpretations based on the high-point code. We have identified at least four factors that could play a role. The first

[4]It is important to remember that this is a different issue than clinician agreement or the reliability of chart reviews.

three represent various ways in which clinicians could have come to overestimate the validity of MMPI codes; the last suggests the research does not adequately reflect the clinical use of high-point codes.

A large literature is available that discusses factors that interfere with accurate clinical judgment (e.g., Arkes, 1981). Covariation misestimation refers to the tendency to weight true positives more heavily than negatives or false positives when judgments are made about the accuracy of clinical decisions. This factor could be particularly active in the formation of opinions about high-point codes. As the most commonly administered psychological instrument, many clinicians have frequent contact with the MMPI. This creates the opportunity for frequent exposure to true positives based on the high-point code even if the error rate is high, which could lead to subjective overestimation of the code's validity.

Second, clinicians may tend to overestimate the power of the codes because of the common misunderstanding of significance. Significance is a function of several factors, including alpha level, sample size, and effect size. Several authors have demonstrated that psychologists assume an effect is larger when it is associated with a significant outcome than when it is associated with a nonsignificant outcome, regardless of sample size (e.g., Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986). One of the hallmarks of the high-point code system studies is a large sample size (see McGrath & Ingersoll, this issue), which means relations can be significant even when effect sizes are small. When multiple researchers find that codes involving Scales 1 and 3 are significantly related to excessive somatic focusing, it can create the illusion of a powerful effect for readers who ignore the effect of sample size on power.

A third factor that could contribute to the misperception of code validity is the availability of MMPI clinical guides implying a high degree of confidence in code-based descriptions. The clinical texts on the MMPI only partially meet the goals of the empirical cookbook originally described by Meehl (1956). The seamless combination of clinical lore with research-based statements, definitive statements about the interpretation of codes, and disregard of inconsistencies in the research outcomes, can combine to create an impression of the codes' validity not supported by this research.

Finally, the existing high-point code system studies might not provide an adequate test of code-based interpretation. Clinicians tend to interpret many of the codes as indicators of certain character types. It may not be surprising to find that the effects are generally small when criteria are chosen based on their availability rather than their appropriateness as indicators of the character types clinicians believe the codes represent.

This is a different validity issue than that raised in the last section. Cone's (1995) distinction between representational and elaborative validity helps to clarify the difference. Representational validity has to do with whether the measure as-

sesses what it is supposed to assess. This was the issue that was raised earlier in questioning the reliability and validity of an intake worker's ratings of hostility.

Elaborative validity has to do with the extent to which a measure bears on the understanding of related phenomena. For example, the character type typically associated with the 3-4 code is a person who is conflicted about the expression of hostility. How strongly a rating of "hostility" from the initial interview will be related to this code, regardless of how valid that rating is, is questionable.

CONCLUSIONS

Despite clinical impressions of the MMPI and previous reviews suggesting it is a powerful predictor of clinical phenomena, our findings indicate the high-point code system studies do not suggest very high levels of validity for MMPI codes. The result is particularly surprising in that these studies provide the basis for the general interpretive strategy recommended by most MMPI texts. Whether and to what extent differences in methods across studies, overestimation of the validity of high-point codes by clinicians, and the use of criteria of low validity contribute to these discrepancies is unclear. We suspect the latter factor is particularly important. Plainly put, it is one thing to find that a new patient with a 2-4 code is not rated as "hostile" by an intake worker after a 1-hr interview. It is quite another if after working with the patient for a week several clinicians fail to concur that the patient meets the description of an "angry depressive," who may or may not appear overtly hostile, but who resists treatment and fosters countertransferential reactions (Greene, 1991).

This discussion suggests that the reliance on criteria of convenience has potentially compromised both the power of these studies and their congruence with the clinical use of high-point codes. The ecological validity of these studies could be improved in several ways, such as by comparing the accuracy of automated reports and random or generic reports (Moreland & Onstad, 1985), selecting criteria on the basis of expected relations with codes, and using statistical techniques that estimate a common latent variable for the criterion variables such as discriminant function or factor analysis (McGrath et al., 1998). Until confirmatory studies are available that evaluate current beliefs about the interpretation of high-point codes, the validity of the MMPI high-point codes has not been adequately tested.

ACKNOWLEDGMENTS

The completion of this study would not have been possible without the help of many people. In particular, we thank John Graham, Yossef Ben-Porath, John McNulty, and Robert Archer.

REFERENCES

*These studies provided data incorporated into the synthesis of effect size estimates.

*Altman, H., Gynther, M. D., Warbin, R. W., & Sletten, I. W. (1972). A new empirical automated MMPI interpretive program: The 6-8/8-6 code type. *Journal of Clinical Psychology, 28,* 495–498.

*Altman, H., Warbin, R. W., Sletten, I. W., & Gynther, M. D. (1973). Replicated correlates of the MMPI 8-9/9-8 code type. *Journal of Personality Assessment, 37,* 369–371.

*Archer, R. P., Griffin, R., & Aiduk, R. (1995). MMPI–2 clinical correlates for ten common codes. *Journal of Personality Assessment, 65,* 391–407.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49,* 323–330.

Atkinson, L. (1986). The comparative validities of the Rorschach and MMPI: A meta-analysis. *Canadian Psychology, 27,* 238–247.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cone, J. D. (1995). Assessment practice standards. In S. C. Hayes, V. M. Follette, R. M. Dawes, & K. Grady (Eds.), *Scientific standards for psychological practice: Issues and recommendations* (pp. 201–224). Reno, NV: Context.

Erdfelder, E., Faul, F., & Buchner, A. (1996). G*POWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers, 28,* 1–11.

Garb, H. N., Florio, C. M., & Grove, W. M. (1998). The validity of the Rorschach and the Minnesota Multiphasic Personality Inventory: Results from meta-analyses. *Psychological Science, 9,* 402–404.

*Gilberstadt, H., & Duker, J. (1965). *A handbook for clinical and actuarial MMPI interpretation.* Philadelphia: Saunders.

Graham, J. R. (1987). *The MMPI: A practical guide* (2nd ed.). New York: Oxford University Press.

*Graham, J. R., Ben-Porath, Y. S., & McNulty, J. L. (1999). *MMPI–2 correlates for outpatient community mental health settings.* Minneapolis: University of Minnesota Press.

Greene, R. L. (1991). *The MMPI–2/MMPI: An interpretive manual.* Boston: Allyn & Bacon.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Gynther, M. D., Altman, H., & Sletten, I. W. (1973a). Development of an empirical interpretive system for the MMPI: Some after-the-fact observations. *Journal of Clinical Psychology, 29,* 232–234.

*Gynther, M. D., Altman, H., & Sletten, I. W. (1973b). Replicated correlates of MMPI two-point code types: The Missouri actuarial system. *Journal of Clinical Psychology, 29,* 263–289.

*Gynther, M. D., Altman, H., & Warbin, R. W. (1972a). A new actuarial–empirical automated MMPI interpretive program: The 4-3/3-4 code type. *Journal of Clinical Psychology, 29,* 229–231.

*Gynther, M. D., Altman, H., & Warbin, R. W. (1972b). A new empirical automated MMPI interpretive program: The 2-4/4-2 code type. *Journal of Clinical Psychology, 28,* 498–501.

*Gynther, M. D., Altman, H., & Warbin, R. W. (1973a). Behavioral correlates for the Minnesota Multiphasic Personality Inventory 4-9, 9-4 code types: A case of the emperor's new clothes? *Journal of Consulting and Clinical Psychology, 40,* 259–263.

*Gynther, M. D., Altman, H., & Warbin, R. W. (1973b). Interpretation of uninterpretable MMPI profiles. *Journal of Consulting and Clinical Psychology, 40,* 78–83.

*Gynther, M. D., Altman, H., & Warbin, R. W. (1973c). A new empirical automated MMPI interpretive program: The 2-7/7-2 code type. *Journal of Clinical Psychology, 29,* 58–59.

*Gynther, M. D., Altman, H., & Warbin, R. W. (1973d). A new empirical automated MMPI interpretive program: The 6-9/9-6 code type. *Journal of Clinical Psychology, 29,* 60–61.

*Gynther, M. D., Altman, H., Warbin, R. W., & Sletten, I. W. (1973). A new empirical automated MMPI interpretive program: The 1-2/2-1 code type. *Journal of Clinical Psychology, 29,* 54–57.

Hathaway, S. R., & McKinley, J. C. (1983). *Manual for the administration and scoring of the MMPI.* Minneapolis, MN: National Computer Systems.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88,* 359–369.

Hedlund, J. L. (1977). MMPI clinical scale correlates. *Journal of Consulting and Clinical Psychology, 45,* 739–750.

*Kelley, C. K., & King, G. D. (1978). Behavioral correlates for within-normal-limit MMPI profiles with and without elevated *K* in students at a university mental health center. *Journal of Clinical Psychology, 34,* 695–699.

*Kelley, C. K., & King, G. D. (1979a). Behavioral correlates of infrequent two-point MMPI code types at a university mental health center. *Journal of Clinical Psychology, 35,* 576–585.

*Kelley, C. K., & King, G. D. (1979b). Behavioral correlates of the 2-7-8 MMPI profile type in students at a university mental health center. *Journal of Consulting and Clinical Psychology, 47,* 679–685.

*Kelley, C. K., & King, G. D. (1979c). Cross validation of the 2-8/8-2 MMPI code type for young adult psychiatric outpatients. *Journal of Personality Assessment, 43,* 143–149.

*Kelley, C. K., & King, G. D. (1980). Two- and three-point classification of MMPI profiles in which Scales 2, 7, and 8 are the highest elevations. *Journal of Personality Assessment, 44,* 25–33.

*King, G. D., & Kelley, C. K. (1977a). Behavioral correlates for spike-4, spike-9, and 4-9/9-4 MMPI profiles in students at a university mental health center. *Journal of Clinical Psychology, 33,* 718–724.

*King, G. D., & Kelley, C. K. (1977b). MMPI behavioral correlates of spike-5 and two-point code types with scale 5 as one elevation. *Journal of Clinical Psychology, 33,* 180–185.

*Lachar, D. (1968). MMPI two-point code-type correlates in a state hospital population. *Journal of Clinical Psychology, 24,* 424–427.

*Lewandowski, D., & Graham, J. R. (1972). Empirical correlates of frequently occurring two-point MMPI code types: A replicated study. *Journal of Consulting and Clinical Psychology, 39,* 467–472.

*Marks, P. A., & Seeman, W. (1963). *The actuarial description of abnormal personality: An atlas for use with the MMPI.* Baltimore: Williams & Wilkins.

*Marks, P. A., Seeman, W., & Haller, D. L. (1974). *The actuarial use of the MMPI with adolescents and adults.* Baltimore: Williams & Wilkins.

McGrath, R. E., Sweeney, M., O'Malley, W. B., & Carlton, T. K. (1998). Identifying psychological contributions to chronic pain complaints with the MMPI–2: The role of the *K* scale. *Journal of Personality Assessment, 70,* 448–459.

McNulty, J. L., Ben-Porath, Y. S., & Graham, J. R. (1998). An empirical examination of the correlates of well-defined and not defined MMPI–2 code types. *Journal of Personality Assessment, 71,* 393–410.

Meehl, P. E. (1956). Wanted—A good cookbook. *American Psychologist, 11,* 262–272.

Moreland, K. L., & Onstad, J. A. (1985, March). *Validity of the Minnesota Clinical Report. I: Mental health outpatients.* Presented at the 20th Annual Symposium on Recent Developments in the Use of the MMPI, Honolulu, HI.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299–1301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester, England: Wiley.

Parker, K. C. H., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103,* 367–373.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

*Warbin, R. W., Altman, H., Gynther, M. D., & Sletten, I. W. (1972). A new empirical automated MMPI interpretive program: 2-8 and 8-2 code types. *Journal of Personality Assessment, 36,* 581–584.

Wiggins, J. S. (1972). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Williams, C. L., & Butcher, J. N. (1989a). An MMPI study of adolescents: I. Empirical validity of the standard scales. *Psychological Assessment, 1,* 251–259.

*Williams, C. L., & Butcher, J. N. (1989b). An MMPI study of adolescents: II. Verification and limitations of code type classifications. *Psychological Assessment, 1,* 260–265.

APPENDIX

General Issues

For each study we used one of two procedures to estimate effect sizes. We used the computational procedure in studies where the full set of original statistics was available. We computed an effect size (*d* for continuous criteria, ϕ for dichotomous criteria) for each relation we concluded could be justified on an a priori basis. Using a formula provided by Cohen (1988) for cases involving populations of unequal size, we then converted *d* to a point-biserial correlation.

We used the estimation procedure (Hedges & Olkin, 1980) when original statistics were incomplete. First, within a set of relations we had concluded could be expected on a priori grounds, we computed the proportion of significant outcomes. This served as an estimate of the mean power for the analyses. Based on this estimate, and information provided in the original studies about alpha level and group sizes, we generated a mean effect size (*d* for continuous data, *r* for categorical data or data analyzed via correlations) by interpolation or extrapolation from power tables in Cohen (1988).

We then reversed the process. We used a program that provides exact power estimates called G*POWER (Erdfelder, Faul, & Buchner, 1996) to generate an estimate of the mean power of those analyses based on the effect size estimate taken from Cohen (1988). We modified the effect size estimate and reentered it in an iterative fashion until the power value generated by G*POWER agreed to at least three decimal places with the initial power estimate. Again, if the effect size measure was *d*, we converted this to *r* using the formula for unequal *n*s Cohen provided.

We based the mean effect size for each study on correlation estimates weighted by the sample size and the number of relations represented by the estimate (because some were estimates of the mean correlation for a set of analyses). We also

used this procedure to combine data across studies to examine potential moderator variables.

Marks and Seeman

Marks and Seeman (1963) did not use significance tests in their study, so we could not use the estimation procedure. The first author indicated the original data and statistics no longer existed, making the statistics found in Appendixes A and B of the text the only available information. Because only means were provided for Q-sort variables, we could not compute standardized effect size measures. Proportions were available for a large number of categorical chart review variables, however, and these represented the data source for computations. Samples sizes varied for different classes of variables. Given the results we report for the Gilberstadt and Duker study (1965), the use of the most representative cases for the Q-sort analyses means their inclusion in our analysis probably would have increased the mean effect size.

Gilberstadt and Duker

We estimated mean effect sizes because a large number of the original statistics were missing from the table of results in the text. Because comparisons were made to a fixed comparison group and the number of cases in code groups differed, the sample size varied for each code; therefore, we generated mean estimates separately by code.

Lachar

Complete data were available from the article to compute the proportion of cases in each code group that fell in each diagnostic group. We computed ϕ for each instance in which a relation was hypothesized between the code and diagnosis, using cases from all other code groups as the basis for comparison.

Lewandowski and Graham

We computed effect sizes. Means, standard deviations, and group sizes were available from the second author for each of their two samples, allowing direct computation of *d*.

The Missouri System

Because of the replication procedure employed in these studies, we used a more complicated variant of the estimation procedure for the Missouri system. We generated an initial power estimate based on the number of replicated outcomes for Samples 1 and 2 combined and for Samples 3 and 4 combined. This value can be conceived as the product of the powers for the two samples as follows:

$$(1 - \beta \mid N_1, \alpha = .10, \rho)(1 - \beta \mid N_2, \alpha = .10, \rho) = 1 - \beta \qquad \text{(A1)}$$

for replicated analyses.

The goal of the analysis was to determine a value for ρ that produced power levels for the two samples that, when multiplied, produced the mean power estimate for the replicated analyses.

This was simpler to compute for Samples 3 and 4 because the similar sample sizes meant the mean power for each sample was approximately the square root of the mean power for the replicated analyses. We generated an initial estimate of ρ from Cohen (1988) using the combined sample sizes, the power for the hypothesized relations, and $\alpha = .01$. We then employed the iterative process using G*POWER to identify the value for ρ that produced the square root of the initial power estimate when $\alpha = .10$. We used the same procedure for Samples 1 and 2, except that differences in sample sizes meant we had to conduct the analysis separately for the samples a number of times until the product of the two sample mean power levels equaled the mean power for the replicated analyses.

Marks, Seeman, and Haller

We estimated mean effect sizes. Marks et al. (1974) used two standards for significance. Replicated analyses were required to demonstrate a joint alpha level of .06, and a .04 level was used when the results were not replicated. Information was not available indicating which analyses met each criterion, so the power of the analyses was approximated using $\alpha = .05$ throughout. This study included the largest set of criterion variables of any study examined. To control the number of relations to be examined, we included only therapist data in our analyses.

Kelley and King

We estimated mean effect sizes. Although one study used an alpha level of .01, the rest of the series used .05 as the criterion for significance. To simplify the analysis we assumed .05 as the alpha level for all analyses, resulting in a small underestimation of the effect size in several analyses that were tested at .01.

Williams and Butcher

We estimated mean effect sizes. Mean effect sizes were generated for different classes of criterion variables because sample sizes varied. For this study the original authors identified expected relations and indicated the proportion that proved significant. We used their judgments for this analysis as well.

Archer, Griffin, and Aiduk

We estimated mean effect sizes. We omitted the within-normal-limits code because we expected most relations to be negative. The authors provided copies of all dependent measures except their demographic information form and the Ohio Literacy Test. One of their measures, the modified Nursing Behavior Index, was completed by a member of the nursing staff on a 3-point scale from *not observed* to *observed frequently or often.* Relations between these items and code group were evaluated in the original study with two degree of freedom chi-square tests. Cramer's ϕ' was estimated to be .14 on average. However, ϕ' is not a correlation coefficient and cannot be directly compared to the correlations resulting from the other analyses. Therefore, we did not include the results of these analyses in the synthesis.

Graham, Ben-Porath, and McNulty

We estimated mean effect sizes. The authors provided copies of all dependent measures, as well as information about significant analyses. Thirteen of the 17 codes were significantly related to 10 or more items from the 188-item PDF, but the authors only listed up to 10 significant correlates from this measure for each code. Because any other significant items were treated as if they were not significant, the result is an underestimation of the true mean effect. On the other hand, given the information provided, we were not able to determine variables generated from the intake data unless there was at least one significant outcome. The omission of items from the intake that never proved to be significant would tend to overestimate the mean effect.

Robert E. McGrath
School of Psychology
Fairleigh Dickinson University
Teaneck, NJ 07666
E-mail: mcgrath@alpha.fdu.edu

Received April 28, 1999