Chapter 1

The Evolution of the MMPI, MMPI-2, and MMPI-2-RF

The Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1940) and its successors, the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and the MMPI-2 Restructured Form (MMPI-2-RF: Ben-Porath & Tellegen, 2008), are currently the most widely used and researched self-report measures of psychopathology. Dahlstrom, Welsh, and Dahlstrom (1975) included almost 6,000 references on the clinical and research applications of the MMPI. Lubin, Larsen, Matarazzo, and Seever (1985) reported that the MMPI is the most frequently used test in professional settings and Watkins, Campbell, Nieberding, and Hallmark (1995) reported similar findings for the MMPI-2. Butcher and Rouse (1996) found over 4,300 references to the MMPI over the 20 years from 1974 to 1994. An electronic search of the psychology databases in January 2010 using the search term "MMPI" identified 24,171 citations and the search term "MMPI-2" identified 4, 216 citations. There is a prolific research and clinical literature on the MMPI and MMPI-2 that reflects its widespread use over 70 years.

A person taking the original MMPI responded "true" or "false" to 566 statements. The person's responses to these statements were then scored on 10 clinical scales that assess major categories of psychopathology. In addition, 4 validity scales assessed the person's test-taking attitudes. Table 1.1 illustrates the scale names and numbers of the 4 validity and

TABLE 1.1 MMPI Validity and Clinical Scales

SCALE NAME	NUMBER	ABBREVIATION	NUMBER OF ITEMS
Validity			
Cannot Say		?	
Lie		L	15
Infrequency		F	64
Correction		K	30
Clinical			
Hypochondriasis	1	Hs	33
Depression	2	D	60
Hysteria	3	Hy	60
Psychopathic Deviate	4	Pd	50
Masculinity-Femininity	5	Mf	60
Paranoia	6	Pa	40
Psychasthenia	7	Pt	48
Schizophrenia	8	Sc	78
Hypomania	9	Ma	46
Social Introversion	0	Si	70

10 clinical scales on the original MMPI. A standard MMPI profile sheet (Profile 1.1) was used for plotting the person's scores on these validity and clinical scales.

After a brief review of the history of self-report personality inventories, the rationale underlying the development of the MMPI and the empirical methods used for item selection and scale construction will be described. Next, the appropriateness of the original norms for the MMPI for contemporary use (cf. Pancoast & Archer, 1989; Colligan, Osborne, Swenson, & Offord, 1983, 1989) will be discussed, followed by describing the development of the MMPI-2 (Butcher et al., 1989), and then the MMPI-2 Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008). The Chapter will end with a quick overview of the interpretive process for the MMPI-2 and MMPI-2 RF that provides the rationale for the organization of the Chapters within this text. MMPI-2 will be used throughout this text as a single term to refer to the original MMPI, the MMPI-2, and the MMPI-2-RF, other than when dictated by the context.

THE EARLY HISTORY OF SELF-REPORT PERSONALITY INVENTORIES

Woodworth Personal Data Sheet

Personality assessment, like intellectual assessment, received its first major impetus during World War I when a need arose for assessment procedures to screen large numbers of individuals. In response to this demand, Woodworth and Poffenberger developed the Woodworth Personal Data Sheet (Woodworth, 1920), a self-rating scale for detecting mental instability. They assembled 116 questions to which the person answered "yes" or "no," such as "Do you usually feel well and strong?" or "Does it make you uneasy to sit in a small room with the door shut?" Woodworth and Poffenberger found that normal individuals would provide about 10 pathological answers to these questions. Any individual who provided a pathological answer to 20 or more questions was to be interviewed by a psychiatrist. Some questions were considered so pathognomonic that a "yes" response to any of them prompted an individual interview such as "Do you know of anybody who is trying to do harm to you?" or "Do you feel like jumping off when you are on a high place?" The items were heterogeneous in content because they tapped every symptom of mental instability that Woodworth and Poffenberger could identify. Neither a systematic empirical method nor a theoretical perspective was employed in selecting questions to be included on the test. The questions were chosen because Woodworth and Poffenberger thought that they assessed mental instability. They did eliminate some questions that did not differentiate between normal individuals and a small group of individuals with mental instability such as schizophrenia, epilepsy, psychopathic personality, and so on. Although the Personal Data Sheet was developed too late to be very useful in selecting recruits, because the United States was already involved in World War I, it did identify those recruits who needed to be interviewed to determine whether they had sufficient mental stability for service in the army under wartime conditions.

Bernreuter Personality Inventory

The success of psychological testing during World War I stimulated the development in the next decade of several personality inventories similar to the Personal Data Sheet. Probably the best known of these instruments were the Bernreuter Personality Inventory (Bernreuter, 1933), which measured neuroticism, dominance, introversion, and self-sufficiency, and the Humm-Wadsworth Temperament Scale (1935), which measured components of temperament that were important for personnel selection in industries. These inventories started the trend in the field of self-report inventories to assess multiple dimensions or components rather than a single dimension such as mental stability or adjustment. Like other personality inventories of this era, the Bernreuter Personality Inventory was constructed on a logical rather than an empirical basis. That is, the test developer would include questions on a particular scale that, on the basis of clinical experience, were thought to measure a specific trait or construct. Likewise, the test developer would determine the scoring direction for any particular question on a logical basis. For example, if the test developer felt that a "yes" response to the question "Do you daydream a lot?" indicated neuroticism, that question would be added to the neuroticism scale with "yes" as the "deviant" response. The total number of these "deviant" responses—responses that the test developer felt tapped the specific trait or construct being assessed—became the score on the scale.

Strong critiques (cf. Landis & Katz, 1934; Super, 1942) devastated the Bernreuter Personality Inventory. For example, to investigate how certain groups would perform on the Bernreuter Personality Inventory, Landis and Katz (1934) examined the scores of 224 patients with a known clinical diagnosis. On the neuroticism scale, 39 percent of the neurotic patients scored above the 90th percentile; however, 23 percent of the schizophrenic patients and 21 percent of the manic-depressive patients also scored above the 90th percentile. Thus, this scale is inadequate because in addition to identifying some neurotic patients correctly, it also misclassified several groups of psychotic patients as neurotic.

Analyzing responses to individual questions revealed additional problems. Bernreuter determined a

MMPI™

MINNESOTA MULTIPHASIC™ PERSONALITY INVENTORY
S.R. Hathaway and J.C. McKinley
PROFILE

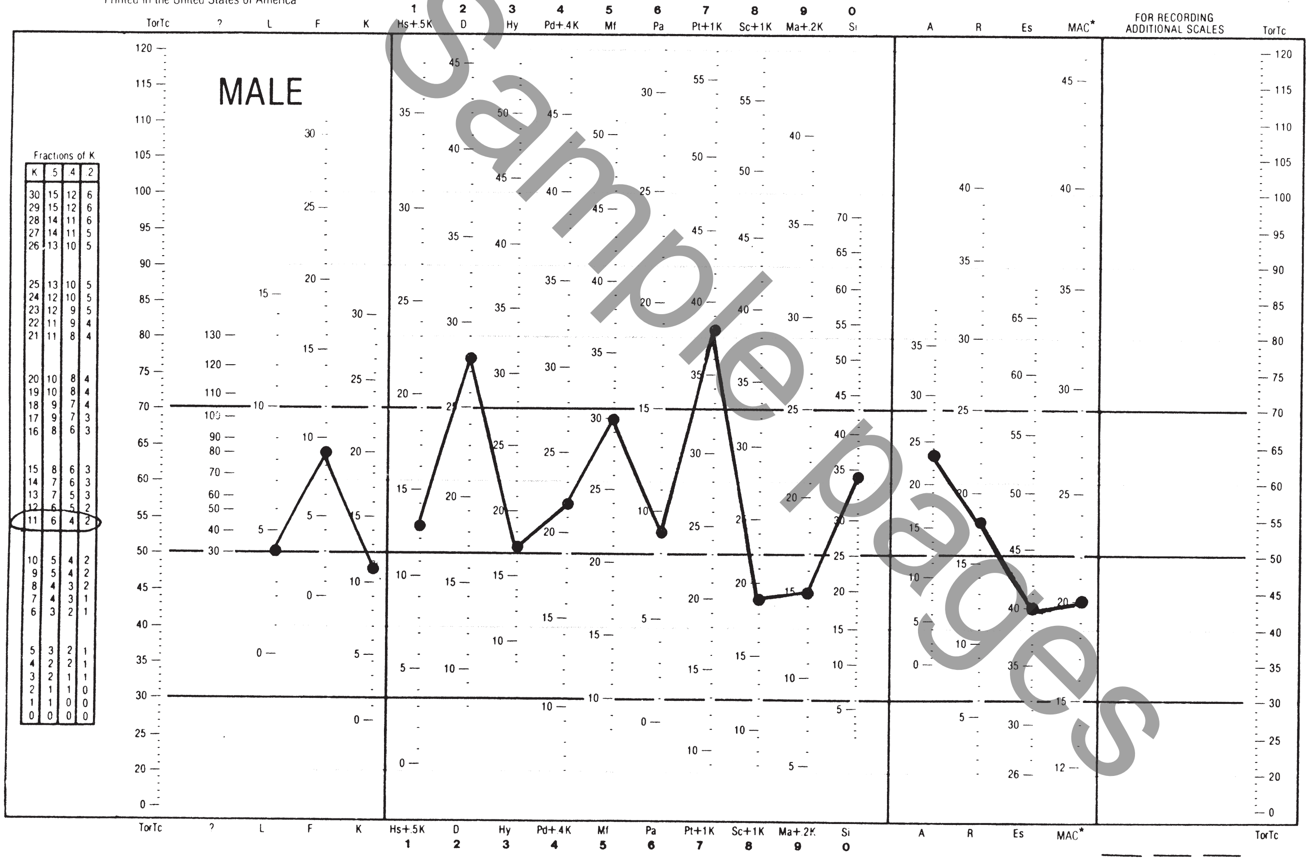
Name **Tim Smith**

Address **547 Geneva Avenue**

Occupation **Clerk** Date Tested **5/ 6/89**

Education **12th** Age **47**

Marital Status **Married** Referred by **Dr. Clark**



Fractions of K

K	.5	.4	.2
30	15	12	6
29	15	12	6
28	14	11	6
27	14	11	5
26	13	10	5
25	13	10	5
24	12	10	5
23	12	9	5
22	11	9	4
21	11	8	4
20	10	8	4
19	10	8	4
18	9	7	4
17	9	7	3
16	8	6	3
15	8	6	3
14	7	6	3
13	7	5	3
12	6	5	2
11	6	4	2
10	5	4	2
9	5	4	2
8	4	3	2
7	4	3	1
6	3	2	1
5	3	2	1
4	2	2	1
3	2	1	1
2	1	1	0
1	1	0	0
0	0	0	0

MMPI Code

Scorer's Initials **HG**

	?	L	F	K	Hs+5K 1	D 2	Hy 3	Pd+4K 4	Mf 5	Pa 6	Pt+1K 7	Sc+1K 8	Ma+2K 9	Si 0	A	R	Es	MAC*
Raw Score	0	4	9	11	7	28	17	18	30	9	27	8	13	34	23	18	40	20
K to be added					6			4			11	11	2					
Raw Score with K					13			22			38	19	15					

*49 item version

NATIONAL COMPUTER SYSTEMS

27209

PROFILE 1.1 MMPI Handscoring Profile Form. Reproduced by permission of the University of Minnesota Press. All rights reserved. "Minnesota Multiphasic Personality Inventory®" and "MMPI®" are trademarks owned by the University of Minnesota.

"deviant" response to the questions on a logical basis without checking to see whether it was accurate for neurotic patients. Landis and Katz (1934) found that other groups endorsed some items as much or more frequently than neurotics. For example, the question "Are you critical of others?" elicited a "yes" response from 69 percent of the normal sample as compared with 32 percent of the neurotic sample and 39 percent of the psychotic sample. Similarly, the question "Do you daydream frequently?" was answered "yes" by 43 percent of the normal sample, 40 percent of the neurotic sample, and 31 percent of the psychotic sample. Clearly, Bernreuter's logical impressions about how the questions on his Inventory would be answered needed to be validated before it was put into widespread use. The omission of this vital step in the test construction process led to the inevitable demise of the Bernreuter Personality Inventory.

Humm-Wadsworth Temperament Scale

Humm and Wadsworth (1935) developed their Temperament Scale, based on Rosanoff's theory of personality (1927), to identify those components of temperament that assessed the ability to adjust to the social requirements of the workplace. Rosanoff's theory of personality conceptualized abnormal behavior as the expression of the uncontrolled manifestations of the same components of temperament seen in normal individuals. That is, there are quantitative rather than qualitative differences in temperament between individuals with normal and abnormal behavior in Rosanoff's theory.

Seven components of temperament were identified by Humm and Wadsworth (1935):

> Normal (N): primarily a control mechanism providing a rational balance and temperamental equilibrium
>
> Antisocial (H): concerned essentially with self-preservation
>
> Cycloid (bipolar) manic (M) or Cycloid (bipolar) depressed (D): characterized by fluctuations in emotionality and activity
>
> Schizoid autistic (A) or Schizoid paranoid (P): characterized by heightened imagination Epileptoid (E): characterized by inspirations to achievement that are meticulously developed and pushed through to completion

Next, Humm and Wadsworth proceeded to write a large number of questions that covered the constituent traits of the seven components. They quickly learned that their logical impressions about how individuals would answer their questions were not supported. When they examined the frequency with which various groups of individuals who had or who did not have the component responded to each question, they found that about one question in four survived. That is, their logical impressions on how the questions would be answered were wrong three out of four times. This finding illustrates how important it is to actually validate how groups of individuals respond to the questions rather than rely on logical impressions.

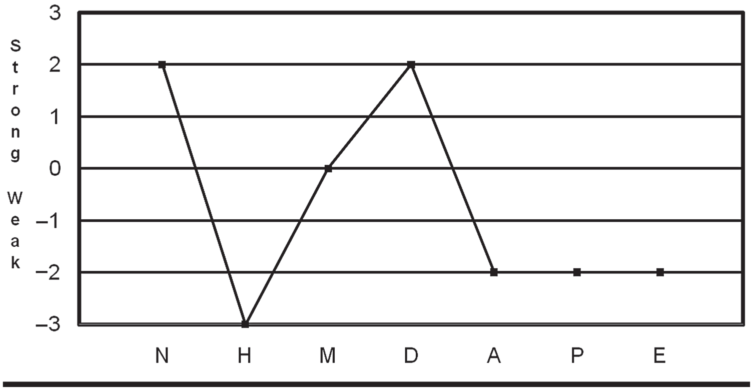
Humm and Wadsworth assigned each question a value of 1 if it was in the bottom fifth of the distribution for that component of temperament, a value of 2 if it was in the next fifth of the distribution, and so on, to a value of 5 if it was in the top fifth of the distribution. The total score for the individual on each component of the scale was the sum of these values for all the questions. Although there were 318 questions on the Humm-Wadsworth Temperament Scale, values were assigned to only 159 (50%) of the questions. Humm and Wadsworth were reluctant to remove the questions that had no values and were not used to determine the total scores on each component because they did not want a scale in which all questions were scored.

The distribution of these total scores for each component was divided into three categories: strong, borderline, and weak. The strong and weak categories were subdivided further into three additional categories:

1. +/−3: one-half of a standard deviation more extreme than the average category
2. +/−2: an average strong/weak score
3. +/1: one-half of a standard deviation less extreme than the average category

Humm and Wadsworth also were concerned with determining the validity of the person's responses to the questions that reflected the first attempt to assess the validity of the specific administration of a self-report inventory. Prior to this time, the test developers simply assumed that the person would provide valid responses to the inventory. This assumption might have been tenable under wartime conditions but it quickly became apparent that validity scales were needed for self-report inventories. Humm and Wadsworth used two criteria to determine the validity of this specific administration of the Scale: the number of "No" responses and the number of skipped questions. The total number of "No" responses to the 318 questions was classified into three categories: (1) acceptable: 145–193; (2) doubtful: 132–144 or 194–214; and (3) unacceptable: < 132 or >214. If more than 25 questions were skipped by the person, the Scale should not be scored.

The diagnosis of temperament was based on the person's highest component. If the person's highest component was M, then the assumption was made that the person has predominantly a cycloid (bipolar) manic temperament. Humm and Wadsworth noted that it was very unusual for an individual to have only one component emphasized in his or her temperament. Most individuals had one component emphasized along with other components secondary in strength. This finding of simultaneous elevation of more



PROFILE 1.2 Humm-Wadsworth Temperament Scale Profile

than one component of temperament is similar to what will be described later with the MMPI in which individuals usually elevated two or more clinical scales.

Profile 1.2 illustrates these seven components of temperament for an individual. This person has generally a "normal" (N) temperament with an associated depressive (D) temperament. The person clearly has very weak antisocial (H), autistic (A), paranoid (P), and epileptoid (E) temperaments. The person's manic (M) temperament is in the borderline range, so it might be worthwhile to review the person's history for a bipolar temperament.

The ensuing literature on the Humm-Wadsworth Temperament Scale through the 1940s primarily was supportive of its use. Humm and Wadsworth published eight studies describing refinements in the Scale and additional uses of it. Other authors published nine studies extending the use of the Scale to other settings such as college students and adolescents. By the early 1950s, there were a total of 14 critiques and responses to them that focused on several main themes:

1. Groups of "problem" and "satisfactory" employees had similar profiles.
2. The Scale did not identify "successful" applicants in some settings, such as pilots.
3. Statistical analyses of new sets of data did not match the data reported by Humm and Wadsworth.

By the mid-1950s, published research on the Humm-Wadsworth had disappeared and the Scale rode off into the sunset, never to be seen again.

Humm and Wadsworth made several innovations with their Temperament Scale that will be continued with the MMPI:

1. Questions (items) were selected only if they actually differentiated among known groups of individuals.
2. Questions (items) for a number of different components were combined into a single inventory.
3. Questions (items) for a specific component of temperament were summed for a total score.
4. Total scores on each of the components were roughly equated so that differences within individuals could be examined.
5. Total scores were plotted on a profile.

Summary

The early personality inventories constructed primarily on a rational basis were generally unsuccessful for a variety of reasons, and the pendulum started swinging toward personality inventories constructed on an empirical basis such as the MMPI. These shortcomings should not be interpreted as an indictment of the general procedure, however. In the following decades, several widely used personality tests were developed at least partly on a rational basis, such as the Edwards Personal Preference Schedule (Edwards, 1959) and the Personality Research Form (Jackson, 1968). The MMPI gradually embraced scales developed on a rational basis beginning when Wiggins (1966) successfully constructed 13 content scales for the MMPI. Butcher, Graham, Williams, and Ben-Porath (1990) continued this movement with their new content scales for the MMPI-2. Wiggins (1973) provides an excellent, in-depth analysis of the relative merits of empirically and rationally derived scales.

CONSTRUCTION OF THE MMPI

Out of the psychometric wilderness of the early 1930s appeared two men, Starke Hathaway and J. C. McKinley, who, under the banner of empiricism, waged a new battle for the scientific advancement of the assessment of psychopathology. They sought to develop a multifaceted or multiphasic personality inventory, now known as the MMPI, that would surmount the shortcomings of the previous personality inventories, some of which were described earlier. Instead of using independent sets of tests, each with a special purpose, Hathaway and McKinley (1940) included in a single inventory a wide sampling of behavior of significance to psychologists. They wanted to create a large pool of items from which various scales could be constructed, in the hope of evolving a greater variety of valid personality descriptions than was currently available.

To this end, Hathaway and McKinley (1940) assembled more than 1,000 items from psychiatric textbooks, other personality inventories, and clinical experience. After deleting duplicate items and items that they considered relatively insignificant for their purposes, they arrived at a sample of 504 items. Approximately 25 percent of these 504 items are very similar to questions that were found on the Humm-Wadsworth Temperament Scale. Hathaway and McKinley did not provide a rationale for deleting insignificant items. Although potentially useful items may have been discarded that would have made the MMPI item pool more comprehensive, this issue was not important to them because they used an empirical method of item selection. The items were written as declarative statements in the first-person singular, rather than as questions that had been standard practice in previous inventories. Most of the items were written in the affirmative because they were easier for the person to understand. Hathaway and McKinley (1940) arbitrarily classified the items under 25 headings as a convenience in handling and in an effort to avoid duplication (Table 1.2). However, they did not attempt to obtain any particular number of items for a category or to ensure that an item was actually properly classified in a category. Table 1.2 shows that some categories are heavily overrepresented and other categories are underrepresented. The issue of the comprehensiveness of the MMPI item pool, and more accurately, its lack of comprehensiveness, however, has become more relevant in recent years because of the increasing focus on item content with Wiggins' (1966) content scales, the MMPI-2 content scales (Butcher et al., 1990), and the MMPI-2-RF (Ben-Porath & Tellegen, 2008).

TABLE 1.2 Content Categories for MMPI Items

CONTENT CATEGORY	NUMBER OF ITEMS
Social attitudes	72
Political attitudes, law and order	46
Morale	33
Affect, depressive	32
Delusions, hallucinations, illusions, ideas of reference	31
Family and marital	29
Phobias	29
Affect, manic	24
Habits	20
Religious attitudes	20
General neurologic	19
Sexual attitudes	19
Occupational	18
Lie	15
Obsessive, compulsive	15
Educational	12
Cranial nerves	11
Gastrointestinal	11
Vasomotor, trophic, speech, secretory	10
General health	9
Sadistic, masochistic	7
Genitourinary	6
Motility and coordination	6
Cardiorespiratory	5
Sensibility	5
Total	504

Note: The category names and sizes are from Hathaway and McKinley (1940).

Using these 504 items, Hathaway and McKinley (1940) next constructed a series of quantitative scales that could be used to assess various categories of psychopathology. In selecting items for a specific scale (e.g., Scale 1 [Hypochondriasis (Hs)]), they used an empirical approach. The items had to be answered differently by the criterion group (e.g., hypochondriacal patients) as compared with normal groups. Their approach was strictly empirical (i.e., no theoretical rationale was posited as the basis for accepting or rejecting items on a specific scale). Because of this empirical approach, it is not always possible to discern why a particular item distinguishes the criterion group from normal groups. Rather, items were selected solely because the criterion group answered them differently than other groups.

Scale 1 (Hypochondriasis [Hs]) was constructed first (McKinley & Hathaway, 1940). (It is now customary to identify each scale by its number rather than its name. The use of the scale number reduces the emphasis placed on diagnostic labels such as hypochondriasis, schizophrenia, and so on, and encourages the clinician to be aware of the empirical correlates of specific scores on each scale.)

This choice to construct Scale 1 first was not simply fortuitous. Hypochondriasis is one of the simpler, more distinct diagnostic categories, and hypochondriacs also were one of the largest groups of patients available to McKinley and Hathaway. Because the procedure for developing Scale 1 typifies the procedure for most of the clinical scales, it will be described in detail. Later, the development of the other clinical scales will be described only in cases where the procedure differs.

The first step in developing Scale 1 (Hypochondriasis [Hs]) was to select an appropriate criterion group. Using a diagnostic classification as the basis for the criterion group selection was logical because McKinley and Hathaway's intent was to develop an inventory to aid in differential diagnosis. They defined hypochondriasis as an *abnormal neurotic concern over bodily health*, excluding the symptomatic occurrence of hypochondriacal features in psychotic individuals. Using this definition, they selected 50 cases of pure, uncomplicated hypochondriasis as their criterion group.

The next step was to select groups of normal individuals. The primary normative group, which served as the reference group for determining the standard MMPI profile for over 50 years, consisted of 724 individuals who were friends or relatives of patients in the University Hospitals in Minneapolis. The only criterion for exclusion was if an individual was currently receiving treatment from a physician. This group reflected a fairly representative cross-section for gender and marital status of the Minnesota population aged 16 to 55 in the late 1930s. Dahlstrom, Welsh, and Dahlstrom (1972) reported that all of the persons in the primary normative group were white because very few members of any ethnic minority other than American Indians resided in Minnesota at that time. The normative groups for the MMPI-2 will be described later in this Chapter, and the use of the MMPI and MMPI-2 with ethnic minorities will be described in Chapter 11.

Four additional normative groups were used in the development of Scale 1 (Hypochondriasis [Hs]) and other clinical scales. Two normative groups were formed to assess whether "nuisance" variables such as age, socioeconomic class, or education were influencing differential item endorsement by members of the criterion group and the primary normative group. One group consisted of 265 precollege high school graduates who came to the University of Minnesota Testing Bureau for precollege guidance. The other was composed of 265 skilled workers from local Works Progress Administration projects. A third normative group consisted of 254 patients who were hospitalized for some form of physical disease in the general wards of the University Hospitals. None of these patients had obvious psychiatric symptomatology. The fourth general normative group consisted of 221 patients in the psychopathic unit of the University Hospitals, regardless of diagnosis.

Once the criterion group and the other normative groups were established, the process of item selection began. For the criterion group and each of the normal groups, the frequency of "true" and "false" responses was calculated for each item. An item was considered significant and was tentatively selected for a scale if the difference in frequency of response between the criterion group and the normative groups was at least twice the standard error of the proportions of "true/false" responses of the two groups being compared. For example, the response frequencies for two potential items for Scale 1 are provided in Table 1.3. In this example, only two groups, the criterion group of hypochondriacs and the original normative group, are compared; before any items were finally selected, the criterion group would be compared with the other normative groups as well.

The following formula was used for the test of the significance of the difference between two independent proportions:

$$Z = p_1 - p_2/\sqrt{(pq[(1/n_1) + (1/n_2)])}$$

where

p = the proportion of "true" responses in the total group = $211 + 17/262 + 50 = 228/312 = .73$

p_1 = the proportion of "true" responses in the first sample = $211/262 = .81$

TABLE 1.3 Frequency of Response by Group for Two Possible Items for Scale 1 (Hypochondriasis [Hs])

	GROUP			
	Normals[a]		*Hypochondriacs*[b]	
ITEM[c]	*True*	*False*	*True*	*False*
1. I am in good health most of the time.	211(81%)	51(19%)	17(34%)	33(66%)
2. I have headaches more often than most people.	10(4%)	252(96%)	5(10%)	45(90%)

[a] $n = 262$
[b] $n = 50$
[c] MMPI items are copyright. These items are similar to items on Scale 1.

p_2 = the proportion of “true” responses in the second sample = 17/50 = .34

$q = 1 - p = 1 - p = 1.0 - .73 = .27$

n_1 = the number of persons in the first sample = 262

n_2 = the number of persons in the second sample = 50

Substituting these values in the preceding formula results in the following:

$$\begin{aligned} Z &= .81 - .34/\sqrt{((.73)(.27)[(1/262)} + (1/50)]) \\ &= .47/.07 = 6.81 \end{aligned}$$

Checking a standard table of *Z* values shows that this *Z* value has a probability less than .001. Hathaway and McKinley considered significant any percentage difference of at least twice the standard error of the independent proportions, or any *Z* equal to or greater than +/−2. Since a *Z* of +/−2 has a probability slightly less than .05 using a two-tailed test, they essentially selected only items that were significant beyond the .05 level. Thus, the first item in the preceding example would be tentatively included in Scale 1 (Hypochondriasis [Hs]), and a “false” response would be the “deviant” answer because the hypochondriacal patients responded more frequently in the “false” direction. If this item also differentiated the hypochondriacal group from the other normative groups using an identical procedure, it would then be included on Scale 1.

Using the same procedure for the second sample item would result in substituting the following values in the formula:

$$\begin{aligned} Z &= .04 - .10/\sqrt{((.048)(.952)[(1/262)} + (1/50)]) \\ &= -.06/.10 = -0.60 \end{aligned}$$

This item would not be included on Scale 1 (Hypochondriasis [Hs]) because the proportions of endorsement are not significantly different between the two groups.

Having selected items according to this procedure, Hathaway and McKinley then eliminated some of them for various reasons. First, the frequency of the criterion group’s response was required to be greater than 10 percent for nearly all items; those items that yielded infrequent “deviant” response rates from the criterion group were excluded even if they were highly significant statistically because they represented so few criterion cases. Additionally, items whose responses appeared to reflect biases on variables such as marital status or socioeconomic status were excluded.

Finally, Hathaway and McKinley rejected a few more of the tentatively selected items that, after a rational inspection of the list, they concluded were not germane to the construct of hypochondriasis. Correlations between each item and the total score on the scale were not calculated nor were any other psychometric bases used in selecting items. The psychometric problems that later were discovered with some of the validity and clinical scales arose because these issues were not considered when each scale was constructed. These problems will be discussed later as appropriate when each scale is reviewed.

The preliminary Scale 1 (Hypochondriasis [Hs]) consisted of 55 items that had been identified by this procedure. The next step was weighting or combining them into a scale. Evaluation of several methods of weighting individual items showed no advantage over using unweighted items. Therefore, each item simply received a weight of “1” in deriving a total score. In other words, a person’s score on Scale 1 is equal to the total number of items that the individual answers in the same manner as the criterion group.

The responses of the normative group consisting of general psychiatric patients helped to refine Scale 1 (Hypochondriasis [Hs]). A fair number of these psychiatric patients obtained high scores on this scale although the psychiatric staff had not noted the presence of hypochondriasis. To eliminate this potential source of bias, the responses of 50 patients who had no hypochondriacal symptoms but who obtained the highest scale scores on the preliminary Scale 1 were contrasted with the original criterion group of 50 hypochondriacal patients. Items showing a significant difference in frequency of endorsement between these two groups were located and combined into a separate grouping, known as the correction of Scale 1. (This correction of Scale 1 should not be confused with the K-correction of Scale 1, which will be discussed later.) For each of these correction items that an individual answered in the nonhypochondriacal direction, one point was subtracted from the total score on Scale 1. Cross-validation revealed that the corrected score on Scale 1 was more effective in differentiating the groups than the original uncorrected score.

The normative group with physical disease also was used in developing Scale 1 (Hypochondriasis [Hs]). This group scored more like the normal group than like the hypochondriacal group on the corrected Scale 1. Thus, their actual physical symptoms appeared to alter their total scores only moderately in the direction of hypochondriasis.

Scale 1 (Hypochondriasis [Hs]) was modified again in order to differentiate Scale 1 more clearly from Scale 3 (Hysteria [Hy]). McKinley and Hathaway (1944) eliminated from Scale 1 those correction items that also appeared on Scale 3, thus arbitrarily making Scale 1 into a somatic ailments scale. They also eliminated some of the original items from Scale 1 that did not separate hypochondriacs from the normative group under subsequent analyses. This final step resulted in the 33 items that are currently used on Scale l on the MMPI.

Soon after the development of Scale 1 (Hypochondriasis [Hs]; McKinley & Hathaway, 1940), five other clinical scales were developed: Scale 2 (Depression [D]; Hathaway & McKinley, 1942); Scale 7 (Psychasthenia [Pt]; McKinley & Hathaway, 1942); Scale 3 (Hysteria [Hy]); Scale 4 (Psychopathic Deviate [Pd]); and Scale 9 (Hypomania [Ma]; McKinley & Hathaway, 1944). The description of the construction of three other clinical scales—Scale 5 (Masculinity-Femininity [Mf]), Scale 6 (Paranoia [Pa]), and Scale 8 (Schizophrenia [Sc])—was not published until 1956 (Hathaway, 1956), although these three scales had been used routinely for more than a decade. (More detailed information on each of these scales will be provided in Chapter 4.)

Scale 5 (Masculinity-Femininity [Mf]) was developed somewhat differently than the other clinical scales. Another 55 items, mostly related to sexual orientation, were added to the MMPI item pool after the data already had been collected from the original normative sample. (The addition of 55 items to the original 504 items on the MMPI would produce an item pool of 559 items. Because the MMPI contains only 550 items, it is not clear what happened to the other 9 items [W. G. Dahlstrom, personal communication, 1979].) Thus, the criterion group of male homosexuals who were used in developing Scale 5 could not be contrasted with the original normative group on these 55 items. Consequently, 54 male soldiers were used as one of the normative groups for this scale, and items that distinguished them from the male homosexuals were included on Scale 5. In addition, items that differentiated men from women within the normative sample were included on this scale. The effects of these different construction procedures for Scale 5 will be explored more fully in Chapter 4.

In 1946, Scale 0 (Social Introversion [Si]) was added to the MMPI (Drake, 1946), completing the standard MMPI clinical profile. Scale 0 also was constructed differently from the other clinical scales. Drake selected MMPI items that differentiated 50 college students who scored above the 65th percentile on the Minnesota T-S-E Inventory (Evans & McConnell, 1941) from 50 students who scored below the 35th percentile.

The Minnesota T-S-E Inventory assesses introversion-extroversion in three areas: thinking (T), social (S), and emotional (E). Drake (1946) limited his initial work to the social introversion-extroversion area, or, more specifically, he investigated introversion-extroversion only in the social area as assessed by the Minnesota T-S-E Inventory. Although Drake conducted his analysis on men and women separately, their norms were so similar that he combined the normative data for the two genders into a single group before finally incorporating it into the standard MMPI profile. (This issue will be explored more fully in Chapter 4.)

Before proceeding further it is necessary to define what is meant by an MMPI *codetype* because this term will be used frequently before codetypes actually are discussed in Chapter 5. Because the items on the MMPI scales were selected empirically, as described earlier, it is not appropriate to say that the client is hypochondriacal because he or she has an elevated score on Scale 1 (Hypochondriasis [Hs]). Rather, it must be said that the client endorses the items *like* the hypochondriacal patients that were used in developing Scale 1, and any correlates of an elevated score also must be determined empirically. Consequently, it has become common practice to refer to the MMPI clinical scales by number (1, 2, 3, 4, 5, 6, 7, 8, 9, or 0) rather than by name to help clinicians avoid this interpretive error. (By convention, Scale 0 [Social Introversion [Si]] is called Scale “zero,” not Scale “ten.”) Interpretation of the MMPI or MMPI-2 then is based on the codetype, which is the two highest elevated clinical scales at a T score of 65 or higher. An MMPI-2 codetype is defined by writing the numbers of the two scales involved with the most elevated one first. For example, if a client’s two highest scores on the MMPI-2 are on Scales 2 (Depression [D]) and 7 (Psychasthenia [Pt]), and both are above a T score of 65 but Scale 7 is higher than Scale 2, then the client’s codetype would be 7-2. If the two highest clinical scales have identical T scores, they are listed in numerical order. In this example, if both Scales 2 and 7 had identical T scores of 75, the client’s codetype would be 2-7. If only one clinical scale is elevated to a T score of 65 or higher, the codetype is called a “Spike” codetype. If Scale 2 were the only clinical scale elevated to a T score of 65 or higher, the codetype would be a Spike 2. There are 90 possible two-point and spike codetypes on the MMPI and MMPI-2 following this procedure. Finally, if no clinical scale is elevated above a T score of 65 or higher, it is called a “Within-Normal-Limits” codetype. Codetypes will be explored more completely in Chapter 5.

ASSESSMENT OF TEST-TAKING ATTITUDES

Although self-ratings provided through item responses can be useful because direct observations of behavior are often impractical, impossible, or inefficient, individuals sometimes fail to provide an accurate self-report (i.e., a self-report that accurately reflects how others perceive their behavior). There are several possible reasons for their inaccurate self-description. First, although persons constructing test items generally assume that each item has essentially the same meaning to all persons taking the test, this assumption is not always appropriate. For example, for a test item such as “I dream frequently,” persons may interpret “frequently” to mean once a day, once a week, or once a month and respond “true” or “false” accordingly. One client might endorse this item as being “true” because he has dreams at least once a month; another might endorse this item as being “false” because she has dreams only once a week. The ambiguity inherent in any single item makes it extremely difficult to obtain an accurate self-description because the person answering a specific item and an observer rating the person on that item’s content may interpret the item somewhat differently. Second, individuals vary in their self-awareness and in their ability or willingness to report the appropriate behaviors. Third, the rational method of test construction also requires that the test developer be knowledgeable about the relationship between people’s responses to individual items and the construct being assessed. The fallacies and errors in earlier rationally derived personality inventories suggest that it is very difficult for the test developer to have this depth of understanding of the dynamics of a personality inventory.

These problems can be demonstrated by the response of schizophrenic patients to an item on the Bernreuter Personality Inventory (Bernreuter, 1933) such as “I sometimes cross the street to avoid meeting people.” A test developer would likely make the a priori assumption that schizophrenic patients would respond “true” to this item. In fact, they answer such an item “false” more often than normal individuals (Landis & Katz, 1934). This response does not mean that this specific behavior is actually characteristic of schizophrenic patients; rather, it means schizophrenic patients say it is characteristic of themselves. As such, it can be treated like any other verbalization the individual makes. It indicates how the person interprets the item and how the person thinks, perceives, and feels even though it may actually be untrue. Although this statement is not literally true, it still provides useful diagnostic information about the individual.

Although these issues unquestionably exist in the interpretation of the content of individual items on the MMPI, they do not invalidate it. The empirical approach to item selection used by Hathaway and McKinley (1940), in fact, freed them of these problems because it assumes that the client’s self-report is just that and makes no a priori assumptions about the relationships between the client’s self-report and the client’s behavior. Items are selected for inclusion in a specific scale only because the criterion group answered the items differently than the normative groups regardless of whether the item content is actually an accurate description of the criterion group. Any relationship between clients’ responses on a given scale and their behavior must be demonstrated empirically. The interested reader should consult Meehl’s (1945) article, which explores this issue in greater depth, and the section on critical items in Chapter 7.

Because Hathaway and McKinley (1940) developed the MMPI under the banner of empiricism, they recognized that the honesty or frankness with which the client responds to the items needs to be assessed empirically each time the MMPI is administered rather than blithely assume that the client has answered the items appropriately. It is possible that a client might adopt a test-taking attitude other than that desired by the test developer. Clients may decide, for whatever reason, to increase (exaggerate, or “fake bad”) or decrease (deny, or “fake good”) their estimate of the frequency of a given behavior or the severity of a specific symptom being assessed by the test instrument, or clients may respond randomly to the test items because of an unwillingness or inability to respond appropriately. In either case it is important for the interpreter of the test inventory to be aware of the possibility that clients have responded inappropriately. Previous test developers often paid lip service to the importance of appropriate test-taking attitudes, but they did not provide specific directions on how to develop or maintain those attitudes. More importantly, they did not provide a means of assessing whether those attitudes actually were present. In the development of the MMPI this problem was assessed directly through what are now called the *validity scales.* The robustness of the MMPI validity scales has been one of the major factors contributing to its longevity in the field of self-report inventories.

Meehl and Hathaway (1946) were convinced of the necessity of assessing two dichotomous categories of test-taking attitudes: defensiveness (“faking-good”) and excessive reporting of symptoms (“faking-bad”). (These two categories will be called “self-favorable” and “self-unfavorable,” respectively, throughout later sections of this book to avoid the connotations inherent in the terms of “faking-good” and “faking-bad,” because it is not always clear whether the person’s motivation for distorting responses is intentional.)

Meehl and Hathaway (1946) considered three possible approaches to assess these two categories of test-taking attitudes. First, they could give the client an opportunity to distort the responses in a specific way and observe the extent to which the client did so. One way of implementing this approach would be to repeat items within the MMPI, phrased either in an identical manner or in the negative rather than the affirmative. A large number of inconsistent responses would suggest that the client was either unable or unwilling to respond consistently.

Although Meehl and Hathaway rejected this solution, the MMPI group booklet form included 16 identically repeated items that was used to detect inconsistent responding. However, these 16 items were deleted and not replaced in developing the MMPI-2, so this solution has been rejected again.

Second, Meehl and Hathaway (1946) considered providing an opportunity for the client to answer favorably when a favorable response would almost certainly be untrue. This solution would involve developing a list of extremely desirable but very rare human qualities. If a client endorsed a large number of these items, it is highly probable that the responses would be dishonest. The Lie (L) scale was developed specifically for this purpose. Items for the L scale, based on the work of Hartshorne and May (1928), reflect behaviors that, although socially desirable, are rarely true of a given individual. A large number of responses in the deviant direction on the L scale would suggest response distortion.

The Infrequency (F) scale was developed according to a variant of this second approach for assessing test-taking attitudes. Items for the F scale were selected primarily because they were answered with a relatively low frequency by a majority of the original normative group. In other words, if a client endorsed a large number of the F scale items, the client would be responding in a manner that was atypical of most people in the normative group. In addition, the items include a variety of content areas so that any specific set of experiences or interests would be unlikely to influence any specific individual to answer many of the items in the deviant direction. The F scale effectively identified individuals who were intentionally reporting psychopathology. However, schizoid individuals and persons who were overly pessimistic about themselves also obtained high scores. Therefore, additional procedures were needed to separate these two groups of persons from those who intentionally reported their psychopathology or misunderstood the items. Meehl and Hathaway (1946) thought the Lie (L) scale would serve this function, which provided another reason for its use as a validity scale.

Third, Meehl and Hathaway considered using an empirical procedure to identify items that elicit different responses from persons taking the test in an appropriate fashion and those who have been instructed to “simulate” psychopathology. Gough’s Dissimulation scale (Ds; Gough, 1954, 1957), which was based on this procedure, will be described in Chapter 3.

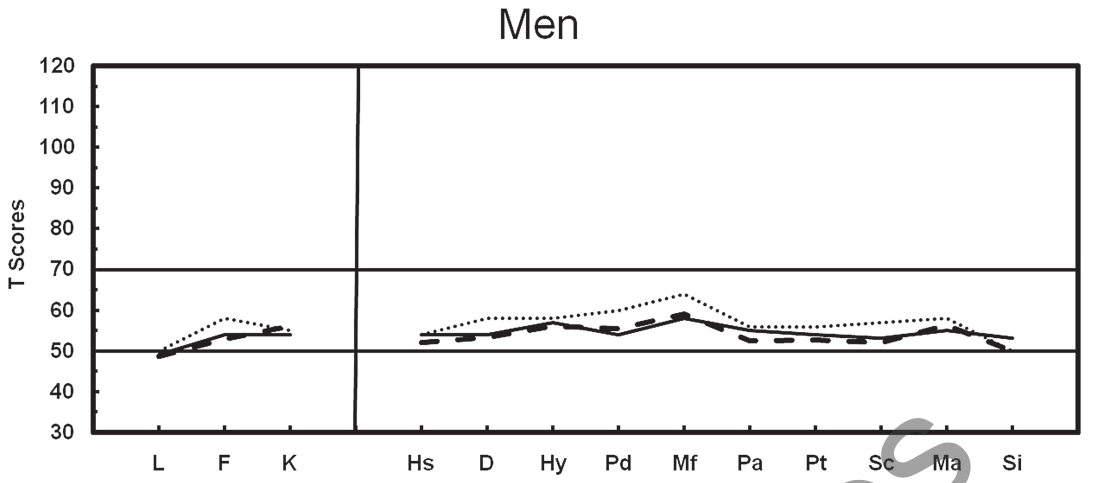
Meehl and Hathaway (1946) adopted a variant of this third approach in developing a third validity scale, the Correction (K) scale. Their task was to differentiate persons known to have psychopathology, who were hospitalized and yet obtained normal profiles (no clinical scales at or above a T score of 70), from normal individuals who for some reason obtained elevated profiles. They selected 25 male and 25 female patients diagnosed as having psychopathic personalities, alcoholism, and other behavior disorders who (1) had a T score of 60 or higher on the Lie (L) scale, which would indicate some form of response distortion, and (2) had diagnoses indicating that they should have elevated profiles, but (3) had actual profiles in the normal range. Based on a comparison of this group with the original normative sample on all items, 22 items were selected that showed at least a 30 percent difference in the response rates of the two groups.

It was later found that these 22 items generally did an adequate job of identifying defensiveness in most patients; however, depressed and schizophrenic patients tended to get low scores. To counteract this tendency, 8 items were added and scored to differentiate these two groups from the original normative group. This final step resulted in the 30-item Correction (K) scale, which is currently used. Meehl and Hathaway (1946) also empirically determined the proportions of K that when added to a clinical scale would maximize the discrimination between the criterion group and the normative group. Because Meehl and Hathaway determined the optimal weights of K to be added to each clinical scale in a psychiatric inpatient population, they warned that with maladjusted normal populations and other clinical populations, other weights of K might serve to maximize the identification of individuals with psychopathology. This issue of the optimal weights to be added to each clinical scale in different populations will be discussed in Chapter 3 when the K scale is examined in more depth.

APPROPRIATENESS OF MMPI NORMS

The issue of whether the items and norms for the MMPI that were developed in the early 1940s are appropriate for contemporary use has been raised repeatedly and debated widely (cf. Butcher, 1972; Colligan et al., 1983; Faschingbauer, 1979). Since the typical individual in the original Minnesota normative group was “about thirty-five years old, was married, lived in a small town or rural area, had had eight years of general schooling, and worked at a skilled or semiskilled trade (or was married to a man with such an occupation level)” (Dahlstrom et al., 1972, p. 8), it seems apparent that there have been numerous changes in our society over the ensuing decades.

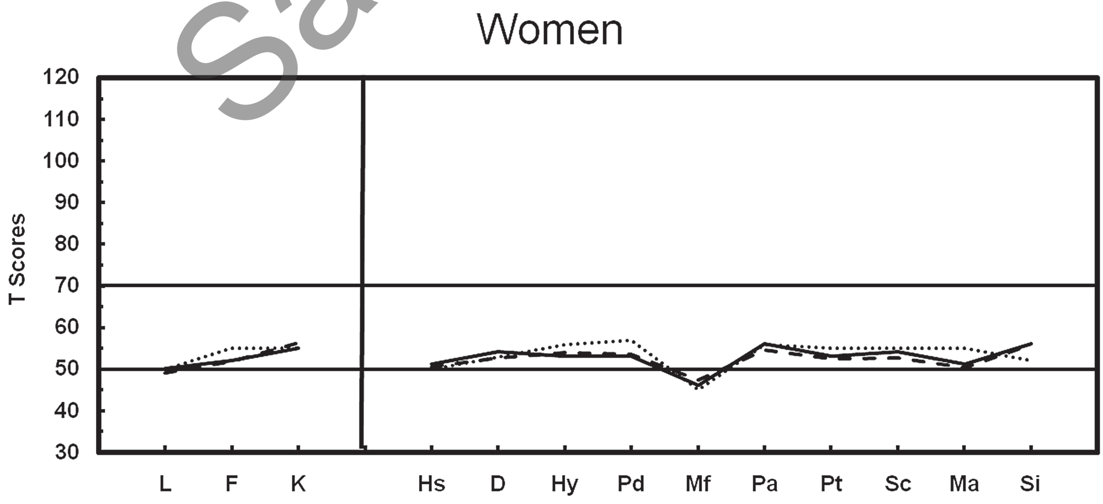
Pancoast and Archer (1989) collated the existing literature on the performance of normal individuals on the MMPI to assess the adequacy of the norms based on the original Minnesota normative group. The mean MMPI



PROFILE 1.3 MMPI Basic Validity and Clinical Scales in Men
----- Pancoast and Archer (1989) ___ Colligan et al. (1983) ...... Butcher et al. (1989)

profile for these normal men (Profile 1.3, dashed line) and women (Profile 1.4, dashed line) showed T scores near 55 for the Scales Correction (K), 3 (Hysteria [Hy]), 4 (Psychopathic Deviate [Pd]), and 9 (Hypomania [Ma]). Only on the Scales Lie (L) and 1 (Hypochondriasis [Hs]) did the mean T scores approach 50. Pancoast and Archer found that studies as early as 1949 demonstrated that normal individuals showed generally small but consistent variations from the mean scores of the original Minnesota normative group. Two conclusions can be drawn from the data summarized by Pancoast and Archer. First, the scores of normal individuals may have been slightly different from the original Minnesota normative group on the standard validity and clinical scales since the MMPI was first developed. Second, there have been only small changes in normal individuals across the five subsequent decades as reflected by their mean T scores on the standard validity and clinical scales.

Greene (1990) examined the changes in the standard validity and clinical scales on the MMPI within four frequently occurring codetypes (Spike 4, 2-4/4-2, 2-7/7-2, and 6-8/8-6) in samples of psychiatric patients over a time span of 40 years. The mean MMPI profiles were virtually identical within all four codetypes for all four samples. It appears that the MMPI scale scores of psychiatric patients have been very stable over this time span. Greene's data did not address whether the empirical correlates of these codetypes remained unchanged across the 50 years that the



PROFILE 1.4 MMPI Basic Validity and Clinical Scales in Women
----- Pancoast and Archer (1989) ___ Colligan et al. (1983) ...... Butcher et al. (1989)

MMPI has been in use. However, the stability of the MMPI scale scores across these years would at least suggest that the correlates probably have not changed. Of course, empirical data are needed to answer this question.

Greene (2000, Table 1.4, pp. 14–15) compared four MMPI-2 codetypes in a sample of psychiatric inpatients (Greene & Schinka, 1995) and psychiatric outpatients (Caldwell, 1997a) collected over different time periods in the 1990s. These MMPI-2 profiles were very similar for all of the codetypes except for the 6-8/8-6 codetypes in which the inpatients score 5 to 10 T points higher than the outpatients on a number of the scales. The stability in these MMPI-2 codetypes would be expected given the stability that was found for MMPI codetypes. When the MMPI-2 profiles were compared to the corresponding MMPI profiles, the common pattern was for T scores to be 5 to 10 T points higher on the Infrequency (F) scale and about 5 T points lower on the Correction (K) scale, and Scales, 3 (Hysteria [Hy]), and 9 (Hypomania [Ma]). The variations in the T scores on Scales F and K between the MMPI and MMPI-2 would be expected given that Hathaway and McKinley assigned the T scores to the validity scales arbitrarily and they recognized that these T scores were inaccurate (Hathaway & McKinley, 1951). The variations on the clinical scales usually do not affect the scales that are defining the codetype and thus would have only minor impact on the interpretation of the codetype. This complex issue of the relationship between MMPI and MMPI-2 codetypes is explored in Chapter 6 of Greene (2000).

The finding that normal individuals and psychiatric patients have shown only minor changes on the standard validity and clinical scales of the MMPI and MMPI-2 across 50 years is very surprising, and would suggest that the MMPI may not be as outdated as many people have thought. The changes that have occurred since the MMPI was developed have been examined by Colligan et al. (1983) in developing contemporary norms for the MMPI and the current restandardization of the MMPI that resulted in the MMPI-2 (Butcher et al., 1989). These two projects will now be examined in turn.

A CONTEMPORARY NORMATIVE STUDY OF THE MMPI

Colligan et al. (Colligan et al., 1983, 1989; Colligan, Osborne, Swenson, & Offord, 1984) investigated whether the original MMPI norms were appropriate for contemporary use. They essentially replicated the data-collection procedures employed by McKinley and Hathaway (1940) and gathered a representative sample of individuals living within 50 miles of the Mayo Clinic in Rochester, Minnesota. "Persons having chronic diseases (e.g., diabetes) were excluded from the study, as were patients receiving cancer treatment, those with rheumatoid or other types of arthritis, those described as being chemically dependent, having a learning disability, or being mentally retarded, and persons undergoing psychotherapy" (Colligan et al., 1983, pp. 74–75).

Their final sample consisted of 1,408 white individuals (646 men and 762 women), whose mean age was in their mid-40s and who had a mean of 13 years of education. Nearly three-fourths of them were married. These individuals were somewhat older and better educated than the original Minnesota normative sample that was described earlier. Colligan et al. also selected a subset of these individuals "in proportion to the age and sex in the general population of adult whites in the United States, as determined by the 1980 census" (1983, p. 87) so that they could make more direct comparisons with the original normative group because the population of the United States had increased in age and become better educated in the ensuing four decades.

Profiles 1.3 (solid line) and 1.4 (solid line) provide the mean profile for these contemporary men and women, respectively, plotted on the original MMPI norms. As can be seen in these two profiles, the men average 3 to 8 T score points higher on the clinical scales and the women average 1 to 6 T score points higher (except on Scale 5 [Masculinity-Femininity (Mf)] where they are 4 points lower) on the clinical scales than the original Minnesota normative group.

Two basic points can be made based on the data presented in Profiles 1.3 and 1.4. First, there are some differences in MMPI performance across the five decades that the MMPI has been in use, although these differences are not as substantial as might have been expected, given the changes in our society in the last 50 years. When the data in these two Profiles reported by Colligan et al. (1983) (solid lines) are compared with Pancoast and Archer (1989) (dashed lines), there is further support for the statement that the scores of normal individuals may have been slightly different from the original Minnesota normative group on the standard validity and clinical scales since the MMPI was first developed. Second, it appears that these changes average less than one-half standard deviation (5 T-points) and these small changes in profile elevation are not likely to have major impact on the clinical interpretation of the MMPI.

Colligan et al. (1983) continued the procedure of using K-corrected scores and they used the same correction weights on the same clinical scales (Scale 1 [Hypochondriasis (Hs)]; Scale 4 [Psychopathic Deviate (Pd)]; Scale 7 [Psychasthenia (Pt)], Scale 8 [Schizophrenia (Sc)], and Scale

9 [Hypomania (Ma)]) that had been suggested by Meehl and Hathaway (1946). However, they made one major change in the method whereby raw scores are transformed into T scores by using normalized rather than linear T scores. The results of using different methods to compute T scores will be explored further in the next Chapter.

There has been only limited research with the Colligan et al.'s (1983) norms. Colligan et al. (1985) reported the frequency with which codetypes occurred in four clinical samples, and the concordance between their contemporary norms and the original MMPI norms. Concordance of codetypes between the two sets of norms ranged from 40 to 60 percent for women and from 50 to 70 percent for men, whereas agreement on single scales ranged from 66 to 79 percent for women and from 69 to 79 percent for men.

Miller and Streiner (1986) reported the concordance between profiles generated by contemporary norms and the original MMPI norms in a large sample of psychiatric patients. They found that 48.4 percent of the profiles showed no changes in the two highest clinical scales, and another 15.1 percent of the profiles had the two highest clinical scales reversed. Thus, 63.5 percent of the profiles had the same codetype using the two sets of norms. In 23.6 percent of the profiles, the highest clinical scale remained the same while another clinical scale became the second highest scale. A totally unique codetype was produced in 9.4 percent of the profiles. Although it is important to know the concordance between codetypes generated by the two sets of norms, the primary issue remains whether the original or contemporary norms more accurately reflect external correlates.

Tables for converting raw scores into T scores so that the clinician can compare a patient's performance with a contemporary adult sample are available in Colligan et al. (1983, 1989). Separate tables are provided for men and women, so the clinician should be careful to use the correct table. Hsu and Betman (1986) have provided tables for converting the T scores for the original MMPI normative group into Colligan et al.'s (1983) contemporary norms and vice versa. Colligan et al. (1983) also illustrate a standard profile sheet for use with their contemporary norms (p. 421).

DEVELOPMENT OF THE MMPI-2

The MMPI-2 (Butcher et al., 1989) represents the restandardization of the MMPI that marks the advent of a new era of clinical usage and research of this venerable inventory. Restandardization of the MMPI was needed to provide current norms for the inventory, develop a nationally representative and larger normative sample, provide appropriate representation of ethnic minorities, and update item content where needed. Continuity between the MMPI and the MMPI-2 was maintained because new criterion groups and item derivation procedures were not used on the standard validity and clinical scales. Thus, the items on the validity and clinical scales of the MMPI are essentially unchanged on the MMPI-2 except for the elimination of 13 items based on item content and the rewording of 68 items.

The profile form for the original MMPI (Profile 1.1) and the revised profile for the MMPI-2 (Profile 1.5) are very similar except for the addition of a number of new validity scales. Only on closer examination are any differences seen on the MMPI-2 profile form: the Cannot Say (?) scale has been moved to the bottom of the page, T scores of 65 are considered to be clinically significant instead of T scores of 70, and the T score distributions have been truncated at 30 so that T scores below 30 do not occur. (T scores also are truncated at 30 in the tables converting raw scores to T scores so information is not lost in using the standard profile form [Table A.1 in Butcher et al., 2001]).

In the development of the MMPI-2, the Restandardization Committee (Butcher et al., 1989) started with the 550 items on the original MMPI (i.e., they first deleted the 16 repeated items). They reworded 141 of these 550 items to eliminate outdated and sexist language and to make these items more easily understood. Many of these items were omitted on the original MMPI because clients did not understand them. Greene (1991, p. 57) provides examples of these items. Rewording these items did not change the correlations of the items with the total scale score in most cases (Ben-Porath & Butcher, 1989). The Restandardization Committee then added 154 provisional items that resulted in the 704 items in Form AX, which was used to collect the normative data for the MMPI-2.

When finalizing the items to be included on the MMPI-2, the Restandardization Committee deleted 77 items from the original MMPI in addition to the 13 items deleted from the standard validity and clinical scales and the 16 repeated items. Consequently, most special and research scales that have been developed on the MMPI are still capable of being scored unless the scale has an emphasis on religious content or the items are drawn predominantly from the last 150 items on the original MMPI. The content areas for these 77 items that were deleted plus the 13 items deleted from the validity and clinical scales can be seen in Table 1.4. Levitt (1990) also has grouped these 77 items into logical content categories and listed the actual items within each category.