Meehl, P. E. & Hathaway, S. R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology, 30*, 525-564.

The K Factor as a Suppressor Variable in the Minnesota Multiphasic Personality Inventory *

Paul E. Meehl and Starke R. Hathaway

Division of Psychiatry and the Department of Psychology, University of Minnesota

I. History and Problem

Among the very large number of structured personality inventories which have been published, it is by now quite generally admitted that there are relatively few which are of practical value in the clinical situation. There are a number of reasons, both obvious and subtle, for this fact, some of which will be developed by implication in the present paper. One of the most important failings of almost all structured personality tests is their susceptibility to "faking" or "lying" in one way or another, as well as their even greater susceptibility to unconscious self-deception and role-playing on the part of individuals who may be consciously quite honest and sincere in their responses. The possibility of such factors having an invalidating effect upon the scores obtained has been mentioned by many writers, including Adams (1941), Allport (1928, 1937, 1942), Bernreuter (1933a,b, 1940), Bills (1941), Bordin (1943), Eisenberg and Wesman (1941), Guilford and Guilford (1936), Humm and Humm (1944), Humm and Wadsworth (1935), Kelly, Miles and Terman (1936), Laird (1925), Landis and Katz (1934), Maller (1930), Olson (1936), Rosenzweig (1934, 1938), Ruch (1942), Strong (1943), Symonds (1932), Vernon (1934), Washburne (1935), Willoughby [and Morse] (1936) and others. One of the assumed advantages of the projective methods is that they are relatively less influenced by such distorting factors, although this assumption should be critically evaluated.

The existence of a distorting influence in test taking attitude is so obvious that it has hardly been thought necessary to establish it experimentally, although a number of investigations have demonstrated the effect. Frenkel-Brunswik (1939) investigated tendencies to self-deception in rating oneself, finding in some cases marked negative relations between self-judgments and the evaluation of others. Hendrickson (1932), cited by Olson (1936), reported that a group of teachers earned significantly more stable, dominant, extroverted and self-sufficient scores on the Bernreuter scales when instructed to take the test as though they were applying for a position, than when under more neutral instructions. Ruch (1942) showed that college students could fake extroversion on the Bernreuter to the extent of achieving a median at the 98th percentile of Bernreuter's norms, as contrasted with

* Supported by a research grant from the Graduate School of the University of Minnesota.

a "naive" median at the 50th percentile. Bernreuter (1933b) found that college students could produce marked shifts in their Bernreuter scores in the "socially approved" direction, although he interpreted this finding as indicating the comparative unimportance of the faking tendency. His reasoning was that had the need for giving socially approved responses operated in the first administration to any appreciable extent, the effect of special instructions to take this attitude should not have been great. This reasoning seems rather tenuous, inasmuch as the occurrence of a shift merely shows that conscious and permitted faking can produce greater effects than those which may have been operating in the "naive" original testing. The insignificant correlations between naive and faked scores were also used by Bernreuter to support his view, an argument which is not comprehensible to the present writers, especially in view of the probably gross skewness of the faked scores. What is clear from his investigation is that people are able to influence their scores to a considerable extent if they choose to, and that the average student's stereotype of what is "socially desirable" seems to be an individual who is dominant, self-sufficient, and stable. Maller (1930), Metfessel (1935), Olson (1936) and Spencer (1938) have studied the effects of anonymity on responses to self-rating situations and shown that the requirement of signing one's name has a definite effect on the scores. Kelly, Miles and Terman (1936) demonstrated the great ease with which scores on the Terman-Miles Masculinity-Femininity Test could be "faked" in either direction once the subjects had been let in on the secret of what the test measured. Strong (1943), Bills (1941), Steinmetz (1932), and Bordin (1943) have presented evidence of the ability of subjects to distort their interest patterns when taking the Strong Vocational Interest Blank.

It is a significant sociological fact about the psychologists that in spite of the strong reasons, both a priori and experimental, for accepting the reality of this phenomenon in objective personality testing, very few systematic efforts have been made to correct for it or to overcome it. In published articles one continually finds brief and inadequate references to the "assumption of frankness" and the necessity for arousing a "sincere desire to know oneself better," but the treatment is usually extremely sketchy and no very concrete suggestions are given for producing such test-taking attitudes nor, what is almost as important in practice, for determining the extent to which they have been present. It almost seems as though we inventory-makers were afraid to say too much about the problem because we had no effective solution for it, but it was too obvious a fact to be ignored so it was met by a polite nod. Meanwhile the scores obtained are subjected to varied and "precise" statistical manipulations which impel the student of behavior to wonder whether it is not the aim of the personality testers to get as far away from any unsanitary contact with the organism as possible. Part of this trend no doubt reflects the lack of clinical experience of some psych-

ologists who concern themselves with personality testing, and the very strong contemporary trend which stresses the statistical interrelationships of item responses much more than the relation of the latter to external non-test criteria. The establishment of "validity" (sic!) in terms of various criteria of internal consistency naturally leads to an unconscious neglect of the problem of non-test behavior correlates.

Among the many authors who recognize the problem there are a few who have made specific suggestions for its solution. The inclusion of special exhortations to frankness and objectivity in the test directions themselves is common, but we have no evidence as to its effectiveness. Obviously, if a subject is consciously determined to fake, he will do so; whereas if his motivation to distortion is of a more subtle, non-verbalized nature, such exhortations can hardly be expected to be efficacious. Another method is to attempt disguise of the items, so that the "significance" of a given response is less obvious. Traditional approaches to the measurement of personality render this technique practically impossible, inasmuch as the items are selected to begin with for their *obvious* psychological significance and hence unless changed so greatly as to no longer elicit the desired information, almost inevitably continue to betray their origin. An effective use of a set of "subtle" items is only possible when the initial item pool is very large and the *initial selection* (not only the final validation) of items is ruthlessly empirical. Those items whose significance would not have been guessed by the test-maker will then be equally mysterious to the testee. When the projective and role-playing components of test-taking behavior are clearly seen to be present in objective personality inventories (Meehl, 1945a), this approach to the problem is very fruitful. A simple strategem along the item-disguise line is to state about half of the items negatively, so that an affirmative response is not consistently a "bad" or maladjusted one. However, such techniques cannot eliminate the problem entirely.

A spurious anonymity using secret coding for identifying the testee is a possibility suggested by the studies cited above, but is clinically impractical for obvious reasons. The deception involved is not desirable, and in any case the clinical patient, unlike the sophomore student, knows perfectly well that the examiner is interested in *his* score individually. Lacking anonymity, it has been suggested by Olson (1936) that the name be signed at the conclusion of the administration instead of at the top of the page. This suggestion was carried into practice by Maller (1932) in his *Character Sketches.* This investigator also stated the questions in the "indirect" (third person) form, requiring the subject to indicate whether he was the *same* or *different* from the person described. Maller presents evidence that this procedure aroused considerably less annoyance in his subjects, although direct proof that this decrease in annoyance led to increased validity is lacking. For reasons which have been given in more detail elsewhere (Meehl,

1945a), it is doubtful whether the removal of personal reference is wholly desirable; since there is reason for believing that the same role-playings and self-deceptions which operate to invalidate *some* of our measurements are an important factor in making *other* measurements possible.

Another technique for reducing the effect of signing one's name is to have the items printed on cards which are then sorted by the subject, making all writing unnecessary and possibly lessening the feeling that one is making a permanent record of his personal failings. This has been done by Maller in a revised test (Personality Sketches) and by Hathaway and McKinley (1943) in the Minnesota Multiphasic Personality Inventory. The latter test will be referred to as MMPI.

Although all of these strategems may have a considerable value, especially in the aggregate, the fact still remains that they do not by any means remove the possibility of "faking." What is much more important, they are mainly directed at the sort of *conscious* falsehood which most writers have stressed, while ignoring the more subtle tendencies to self-deception which are probably of even greater importance in affecting scores. In the third place, they neglect to stress the existence of trends in the opposite direction—namely, those trends which exaggerate the apparent abnormality or maladjustment of the individual rather than soft-pedaling it. It is only natural that the tendency of a testee to put himself in a favorable light should have received more attention than the contrary tendency, which makes much less "sense" psychologically at least from a superficial point of view. There is evidence that this latter tendency does exist, however, and that it is a much more important factor in determining scores on personality inventories than has generally been supposed. Some of this evidence will be presented in the present paper, while other indications have been given elsewhere (Meehl, 1945b). It is also probable that certain systematic differences in item-interpretation, not necessarily a function of personality dynamics of the defensive or self-critical sort but relatively "neutral" psychologically (e.g. semantic variation), lead to score deviations that are misleading. Such problems have been investigated by Benton (1935), Eisenberg (1941), and Eisenberg and Wesman (1941).

A more fruitful attitude was taken by Rosenzweig (1934) in which he reiterated the fact of untrustworthiness of self-ratings and indicated that instead of trying to completely eliminate these sources of error we should recognize them and attempt to "correct" for them in interpreting the results. He says,

> "Astute phraseology in the instructions and questions of the test have sometimes been resorted to, but such expedients are rarely very effective. Might it not be more effective to recognize at the outset that such tests have certain limitations that can never be completely circumvented and then go on to the measurement of these limiting factors themselves, thus obtaining information by which a correction may be applied to the subject's answers?" (Rosenzweig, 1934).

Rosenzweig's specific proposal for achieving this end was to include among the usual self-rating items a set of items of the form "I should like to be the sort of man who …." on the theory that if we knew something of the strength of certain "ideal-self" trends in the person, we could make appropriate correction for these trends in interpreting responses to the traditional items. Rosenzweig never carried this idea into practice and there is no way of telling whether or not it would have worked. It seems to the writers that it would be relatively ineffective, since what is desired is not a statement of the strength or number of ideals for the self, but a measure of the extent to which they are allowed to distort responses. In other words, a subject might easily have quite lofty ideals verbally expressed, but might be too honest, insightful, objective, or self-critical to distort his responses into agreement with these ideals. It is, for example, rather characteristic of psychasthenic persons to express high and often unattainable ideals of perfection and achievement; whereas at the same time they are prone to be excessively self-critical, a fact which is psychometrically reflected in the negative correlation of the Pt (psychasthenia) scale of MMPI with some of the subtle "lie" scales which will be discussed below.

Maller (1932) attempted to solve this problem in another way in his *Character Sketches,* by including a small set of items which were supposed to measure the subject's "readiness to confide." The occurrence of very normal, well-adjusted scores in combination with a low measured "readiness to confide" would lead one to be sceptical of the validity of the measurement. This was a material advance in principle, except that the "readiness to confide" items were themselves self-ratings on that very readiness. In the later form called *Personality Sketches* Maller does not make use of this procedure so we may assume that it was unsuccessful or at least did not materially improve validity.

Carrying Rosenzweig's thinking to its logical conclusion, the obvious procedure to follow is to give the subject a good *chance* to distort his answers in accordance with some self-picture or conscious facade, and observe the extent to which he does so. The difficulty here is that such a procedure requires a knowledge of the objective facts (and the subjective facts!) which is usually inaccessible to us. Here there are three possibilities open to the test-builder. First, he may sidestep the problem of getting directly at the objective truth, and attempt to establish falsehood by obtaining internal contradictions. This was another technique employed by Maller in his earlier test. Cady (1923), in his application of a modified form of the Woodworth Psychoneurotic Inventory to the measurement of juvenile incorrigibility, had earlier made use of repeated items to increase reliability of the scores; although the aim of detecting inconsistency of the "fake" sort was not explicit in his rationale. Each question appeared twice, once in each section of the test, except that in the second appearance the question was phrased in the negative. Theoretically the subject's

response should also be reversed; and the number of failures to reverse is an indication of some inconsistency and hence, Maller assumes of non-cooperation or dishonesty. The "inconsistency score" obtained in this way was to be subtracted from the adjustment score to get a sort of corrected score as proposed by Rosenzweig. It is by no means obvious that the shift to a negative form of item will leave the projective properties of the stimulus simply reversed in meaning; so that the fact of an "inconsistency" in the strict logical sense would not necessarily imply lack of cooperation or dishonesty. However, it would seem reasonable that a very large number of such inconsistent pairs would cast grave suspicion upon the scores, either for dishonesty or some equally serious reason. This technique also was abandoned by Maller in his revised instrument. The second method of using distortion is to present opportunities for answering in a very favorable way but in a way which could almost certainly not be true. This idea was employed by Hartshorne and May in the Character Education Inquiry (1928). Since there are very few aspects of behavior for which one could have complete confidence that no subject would be "ideal" in them, it is necessary to present a considerable number of such opportunities and progressively reduce the probability that any flesh-and-blood individual would be as described. Everyone has at least a few highly desirable traits, and no one has all of them. Without knowing anything whatsoever about a particular person, we can write down on common-sense grounds a list of extremely good and rare human qualities which it is statistically absurd to suppose will all or in large part be his. If he says, however, that he has all (or a very great many) of them, we decide that he is not telling the truth. To practically clinch this argument it is only needful to choose desirable attributes which will very rarely belong, even singly, to anyone; and which furthermore relatively few normal persons claim for themselves when given the chance. In the mass the answers to these items may yield very strong evidence for deception. "I sometimes put off until tomorrow what I ought to do today" can be answered *False* by *very* few honest people. If a subject gives such responses with some considerable frequency, the inference is obvious. A more detailed discussion of this approach will be given in section III below.

The Humm-Wadsworth Temperament Scales and the Minnesota Multiphasic Personality Inventory have both made use of this method, the latter more explicitly. Humm and Wadsworth (1935) deserve credit for having been among the first investigators of structured personality measurement to lay great stress upon the problem of detecting non-cooperation and distortion of response when evaluating a particular profile of scores. They were also among the first to adopt an explicit and uncompromising empiricism in selecting items from a large initial pool. The two scales which serve as "checks" or "correctors" for the remainder of the profile on the Humm-Wadsworth are the "Normal"

component and the "no-count." The Normal component is rather difficult to evaluate from the theoretical point of view, for reasons which have been given elsewhere by one of the present writers (Meehl, 1945b). It is sufficient here to indicate merely its function as described by Humm and Wadsworth, which is to assess the strength of a general inhibiting, controlling, or normalizing factor in personality which serves to act as a "brake" upon strong abnormal tendencies on the other variables. This means that in interpreting a given profile, the significance of any deviation on one of the abnormal components must be established with the size of the Normal score in mind. To the extent that the Normal component measures what the authors claim for it, it is not especially relevant to the present problem; but if it actually operates by detecting something other than the personality component they describe, it would perhaps be of significance here. For a more detailed discussion of this question the reader is referred to the study cited above.

The "no-count" is based upon the number of items to which the subject responds in the negative. Inasmuch as approximately 76 per cent of the scored items (87 per cent of the total pool) of the Humm-Wadsworth are "obviously" suggestive of abnormality when replied to affirmatively, the "no-count" is to some extent a measure of the testee's tendency to avoid, consciously or otherwise, saying "bad" things about himself when taking the test. That this relationship obtains is further supported by the tendency for the no-count to correlate positively (.77) with the "Normal" component and negatively (–.39 to –.72) with the various abnormal components (Humm & Wadsworth, 1935). If the no-count is excessively great, the inference is that the subject has responded in a very defensive or possibly (as in some psychotics) stereotyped fashion; and therefore the particular testing is of doubtful validity. In another article, Humm and Wadsworth state that as high as 25 or 30 per cent of normals seem to invalidate their scores in this way, a proportion which would seem to be impractically high for clinical purposes. In a later article (Humm, Storment, & Iorns, 1939) they attempt to reduce the proportion of useless tests by a "correction" for the no-count based upon multiple regression procedures. Humm and Wadsworth state that in a subsequent group of cases "well known" to them, the improved validity of profiles thus corrected was demonstrated. An unpublished study of hospitalized psychiatric cases by Arnold (1943) indicated that even the exclusion of cases with "invalid" no-count did not result in any greater validity clinically than was obtained using all cases. Humm (personal communication) states that improved multiple regression techniques have resulted in a very marked reduction in the proportion of test misses and of uninterpretable profiles. These more recent data on the Humm-Wadwsorth have not been published. On present evidence it is difficult to say to what extent the use of multiple regression technique was successful in improving validity.

Washburne, in revising his "Test of Social Adjustment" (OSPA), included a set of 21 items modeled after the "lie" items of Hartshorne and May and referred to the total score on this set as *objectivity*. This score was included to detect both lying and unintentional inaccuracy, and the author reports that interviews with people showing very low objectivity scores showed that "it was useless to question them." A very low objectivity score was said to invalidate the test as a whole, and a weighted objectivity score was included in the total score on the entire test (Washburne, 1935).

Another application of the second method for detecting invalidity by identifying the presence of distortion was the "lie" scale (and its complement, F) of the MMPI, which will be discussed in detail in section III below.

The third technique available is the empirical derivation of a "fake" scale by making use of the item shifts obtained when persons take a test under normal "naive" conditions and then are retested with instructions to fake. This method has been used by Ruch to construct an "honesty" key for the Bernreuter. It is interesting that a procedure so logical and straight-forward, invented to solve a problem so obvious and insistent, should have been employed for the first time over twenty years after the appearance of the first personality inventory. Ruch says:

> "The argument is rather simple. If answers to items on a test like the Bernreuter can be faked at all, the chances are that some are easier to fake than others. Therefore, it should be possible to give each item a weight to represent the extent to which it can be faked by the average college student. This was done by tabulating the frequency of each answer to each question for the standard condition and for the influenced condition. These frequencies were converted into percentages, and an 'honesty' weight was assigned to each reply according to the magnitude of the critical ratio of the difference between the frequency of the reply in the honest and in the influenced condition" (Ruch, 1942).

In applying this honesty scale to a new group he was able to show that all cases of "real" introverts would be detected in an attempt to make themselves appear extroverted on the test. There are a number of interesting problems presented by this method, such as the extent to which the key would work if the subjects were not under actual instructions to fake extrovert but were being more "subtle" and actually trying to deceive an examiner in a real life situation. Presumably the deviation toward dishonesty would not be as great under such circumstances. The use of the critical ratio as a basis for weighting items might also be open to some question. In any event, Ruch seems to have been the first investigator to attempt empirical derivation of a fake key for a question-answer personality inventory. The results of applying this procedure to work on MMPI will follow in the present article.

As was mentioned earlier, there is some evidence of a tendency in

the opposite direction in taking personality tests. It is difficult to characterize such a tendency, especially since it may occur on several different bases. A patient in the hospital may for instance engage in a sort of "psychiatric malingering" for strictly conscious reasons, presenting a profile on a test such as MMPI which shows abnormalities out of all reasonable proportion to what is apparent from other considerations. Again, there may be somewhat general traits of verbal pessimism or self-deprecation which, while of some relevance personologically, act so as to systematically distort the results of personality measurement. We shall dichotomize the test-attitude continuum by the two opposed terms "defensiveness" and "plus-getting," not implying anything as to the degree of conscious, deliberate deception involved in either. The corresponding *extremes,* where such deliberate deception seems likely, we shall refer to as "faking good" and "faking bad" respectively. It is recognized that, like the defensive tendency, the "plus-getting" tendency may exist in all degrees from a mild self-criticality or merely objectivity to a deliberate, conscious attempt to make oneself look psychiatrically abnormal. Whether this represents simply the extreme of a continuum with faking good at the opposite end, or an entirely new and different factor, we shall for the moment leave aside. In any case it would be desirable to develop a scale for detecting these tendencies to put oneself in a bad light when answering a personality inventory, so that allowance might be made in such cases in the light of a deviant score obtained on such a scale. The F scale of MMPI was not originally developed with this in mind, but subsequent evidence showed that it could be used in this way (see below). Presumably the two "correction" scales Ch (McKinley & Hathaway, 1940) and Cd (Hathaway & McKinley, 1942) which were found necessary in the early attempts to detect hypochondriasis and symptomatic depression were at least partially dependent upon the operation of such a plus-getting tendency.

A systematic investigation of the plus-getting tendency was attempted by one of the writers, which resulted in the development of a somewhat more generalized correction scale which was called N. The details of derivation and interpretation of this scale are reported elsewhere (Meehl, 1945b) and will not be repeated here. Suffice it to say that from a study of the item responses made by a group of presumably normal persons who showed abnormal MMPI profiles as contrasted with a group of clinically abnormal persons with matched profiles, a group of items was isolated which could be used to roughly quantify the plus-getting tendency. It was found that normal persons who show distinctly abnormal (maladjusted) profiles on the personality scales proper, tended to answer this selected set of *N* items in the "obviously" maladjusted direction, which was with few exceptions also the direction of response given by a minority of the unselected normal population. In other words, a person who is clinically normal in spite of

having an abnormal profile shows a tendency to give statistically uncommon answers which are also "maladjusted" answers in the sense that by inspection they would be considered evidence of psychiatric involvement. For example, about 48 per cent of the unselected general population normals answer "True" to the item "A windstorm terrifies me." Yet we find that among those normals selected specifically for showing apparently *abnormal* profiles on the personality scales proper, about 62 per cent give an affirmative answer to this question. Persons having MMPI profiles no more deviant than these plus-getting normals but who are actually abnormal clinically, give an affirmative answer about 26 per cent of the time. Thus if a person shows an otherwise deviant profile but states that he is terrified by windstorms he stands a better chance of being clinically normal than one who gives the a priori more "normal" or "adjusted" response. Similar items on the N scale include such things as "I am afraid of fire," "I have a fear of water," "People often disappoint me," "I did not like school," and so on. Inspection of these items and an examination of the correlations between N and the other MMPI scales led to a conviction that the N scale was actually detecting a diffuse plus-getting tendency of the sort described. It was further shown that either the inspectional or mechanical use of the N scale in order to under-interpret profiles having the plus-getting tendency led to a reduction in the number of false positives in identification of psychiatric cases. However, the N scale was rather long, and was also apparently loaded with genuine psychiatric factors which led to an undesirable under-interpretation of profiles belonging to grossly abnormal persons. It is therefore to be seen merely as a beginning attempt which was supplanted by K as will be described below.

II. MMPI Scale F

The MMPI variables F and L were not formally validated originally, but were presented on face validity, that is, we assumed their validity on a priori grounds. The F variable was composed of 64 items that were selected primarily because they were answered with a relatively low frequency in either the true or false direction by the main normal group; the scored direction of response is the one which is rarely made by unselected normals. Additionally, the items were chosen to include a variety of content so that it was unlikely that any particular pattern would cause an individual to answer many of the items in the unusual direction. A few examples are: "Everything tastes the same." True. "I believe in law enforcement." False. "I see things, animals, or people around me that others do not see." True. The relative success of this selection of items, with the deliberate intent of forcing the average number of items answered in an unusual direction downward, is illustrated in the fact that the mean score on the 64 items runs between two and four points for all normal groups. The distribution curve is, of course, very skewed positively; and the higher scores approach half the

number of items. In distributions of ordinary persons the frequency of scores drops very rapidly at about seven and is at the two or three per cent level by score twelve. Because of this quick cutting off of the curve the scores seven and twelve were arbitrarily assigned T-scored values of 60 and 70 in the original F table.

From the first it was recognized that F represented several things. Most simply, since the subject would need to sort almost all of the items according to expectation in order for these low scores to result, any error in recording, such as mistaking true items for false items and the like, would raise the F score appreciably. Similarly, if a subject could not understand what he was reading adequately enough to make conventional answers to these items, the F score would obviously be higher. It was felt to be axiomatic that this method would eliminate as invalid records of subjects who could not read and comprehend or who refused to cooperate sufficiently to make expected placements.

In addition, however, it was early discovered that schizoid subjects and subjects who apparently wished to put themselves in a bad light also obtained high scores. The schizoid group obtained high scores because, due to delusional or other aberrant mental states, they said very unusual things in responding to the items and thus obtained high F scores. This is referred to as distortion since we feel that an impartial study would not justify the patient's placements. Among more normal persons some high scores were also observed where the individual had rather unusual ways of responding to conventional stimuli such as are represented by the items involved. For example, to the item, "I have had periods in which I carried on activities without knowing later what I had been doing," most persons answered false. Some persons, however, included periods of sleep in the implication of the item. One might argue that such ways of thinking are often allied to schizoid mentation generally and that the answers in this case indicate a true abnormality. At the very least, however, the person is responding to some items in a way that differs from that of most individuals. Such persons might, therefore, not be appropriately approached through this method of personality measurement. It seems a reasonable enough possibility that there are individuals whose habitual ways of reacting to items are so different from their fellows that measurement of their personalities through the use of verbal items of this type would reflect the unusualness of their reactions to the items more than any clinical abnormality. This semantic factor has been treated more completely elsewhere (Benton, 1935; Eisenberg, 1941; Meehl, 1945b). In so far as such a possibility may exist we have not yet separated it from the clinically more important abnormality expressed in the Sc scale. Parenthetically, one of the most persistent difficulties with developing the Sc scale was this very fact, that an appreciable number of individuals obtained high scores on Sc without being marked by a clinically important degree of abnormality. They, nevertheless, as indicated above,

were responding differently from other people about them as represented by the original data from the general population. It appears that the essential difference clinically is concerned with the particular manifestation of unusual mentation in the individual. If this is not too clearly counter to society's mores, the person may not be thought of as schizoid by those about him though he is often recognized as queer.

Clinical experience suggests that the usual critical score of $T = 70$ is too low in the case of F. We have found that scores ranging up to $T = 80$ (raw score 16) are more often a reflection of "validly" unusual symptoms and attitudes than an indication of invalidity in the rest of the profile due to misunderstanding, etc. Raw scores much above this, however, strongly suggest an invalid record.

With the problems of measurement that developed in the armed forces where a subject might be expected frequently to attempt to put himself in a bad light in answering the MMPI, the F score became especially interesting. It was, of course, immediately possible to consider the F score as an evidence of this attempt, to malinger and obtain fallaciously bad scores on other scales. As a check on this interpretation, a group of 54 ASTP [Army Specialized Training Program] men who had completed a considerable portion of their training in psychology were asked to take the Multiphasic twice.[1] The group took it in the usual way and also took it under instructions to assume that they wished to avoid being accepted in the draft and in order to be rejected they were to obtain adverse scores without giving themselves away. Half the group took it with "fake" instructions first, half second. Through this plan we had a "faked bad" record and a "normal" record. The data showed that 96 per cent of the "faked bad" records had a raw score F of 15 or more ($T \geq 78$), indicating that even these men who were somewhat cognizant of psychological measurements betrayed themselves when they attempted to fake a bad record. The resulting profiles were nearly all unusually abnormal in character, and most of them would have appeared invalid from their pattern alone to one who has seen a large number of abnormal profiles.

From this experiment it appeared that F was a very good device for identifying the intentional faking that could be set up in an experimental situation. It still seemed desirable, however, to attempt to separate among the individuals obtaining relatively high F scores those who were of the above-described schizoid type or who were simply over pessimistic in their view of themselves, from those who for one reason or another faked a bad score or did not understand the items.

III. The L Scale

The original 15 L items making up the L scale were selected under the inspiration of the work of Hartshorne and May. Each of the items

[1] We are indebted to Dr. Howard F. Hunt for administration of these tests.

presents a situation desirable socially, but rarely true of the individual. For example in the item, "I read all the editorials in the newspaper every day," we have a type of behavior that might be generally considered desirable but which is true of only a small percentage of the population. It was recognized that very conscientious persons would frequently have more than average of these L items truly positive but for a person to have six or eight of them seemed almost impossibly good. The 15 items of this type scattered among the main body of the items, constituted a fairly subtle trap for anyone who wanted to give an unusually good impression of himself.

Among the various normal groups the mean score on the L items lies between three and five. As in the case of F the frequency curves are all skewed sharply in the positive direction. Very few individuals obtain raw scores of seven or more, and the two or three per cent level is at about ten. These values were arbitrarily called the 60 and 70 T-score points, respectively. As the L score was used in the clinical setting and as some data began to accumulate from personnel workers in industrial situations, it became apparent that the assumptions regarding the meaning of L were in the main correct, but that there were also other valid interpretations of L, at least in the range from T-score 56 to 70. In fact we found ourselves placing considerable emphasis on T-scores of 56 to 60 which indicated that the original arbitrary assignment of T-scores had been too conservative. On the other hand while the positive presence of the rise in the L score seemed quite valid as an indicator that the individual taking the test was being dishonest and might be somewhat unreliable, if no rise in L was observed, the finding could not be so positively and clearly interpreted. The L score was a trap for the naive subject but easily avoided by more sophisticated subjects.

To check the assumption that L would not identify the more sophisticated subject an experiment was performed with ASTP psychology students. As in the study cited under Section II above, 53 men were given the MMPI twice. The "faked good" data were obtained under the instruction to make certain in taking the test that they would be acceptable to army induction. These records showed no appreciable rise in L. It is also true, however, that the majority of the profiles were only slightly, if any, better than the corresponding non-fake profiles. This experiment would have been improved if persons whose true profiles were abnormal had been used. Some data have been collected from such cases but the number is small. At least, one may conclude that the intent to deceive is not often detectable by L when the subjects are relatively normal and sophisticated.

IV. The K Scale

In summary there were two basic lines of experimental approach to the problem of identifying the attitude a subject takes toward the items

that he is faced with in the personality inventory.[2] Each of these two approaches permits a subdivision into several methods. First, we may have the subject deliberately assume a generally defined attitude, as in the study by Ruch. For example, we may ask him to attempt deliberately to obtain adverse scores while not betraying his intention, and secondly, we may choose records in which there is presumptive likelihood that a special attitude has been assumed. The first approach may be subdivided into those experiments in which the "faking" is directed toward obtaining adverse scores and the approach in which the intention is to obtain desirable scores. In both latter cases an additional set of responses must be obtained relatively simultaneously with the "faked" responses in which the individual assumes his ordinary attitude. The "faked" and "normal" records can then be contrasted for study. One may then make an item analysis to discover the items that are most frequently changed from the "normal" records as contrasted to the "fake" records. Using these "fake" approaches, several scales were derived.

It was found that the items indicating an attempt to obtain a bad record are not necessarily those derived by analysis of records where the subjects attempted to obtain a good record. Our first finding in this regard was that either of these procedures provided a scale that would be about as good for the other type of faking as it was for the one from which it was derived when such scales were applied to test cases not used in the original derivations. It was further found that using two such scales separately did not materially increase the predictive value. As has already been pointed out, it was also found that the original F scale was as effective as was needed to identify those persons who intentionally attempted to obtain a bad score at least within the range of the experiments that we conducted. Conversely, the L scale was not effective nor were any of the specially derived scales especially effective in identifying sophisticated persons who deliberately attempted to obtain better scores. In all of these experiments the findings were so complex and the time devoted to many subprojects was so great that we shall only present data for the final scale K (see below).

In the second line of experimental approach there are also several subdivisions. One may find among presumably functional and normal records those records which are so abnormal as to indicate that the individual should have been in a hospital and attempt to discover the items among these records that will differentiate them from the records of actually abnormal persons. For the counterpart to this approach one

---

[2] Harmon and Wiener (personal communication) have investigated the possibility of detecting defensive and plus-getting tendencies through a division of certain MMPI scales into "subtle" and "obvious" items. Separate T-scores may then be calculated for the subtle and obvious scores on each scale so treated, and in terms of the discrepancy between S and O one may form a judgment as to the strength of the defensive or plus-getting test attitude of the subject. This ingenious technique is still in process of investigation by its inventors and a more adequate treatment of the method and its results will presumably be forthcoming from them later.

chooses cases who were in the hospital but whose records show a normal profile. These may likewise be compared by item analysis to the records of hospital patients with suitably abnormal profiles who would be assumed to have had no interfering test taking attitude. Using this approach we also derived several scales and made many experimental tests of them. Again the details of all of these are not worthy of the complex presentation they would require and these preliminary results will merely be summarized.

The first and most important finding was that whichever of these methods was used, as was the case with the "faked" approach above, the resultant scales were about equally effective and about equally unsatisfactory regardless of the approach and of the particular item content. These scales were also rather effective in differentiating the "fake" group and in some cases were just as valid for that purpose as were the scales derived by that approach. After some two years of this experimentation all of the scales that had showed any promise were reconsidered by applying them to various available groups that had not been used in their derivation and from among them all a single scale which was originally called L6 was chosen as the best. It should be recognized that L6 was not entirely satisfactory but its action in several of the sample situations resulted in its tentative adoption. Although as indicated in the above summary the particular derivation does not seem to play an important part since we could not easily distinguish a scale as having been derived by a special process when we examined its action, nevertheless it may be desirable to tell how L6 was derived. It must not be forgotten that several other scales resulting from the other methods were very nearly as good as was L6, especially the plus-getting scale N. However, when the N scale and L6 were compared and even applied to the test situation set up for the N scale, L6 was a close competitor with N and in several instances was actually better.

In brief, L6 was derived by an item analysis of the responses of 25 males and 25 females in the psychopathic hospital whose profiles showed an L score of $T = 60$ or more and who, with the exception of six normal cases, had diagnoses indicating the probability that they should have had abnormal profiles but whose profiles were in reality within normal range. The diagnoses given to these cases by the psychiatric staff were mostly psychopathic personality, alcoholism and allied descriptive terms indicating behavior disorders rather than neuroses. In general one would expect persons with such diagnoses to be rather more likely to be defensive in taking a personality test than cases of psychoneurosis. There are a few exceptions, however, in the case of hysteria where as has been pointed out in previous papers (McKinley & Hathaway, 1944; Meehl, 1945a,b) there is a tendency for the hysteria to be based upon something closely allied to the assuming of an overly perfect attitude in answering personality items. A particular listing of the diagnoses among these cases is not given here because the

diagnostic categories are not clear enough to be of additional value. In summary, two criteria were employed in the selection of the criterion group. Practically all of them were individuals known to be characterized by deviant behavior but they obtained relatively normal profiles and were thus what we have called misses for the Inventory; and all of these criterion cases were also characterized by having a tendency to obtain elevated scores on the original L scale.

The item responses of these fifty cases handled separately for males and females were compared to the male and female item frequencies from the general group of males and females that has been used in past scale derivations. In all, 22 items were chosen as a result of this comparison. All of these items showed a per cent difference of 30 or more between the criterion cases and the control group, males and females being considered separately.

These 22 items are given below with a T or an F to indicate the direction of the answer that would agree with the tendency of the criterion group. Since the criterion group is assumed to be a group with intent to obtain good scores the larger raw scores on these items are in the same direction as the larger raw scores on the original L scale. The content of these items would seem to suggest an attitude of denying worries, inferiority feelings, and psychiatrically unhealthy symptoms, together with a disposition to see only good in others as well as oneself. The over-all impression is one of "impunitiveness" if we may apply Rosenzweig's term in a situation where no particular frustration is involved.

B-55 I have very few quarrels with members of my family. (T)
C-18 I worry over money and business. (F)
C-27 It makes me impatient to have people ask my advice or otherwise interrupt me when I am working on something important. (F)
C-33 It takes a lot of argument to convince most people of the truth. (F)
D-53 I think nearly anyone would tell a lie to keep out of trouble. (F)
D-54 Most people will use somewhat unfair means to gain profit or an advantage rather than to lose it. (F)
E-43 When in a group of people I have trouble thinking of the right things to talk about. (F)
E-44 I find it hard to make talk when I meet new people. (F)
E-52 People often disappoint me. (F)
F-8 It makes me uncomfortable to put on a stunt a at party even when others are doing the same sort of things. (F)
F-33 Often I can't understand why I have been so cross and grouchy. (F)
F-34 Criticism or scolding hurts me terribly. (F)
F-43 At periods my mind seems to work more slowly than usual. (F)
F-46 I frequently find myself worrying about something. (F)
G-18 I have periods in which I feel unusually cheerful without any special reason. (F)
G-29 I get mad easily and then get over it soon. (F)
G-30 At times my thoughts have raced ahead faster than I could speak them. (F)

G-31 At times I feel like smashing things. (F)
I-22 I have often met people who were supposed to be experts who were no better than I. (F)
I-31 I have sometimes felt that difficulties were piling up so high that I could not overcome them. (F)
I-37 I certainly feel useless at times. (F)
I-38 I often think "I wish I were a child again." (F)

Following the final choice of L6 as the best of the scales available, we subjected it to more careful study and went back through hospital and normal records to find out if it seemed to be of any help in interpreting individual profiles. There were relatively few data on normal cases but on hospital cases a fairly extensive symptomatic summary was available that would permit us to judge whether or not a patient should have had a normal profile. We could then look up the profile and if it was normal we could check to see if the L6 deviated in an upward direction indicating that the patient had attempted to place himself in a good light. As a result of this study L6 appeared effective but left much to be desired.

Since in the summary of scales when L6 was chosen for intensive study, it had seemed about as adequate for the detection of plus-getting as was N or any of the other experimental scales, the records of a new series of presumably normal persons showing deviant profiles was examined and it was again true that L6 appeared to work at the plus-getting end of the test-attitude continuum. That is to say, a relatively low score on L6 could be used to under-interpret an otherwise deviant profile and thus avoid some of the presumably false positives in the normal population sample. Thus L6 seemed useful at "both ends" of the test-attitude continuum, defensiveness and plus-getting.

The most outstanding difficulty in such a procedure was that L6 tended to be low on severe depressive or schizophrenic patient records and thus lead to an under-interpretation in spite of the fact that the patients were very grossly abnormal. To partly correct for this tendency, items were added that would work in the opposite direction. To choose these we studied the item tabulations for the group of ASTP men who had attempted to fake good and bad scores. In this study there were many items which showed no tendency to change with an alteration in the test-taking attitude. That is, the per cent of true or false, as the case might be, remained constant whether the attitude was the normal one or the faked one. From among these items, a sub-group was chosen which showed differences between schizophrenic and depressive criterion groups and general population normals. The procedure rested upon the admittedly somewhat shaky assumption that any item that did not appear to be much affected by the test-taking attitude as approached by a normal person attempting consciously to "fake" good or bad but which did occur as a frequent item to differentiate depressed or schizophrenic patients would be useful in correcting the tendency of our L6 scale to go too low for schizophrenic and depressed patients. Of

course such an item was scored in a way that would make it work against the tendency of the L6 scale. Eight items were selected by this method. The effect of adding these eight items to the 22 on L6 was of course to elevate slightly the mean score of normals and make it more nearly approach the mean score of abnormal cases on the complex of all 30 items. The eight items chosen by this procedure are given below. The letters T and F indicate the response scored in the "lie" direction, and in the direction characteristic of schizophrenic and depressed cases.

A-3 I have never felt better in my life than I do now. (F)
C-28 I find it hard to set aside a task that I have undertaken, even for a short time. (F)
D-48 I think a great many people exaggerate their misfortunes in order to gain the sympathy and help of others. (F)
D-51 I am against giving money to beggars. (F)
F-7 What others think of me does not bother me. (F)
F-20 I like to let people know where I stand on things. (F)
G-23 At times I am all full of energy. (F)
J-51 At times I feel like swearing. (F)

As a final step these eight items were combined with the 22 L6 items into a single scale which we have called K. The K scale represents the final outcome of many experiments in the general field of measuring test attitude. The K scale is far from perfect for its purpose as measured by the various available data. Generally speaking it is about as good as any other single scale derived for any one of the single purposes that have been described. In individual applications it is inferior now to one scale and now to another but the differences are never great enough to be very significant practically and the small number of items in this scale gives it a distinct advantage over one or two of the longer scales such as N. Finally, as was stated above it is not expedient to present more than a single scale although a slight advantage could have been gained if two scales analogous to the original L and F scales had been separately presented.

The construction of K being what it was, odd-even or Kuder-Richardson reliabilities were not computed. Test-retest coefficients were .72 and .74 computed on two groups, one of which was retested at intervals varying from one day to over a year, the other after a lapse of 4–15 months.

Since the K scale was derived as a correction scale or suppressor variable (Horst, 1941; Meehl, 1945c) for improving the discrimination yielded on the already existent personality scales, it was not assumed to be measuring anything which in itself is of psychiatric significance. Actually, its relationship with such clinical variables as the subtle Hy items (see below) might suggest an interpretation of K alone; further, it is presumably a significant fact about a person that, in answering a personality inventory, he tends to behave as a "liar" or a "plus-getter." However, the real function of K is intended to be the correction of the

other scores; and validity will be discussed with reference to this function only.

It is first necessary to choose criterion cases of the sort on which K can conceivably be of value. It is clear that such cases will be characterized by the presence of what may be called *borderline* profiles, i.e., those showing T-scores, say, between 65 and 80. The reason for this is that in studying hundreds of deviant profiles after the addition of K, almost no individuals were found with T-scores above 80 in the normal sample, and it was not statistically profitable to correct elevations of such magnitude to the point of calling them normal. On the other hand, when a curve shows no elevations at all above 65, even the presence of a high K score does not enable the clinician to form any adequate notion of what the peak would be, if any, had the K-factor not been operating to distort the results. In other words, there are upper and lower limits beyond which deviations on K cannot effectively operate. Profiles showing scores above 80 are to be interpreted as probably "abnormal" no matter how low K falls; while if a profile shows no scores above 65 we cannot tell whether a high K means the profile should be adjusted toward more severe scores or is merely that of an actually normal person who for some reason or other took a defensive attitude when being tested. The kind of curve which gives interpretative difficulty and which could conceivably be improved by knowledge of the influence of K would be a curve in the doubtful, borderline region. Accordingly, a group of cases from the normal and hospital groups was chosen on the basis of having achieved such borderline curves. We selected for this study all cases in the files showing at least one personality component[3] elevated as high as $T = 65$, but no component elevated to $T > 80$. Among the normals, there were 174 having such borderline curves, of which 71 were males and 103 were females. Corresponding to these cases, we located among our clinically abnormal cases 129 males and 208 females with similar borderline profiles. The data for the two sexes were treated separately.

The analysis of these data was in terms of the ability of the K scale, used mechanically as will be described, to separate the curves of the actual normals from those of the actual abnormals. For each sex group, the procedure was to arrange the whole set (normals and abnormals combined) in order of the magnitude of their K scores. The distribution of K was cut on the basis of the proportion of normals and abnormals in the sample, calling all cases above the cut "abnormal" and all those below "normal." Setting up a fourfold table on this basis, a chi-square of 20.436 for the males and 29.540 for the females was obtained. Both of these are highly significant ($P < .001$) with 1 d.f. If, instead of locating an optimal cutting score the K distribution was cut at the mean of the general population K distribution (i.e., at $T = 50$ regardless of the present samples) the cutting point of the males is unchanged, whereas

[3] Mf is excluded from consideration here and in all that follows.

that for the females shifts enough to lower their chi-square to 17.750, which is still highly significant. In other words, if one considers miscellaneous profiles which lie in the borderline range between 65 and 80, regardless of the kind of elevation and irrespective of the clinical diagnosis of those who are clinically abnormal, he can separate them into "actual" normals and abnormals significantly better than chance by using a cutting score on K. It must be emphasized again that K in this instance is operating chiefly as a suppressor of certain test-taking tendencies, since K by itself does not practically differentiate unselected normal and abnormal cases (1 to 2½ raw score points difference between means for various samples). In terms of percentages, it was found that for the males, 72 per cent of the abnormals and 61 percent of the actual normals were correctly identified. For the females, 66 per cent of the abnormals were identified as such and 59 per cent of the normals were so classified. These percentages are based upon the separations at a K = 50, taking, therefore, no account of the actual normal-abnormal proportions among the present cases.

Evidence from examination of the test misses spotted by K in the above data combined with our knowledge of the correlation between K and other MMPI scales, indicated that the K correction was more important in the case of some scales than of others. Therefore, it was decided to analyze the borderline groups in terms of the peak elevation of their profiles, in the attempt to identify those particular curves on which K could be used with profit.

The entire group of 511 borderline curves (males and females, normals and abnormals pooled) was divided into eight sub-groups, each sub-group being composed of cases having the peak score on the same one of the eight personality components. Thus, there were 60 curves having the peak on Hs, 91 on D, 119 on Hy, 66 on Pd, 38 on Pa, 25 on Pt, 28 on Sc, and 52 on Ma. (The difference between this total of 479 cases and the 511 used in getting the over-all chi-square is due to the exclusion of 32 profiles on which no "peak" could be fairly assigned, since two or more of the components showed identical T-scores and these were the highest on the given curve.)

The normals and abnormals having borderline curves with the same peak score were than separated mechanically by the use of a cutting score on K, the proportion of cases above the cutting score being determined on the basis of the proportion of actual abnormals versus normals in each sub-group. This was unavoidable in the present analysis because the relative proportions of actual normals and abnormals varied widely from scale to scale and the use of the mean of K would have been grossly misleading since in some instances the proportions were extremely asymmetrical (Zubin, 1934). For the eight groups studied in this manner, only three showed a significant chi-square ($P < .01$), namely those having peaks on Hs, Pd, and Sc. The Ma group yielded a chi-square between the 10 per cent and 20 per cent level of

significance. On D, Hy, Pa, and Pt the chi-squares were all below the 20 per cent level of significance; and the pooled chi-square for these five scales (5 d.f.) gave a $P > .22$. It would seem, therefore, that the K-factor may be used with profit in interpreting some kinds of profiles but not others. Of course, the failure to discriminate with K when grouping profiles by peak score does not establish that a K-correction might not be profitably added to the single scores themselves. This problem will be treated at length in a sequel to the present paper.

One other validating study was done on K. In this instance, we made use of a group of 22 normals and 22 abnormals employed in a previous study (Meehl, 1945b). The normals in this set consisted of a random selection from a large group of profiles showing any elevation of 70 or over (excluding Mf). The abnormals consisted of a heterogeneous group also having at least one score over 70, and included seven psychoneurotics, seven schizophrenics, three psychopaths, two alcoholics, two manic-depressive (depressed), and one paranoid state, chosen randomly from recent hospital cases. These groups had been selected for a different purpose and had not entered into the derivation of K in any way. They can also be considered, therefore, a fair test group for validation purposes. Without regard for any other information concerning the profiles, all cases showing $K > 50$ were arbitrarily guessed as abnormals, whereas those with $K < 50$ were called normals. The cutting score was therefore also independent of the statistics of the present group. Here the K scale worked phenomenally well, being much better than the N-scale (which was derived on cases some of which were included in this blind diagnosis study). Of the entire group of 44 cases, 37 were correctly classified when using K in this way, a total of 85 per cent hits. It will be recalled that we are here trying to separate normals and abnormals all of whom have deviant profiles, so that this per cent is quite impressive considering the task set for K. Of the seven errors in classifying, six are "false positives," i.e., cases of normals showing elevated profiles and $K > 50$, called therefore abnormal. The chi-square for the fourfold table of these data is 21.569 which with 1 d.f. is highly significant ($P < .001$). This corresponds to a contingency coefficient of .57. Here we have striking evidence of the validity of K when used to differentiate between deviant curves of actual normals and abnormals. We are not prepared to explain the superiority of this result to that given by the analysis previously discussed, except to say that the range of abnormal scores in the present analysis was from 70 to 90 whereas in the previous analysis we used "borderline" scores defined as lying between 65 and 80. In what way this could make K appear to function more effectively in the one case than the other is not clear. Also the present study involved only males, where K in general seems to work a little better than on females.

The fact that K is less effective as applied to some scales than others would suggest separate interpretations or cutting scores depending

upon the kind of profile with which one is confronted. Furthermore, the rough classification into "normal" and "abnormal" on the basis of a single arbitrary cutting score obviously sacrifices some quantitative information about the actual magnitude of the personality scale elevations with respect to the magnitude of the K score. We do not intend to propose such a rough cutting method as the most efficient manner of application for K, but are using that form here simply to indicate that K has differentiating power for what it was hoped to differentiate. The optimal mathematical procedure in using K as a suppressor involves complex issues which we shall have to reserve for a later publication.

V. Relation of K to Other Test Variables

The correlation of the K scale with other MMPI variables should throw some light upon the question of its differential efficiency on these scales, as well as give us some insight into its psychological nature. Table 1 below shows the intercorrelations of K with the other personality components measured by MMPI. These correlations are based upon 100 cases in each of the four groups indicated, chronological ages 26–45, excluding records having "?" > 70 or F > 80.

Table 1

Intercorrelations of K with Other MMPI Variables

	Hs	D	Hy	Pd	Pa	Pt	Sc	Ma	Mf
Normal males	–.30	.15	.48	–.17	–.07	–.67	–.59	–.36	
Normal females	–.35	–.03	.30	–.06	–.02	–.64	–.58	–.28	
Male abnormals	–.42	–.29	.11	–.26	–.19	–.60	–.60	–.37	–.08
Female abnormals	–.17	–.16	.17	–.21	–.13	–.63	–.58	–.38	.04

Of interest in this table are the following facts. With the exception of Hy and one of the four coefficients of D, the correlations are consistently negative. This is of course to be expected if K represents the defensive, lying, or self-deceptive test-taking attitude it was derived to measure. The negative correlations with Hs combined with the positive correlation with Hy indicate that there must be a fairly high positive correlation between K and those non-somatic items on Hy which have been previously referred to—the "zero" items on Hy or what Harmon and Wiener have called "hy-subtle" (henceforth designated Hy-O).[4] Since this latter set of items, although derived by its empirical separation of clinical hysterias from normals, seems to reflect the self-deceptive and impunitive attitude of the hysterical temperament, it is

[4] These items are called "zero" items because on the scoring templates they are indicated with a letter "O," meaning that one receives a point for the "abnormality" by responding in the direction which, on that single item, characterizes the majority of general normals. This means that the abnormals in question tend to give the "normal" response much more often than the normals do.

consonant with our interpretation of K that it should be markedly correlated with Hy-O. The direct evidence on this point will be reported below. The only correlations of very impressive magnitude which appear in this table are those with Pt and Sc. Here they are high negative—the person who makes responses characteristic of compulsive and schizoid persons has the opposite of the self-deceptive and defensive attitude. In other words, he tends to be a "plus-getter" and in this way is distinctly unlike the hysteric. These correlations are also in harmony with our clinical knowledge of the components in question, especially in the case of the psychasthenia. The Pt scale has never been considered very satisfactory, and it has been shown in unpublished studies that Pt can actually be used as a correction scale in the way in which N was used. It is perhaps significant that of all the MMPI scales, Pt is the only one for which, lacking a sufficiently large criterion group, methods of internal consistency were employed in the item selection. Here again we would expect to get a greater operation of non-clinical test-taking factors of the K variety.

It might be thought that such low correlations as occur in the table above would preclude any possibility of the use of K as a suppressor. There is a tendency for the scales on which K seems "valid" by the chi-square test to show the higher correlations, with the exception of Pt. It will be shown in a subsequent paper that, for the use to which K is put, correlations as low as .20 can be utilized to yield very significant and useful improvements in discrimination.

At this point we may briefly review some of the previously developed scales which are now known to be saturated with what we may call the *K-factor,* since their diverse sources and methods of derivation furnish additional strong evidence for our theoretical interpretation of K. Two of these scales have never been published, so that their derivation and properties must be briefly summarized here. About three years before research on the test-taking attitude was begun, Hathaway and W. K. Estes, using a variant of the method of internal consistency, developed a scale called G. This scale is the only MMPI scale which was derived without the use of any kind of criterion external to the test; like those personality tests being developed by factor analytic methods at the present time, the selection and scoring of items was based wholly upon the intercorrelations among the items themselves. Essentially, the procedure consisted in locating among a group of 101 unselected normals those individuals who, when their answer sheets were used as scoring keys, produced the maximum variance of the other 100 scores. The assumption was that these persons were the most extreme deviates on whatever factor or factors contributed most heavily to the variance and covariance of the total pool of MMPI items. From the evidence adduced by Mosier (1936), it is of course clear that the "purity" or factorial unity of this hypothetical underlying continuum is by no means guaranteed by such a procedure.

Another way of looking at this procedure is to consider the fact that one maximizes the variance of a set of items by scoring them in such a direction as to maximize their mean covariance—since the item variances are unaffected by the direction of scoring. Instead of actually calculating the variances for the 2^{550} ways of scoring the test, we select *individuals* who approximate the optimal scoring key. It was found that the scoring keys for some 10 individuals selected by this method tended to form two distinct clusters, each of which consisted of keys (individuals) showing high correlations with one another and high negative correlations with the members of the other cluster. An item analysis was then carried out on these two small groups, and the items resulting were combined into a scale called G (general factor).

The G scale had a number of interesting properties which were not interpretable at the time of its derivation. It showed a very large variability, both in absolute terms and as indicated by a coefficient of variation. The scores *among normals* ranged from those who answered none of the items in the scored direction, to those who answered all but eight of the 62 items in the scored direction—a phenomenon unheard of in the other MMPI scales. The odd-even reliability of G was about .93, which is considerably higher than the coefficients we typically find in the MMPI scales. The item content was that of the typical "neurotic" or "maladjustment" sort which predominates on *a priori* scales such as the Thurstone or Bernreuter Bl-N. Examples of items are: "When in a group of people I have trouble thinking of the right things to talk about" (T); "I cry easily" (T); "I am certainly lacking in self confidence" (T). It is perhaps significant that the most powerful single item in the internal consistency sense—which happens in the sample studies to have a correlation of 1.00 with the entire G-scale—is almost a distilled essence or prototype of so-called "neurotic schedule" items: "I am easily embarrassed" (T). The G scale, although derived without recourse to any clinical group whatever, nevertheless showed a correlation of .91 with Pt. The mean MMPI curves for unselected normals with high G (the "neurotic" end) showed elevations on F, Hs, D, Pd, Pa, Pt, Sc, and Ma, especially on Pt and Sc; whereas L (raw score) and Hy tended to fall below the mean. The mean profile for normals with low G was almost an exact mirror image of this curve. However, G was not found to be very effective in the detection of any clinical group or to be particularly useful for any purpose; and since at that time no theoretical basis was available for interpreting it, the scale was abandoned. Another scale, called + ("plus"), was derived in a similar but not identical manner.

In the derivation of the original hypochondriasis key, there was developed a correction scale called Ch, the function of which was to separate actual clinical hypochondriacs from a group of non-hypochondriacal abnormals (mostly schizophrenic and depressed) who attained spuriously elevated scores on H. The item content of this Ch key was

quite puzzling, because although the correction was successful, the items did not seem to refer to anything either hypochondriacal or anti-hypochondriacal. In fact it was difficult to see what psychological homogeneity, if any, they possessed. For a more detailed description of this scale (now no longer in use since the appearance of the modified Hs key) the reader is referred to the original article (McKinley & Hathaway, 1940). For present purposes it is merely necessary to state that the great majority of the items on Ch were scored if answered in the statistically rare and obviously "maladjusted" direction and that they apparently measured some non-somatic component of test responses which resulted in spuriously elevated H scores in persons who were not actually hypochondriacal.

Still another scale of the same general sort was derived by Meehl and called N. To briefly repeat what has been said above, this scale differentiated normals showing elevated profiles from clinical abnormals showing no greater profile elevations, and was interpreted as detecting a plus-getting test attitude for which scores on the personality components proper should be corrected. The type of item occurring on the scale N has been discussed above.

Lastly, we recall to mind the Hy-O items which have been described above as reflecting this kind of component, although scored in the opposite direction from N, Ch, and G.

It is of considerable interest to examine the correlations between K and these other variables, derived in their diverse ways. Table 2 presents the correlations between K and the various scales thought to be loaded with the factor in question, based upon scores of 100 individuals ages 26–45 in each of the groups indicated.

Table 2

Correlations of K Scale with Other Variables Thought to be Loaded with the "K-factor"

	+	G	N	Ch	Hy-O
Normal males	–.64	–.76	–.70	–.67	.81
Normal females	–.62	–.73	–.64	–.63	.78
Male abnormals	–.70	–.75	–.69	–.64	.74
Female abnormals	–.70	–.81	–.72	–.71	.74

Considering the relative unreliability of some of these variables, the above is a very impressive group of intercorrelations. We have two scales (G and +) which were derived wholly by internal item relationships and without regard to criteria of any non-test behavior; a scale (N) which corrects for the self-criticality of certain plus-getters who show deviant profiles; a scale (Ch) which differentiates hypochondriacs from non-hypochondriacal abnormals who have elevated H scores; and a subset of items (Hy-O) which were chosen because they differentiate a clinical group—hysteria. There is, however, a considerable item over-

lap among these scales, tending to raise these correlations. On the other hand, it will be recalled that the scale K is not actually "pure" for the hypothetical test-taking attitude because it is a composite of the test-taking scale L6 plus the eight "psychotic" items. This would presumably tend to lower the correlations. Accordingly, we have substituted L6 for K, removed the item overlap among the scales G, N, Ch, L6 and Hy-O, and calculated correlations among these reduced keys. Table 3 shows the intercorrelations among these five non-overlapping keys, based upon the responses of 150 unselected normal males between the ages of 26 and 45, rejecting records with ? > 70 or F > 80. All scales were scored so as to render the correlations positive.

Table 3

Intercorrelations of Five Scales Thought to be Loaded with the Test-taking Attitude, No Item Overlap. $N = 150$ Normal Males

	G	Ch	L6	N
Ch	.82			
L6	.76	.71		
N	.78	.73	.66	
Hy-O	.70	.63	.70	.59

This correlational matrix has been subjected to a factor analysis, repeated three times in successively approximating the communalities because of the small number of tests. The first factor extracted leaves no residuals larger than .049, and the SD of the residuals is .032, which is less than the SE of .041 attached to the mean r in the matrix. Testing the significance of the residuals by the formula chi-square = $\Sigma(z_0 - z)^2 (n - 3)$ (Burt, 1941, p. 339) the chi-square on the deviation of observed r's from those predicted with the first factor loading was not significant (chi-square = 5.101, 5 d.f., $P > .30$). It appears that one common factor is quite sufficient to account for the intercorrelations of these scales. The factor loadings of the scales G, Ch, L6, N, and Hy-O are .927, .868, .847, .818, and .770 respectively. It is interesting to find such a powerful factor running through scales derived by such diverse methods. It is also worth noticing that the largest loading of the K-factor is in the one scale constructed wholly by "internal consistency" methods, whereas the smallest loading is that of the clinical variable Hy-O. If we extract a second factor just to see what it looks like, none of the loadings is over .20 and the meaning of the second factor would be quite uninterpretable on our data. Although we have been thinking in terms of a "K-factor" on the basis of the apparent community of practical function shown by these various scales, it is reassuring to find that the term "factor" may be used here without doing violence to the more technical meaning of that term as used by factor analysts.

Considering the nature of the items which are involved in scales such as L6, N, and G, this finding perhaps sheds some light on the relative inadequacy of "neurotic" inventories such as the Bl-N when applied to clinically diagnosed neurotics. Here we have a kind of item which, while it does not (in its own right) appear to discriminate normal from abnormal individuals very successfully, does reflect some kind of a test-attitude or self-critical component. Those "neurotic" persons who happen to be characterized by this particular manifestation of self-criticism, such as certain compulsives, will probably be differentiated by such a set of items. On the other hand, other equally "neurotic" persons such as hysterics, who are characterized by the opposite attitude, will not be successfully spotted by the scale. If anything, they should be discriminated backward! Furthermore, the central tendency of abnormals in general is the same as that of normals, and it is quite possible that in developing personality questionnaires set up in the traditional, *a priori* fashion and "refined" by statistical manipulation we are merely setting up sets of items to differentiate among people with respect to various test-attitude continua of little or no psychiatric relevance. It will be recalled that the scale G consisted of items having the heaviest loading with whatever factor (or factors) contribute most to the variance and covariance of the entire 550 items in the MMPI pool. Yet this scale turns out to have little or no clinical value (*except* as a suppressor) and to be the scale most saturated with respect to the test-taking attitude. We feel that psychologists have tended to forget the fact that when one constructs a personality inventory by studying the item-associations, whether by old-fashioned methods of internal consistency or by factor analysis of item correlations, he is merely locating certain covariations in verbal behavior. When a final scale based upon that kind of derivation is presented to the clinician, all that the clinician can be assured of is that *persons who say certain things about themselves also have a tendency to say certain other things about themselves*.

Willoughby's argument (1935) that the non-chance covariation of item responses establishes "validity" with respect to *some* underlying, common trait which gives rise to the covariation may be admitted without contradicting what we have just said. That items should exhibit consistency in this covariant sense in spite of not being valid for the traits sought, or in fact even being negatively valid, has been shown by many studies, most particularly those of Landis and his associates (Landis & Katz, 1934; Landis, Zubin, & Katz, 1935; Page, Landis, & Katz, 1934). The "underlying disposition" which leads a subject to respond in a certain way to such questions may or may not be identical with the dispositions we recognize as clinical variables, nor with those that might be suggested by the item content. It is quite clear on present evidence that this identification cannot be established by an assumed equivalence between non-test behavior and the verbal report. Hence, as has been repeatedly stressed by the present writers, both *a priori*

selection of items and the psychological naming of a statistically homogeneous scale from its item content are fraught with possibilities of error.

An obvious line of investigation which is suggested by these considerations is the systematic study of the relationships which exist among variables such as K, G, and N which are fairly definitely known to be chiefly test-taking variables, and other personality scales which have been developed by variants of the method of internal consistency. Because of the influence of socio-economic or educational level upon the K-factor (see Section VI below) such studies should ideally be carried out upon subjects from the general population. At present, we can only report a few preliminary studies which seem to have some bearing upon this question. All of these studies happen to be concerned with the batteries developed by Guilford and Martin (GAMIN, STDCR, and the Personnel Inventory). We wish to emphasize that the presentation of these scattered data on our part is intended simply to raise some questions concerning the construction of scales by internal consistency methods where factors such as K are probably in operation; the validity of the Guilford-Martin scales must of course be assessed upon other grounds. We wish further to stress that in comparing these tests with MMPI we do not intend to set the latter up as a "criterion," although it does of course have the advantage that each item is known to differentiate certain defined criterion groups which literally define the scales on which the item occurs. It should also be made clear that Guilford, as one of the foremost contributors to the factor analytic approach to personality test construction, has explicitly called attention to the importance of the problem of test-taking attitudes as "factors," when he says,

> "We must constantly remember that the response of a subject may not represent exactly what the question implies in its most obvious meaning. Subjects respond to a question as at the moment they think they are, with perhaps a lack of insight in many cases as to their real position on the question. They also respond as they would like themselves to be and as they would like others to think them to be and as they wish the examiner to think them to be. They also respond with some regard to self-consistency among their own answers. Whether these determining factors are sufficiently constant to set up individual differences which are uniform in character and so constitute common factors in themselves is difficult to say. Should any one of them be so pervasive it should introduce an additional vector in the factor analysis" (Guilford & Guilford, 1936, p. 118).

It is our opinion that the data we have presented indicate that the answer to Guilford's question is in the affirmative, and that the inclusion of a few K-type scales in a factor analysis would probably result in a somewhat different interpretation of the other tests and factors than would otherwise be the case.

Wesley (1945) has studied the relationships existing between the Guilford-Martin Personnel Inventory of traits O-Ag-Co and the MMPI

scales, based upon the test records of 110 presumably normal college women. The three traits measured by the Personnel Inventory are called *objectivity, agreeableness,* and *cooperativeness* by their authors. High scores are in the direction of the traits named, and low scores indicate the presence of what is called in composite the "paranoid" personality. Wesley found that the composite Personnel Inventory score correlated only .11 with the MMPI Pa scale which, while still in a preliminary stage, does consist of items which are empirically known to distinguish clearly paranoid groups of persons from people in general. Together with this rather disconcerting finding, she also discovered that the "paranoid" score on the Personnel Inventory correlated .50 and .57 with the MMPI scales Pt and Sc—both of which are relatively weak scales from the standpoint of clinical differentiation but are known to be heavily loaded with the K-factor. The correlations of "objectivity" with Pt and Sc were both –.62, which led her to correlate Trait O with the correction scale N, leading to the same figure. None of the other correlations of the Guilford scales with MMPI scales exceeded .45, and the majority of them were under .20. The mean MMPI profile of subjects selected on the basis of having low raw scores on N (the "defensive" end) showed a pattern hardly distinguishable from that of subjects selected for having high scores on Factor O. It is interesting to note in passing that of the seven items of very similar wording which occur on both the Guilford-Martin Inventory and the MMPI Pa scale, five are scored as "paranoid" in the opposite direction on the two scales. For example, to say that most people inwardly dislike putting themselves out to help others, that most people would tell a lie to get ahead, that some people are so bossy and domineering that one feels like doing the opposite of what they tell him to do, are responses scored as paranoid on the Guilford-Martin; whereas it is found empirically that these verbal reactions are actually significantly *less* common among clinically paranoid persons than they are among people generally. This kind of finding suggests that paranoid deviates are characterized by a tendency to give two sorts of responses, one of which is obviously paranoid, the other "obviously" not. But these two sorts of responses are negatively correlated among people generally, and hence appear scored oppositely on scales developed by internal consistency methods.

It is of course possible to begin the development of scales by internal consistency or item-intercorrelation procedures, and having built a scale by these methods, to apply it to various criterion groups for validation. But it would seem that if the aim is to find items which will optimally perform such a discriminating function, the most direct route to that goal is immediate empirical item selection from the start. It may be agreed that scales developed through item-correlation techniques have more statistical "purity" and hence are in a certain special sense better for what they *do* measure. One's attitude toward this problem is likely to reflect his more fundamental views as to the nature of a so-

called "measurement" in personality testing, complete discussion of which would take us beyond the present paper. It seems clear that the results of factor analysis to date have not, whatever their theoretical validity, made possible the construction of single personality items which can be called even approximately "pure." For example, in Guilford's factor analysis of 89 personality items originally chosen (on the basis of suggestions from a previous factor analysis) to sample seclusiveness, thinking introversion and rhathymia, after the extraction of nine different factors the majority of the items still showed communalities less than .50. Torrens (1944), Wesley (1945), and Loth (1945) all found that the typical scale intercorrelation among the variables of the Guilford-Martin batteries STDCR, GAMIN, and the Personnel Inventory is actually higher than the typical intercorrelations of scales on MMPI which were developed with almost no consideration for questions of scale purity or freedom from item overlap.

Louis Wesley (personal communication) has suggested that the contrast between the two methods of scale derivation is between *maximal measurement* and *meaningful measurement.* By this is meant that internal consistency methods lead to scales which measure whatever they measure with high consistency, large variance, great discrimination. This is "maximal" measurement. It is suggested that the most important non-test behaviors, which it is the aim of the test to predict, may not be associated with the same variables which lead to the kind of consistency involved. We may, as in the case of the Pa scale, have to sacrifice the desire to have high item intercorrelations in order to score items so as to achieve the more fundamental aim of criterion discrimination. Since scales are so very "impure" at best, there does not seem to be any very cogent reason for sacrificing anything in pursuit of the rather illusory purity involved.

There are multiple determiners which enter into a subject's decision when he answers a personality item. One might say that all but a very few personality items have an inherently "multiphasic" character, exceptions being such items as "I am a male." Obviously, if there existed or could be invented verbal items which were even approximately pure, the "scales" of such items could be extremely short and in fact the practical value of substituting an inventory for a few brief oral questions would be much in doubt. But the items are not uniquely determined. This simple behavioral fact imposes certain limitations upon the progress of personality measurement, as has been pointed out by many critics. From the common sense point of view, the situation is not very different from what occurs in medical diagnosis or in the psychiatric interview. Almost all of the symptoms or responses which are in evidence are known to arise upon diverse bases. During a psychological interview, a woman may miscall her husband by the name of a former suitor, a phenomenon which is in itself ambiguous; perhaps she has recently seen the man in question, perhaps she has been reading a

novel in which that name appears, and perhaps—the psychiatrically significant possibility—she feels somewhat regretful for not having married him instead. Later, we find that she developed a headache on her wedding anniversary, also an ambiguous datum if it stands alone. Again, she is excessively effusive about how happy her married life is, and so on. It is through the hypothesis of marital dissatisfaction that these different behaviors find a common explanation. When we accumulate such single items about her behavior, we are merely piling up the probabilities. It seems a little foolish to locate these behavior particles or their "sum" on a continuum of measurement, except in the most crude ordinal and probability sense. It is further quite likely that important configurational properties are also involved here, so that the significance to be assigned to one of these single facts should be a function of the other facts we know. The traditional scoring procedure of simply counting *how many* responses belonging to a certain class have been made seems to be very crude; fortunately it has been repeatedly found that the various weightings, compositions, and non-linear refinements which the behavioristic logic might suggest do not usually make sufficient practical difference in the ordering and sorting of people to be worth doing. The fact that we find it convenient to treat these behaviors in certain mathematical ways (independent scoring, unit weights, summation, linear transformations, etc.) should not mislead us into supposing that we are doing anything very close to what the physicist does when he cumulates centimeters. From this point of view, methods aimed at either "purity" or "internal consistency" are not easy to justify. At the very best, we have a rather heterogeneous collection of verbal responses which have a rough tendency to covary in strength. It may or may not be true that the most important (powerful) determiners of this tendency to covary are clinically relevant or personologically significant. For example, disliking one's husband is not the most powerful "factor" in determining the frequency of headaches, among people generally. Nor is it the most potent factor in determining whether one calls him by the wrong name. Furthermore, the tendency to do these two things may not be covariant at all among people in general. None of these reasons, however, would lead us to reject the two facts in trying to evaluate the hypothesis of marital unhappiness.

From both the logical and statistical points of view, the best set of behavior data from which to predict a criterion is the set of data which are among themselves not correlated. This is well known and made use of in the combination of scales into batteries; but for some reason psychologists are uncomfortable if the same reasoning is applied within scales. The statistical considerations are of course quite general, applying as well to items as to scales. It is likely that the insistence upon high internal consistency and "item validity" in the item-test correlation sense springs in part from a feeling that all of the items ought to be

"doing the same thing." This certainly sounds like a reasonable demand as it stands, but it requires clarification. As is clear from the factor analysis studies, one simply cannot find any appreciable number of non-identical verbal items which all "do the same thing." Every one of them depends upon many things, and the item as a unit is like the old-fashioned atom—uncuttable and hence permanently impure. Items "do the same thing" when they are so combined in pools that it is very unlikely that the subject will answer many of them in the scored direction unless he is characterized by a certain strength or range of non-test behaviors which in turn depend upon the one (or few) "variables" that are common to the items. It may still (unfortunately) be the case that the heaviest contribution to each item consists of variables other than the ones we are interested in. That this is in fact true is indicated by the typical values of item communalities.

It is this state of affairs which we believe imposes limitations upon the efficiency of such suppressor scales as K. Since we cannot find items which depend upon only clinical abnormality, we try to find items which depend upon abnormality to an appreciable extent even though they unavoidably depend upon other things as well. The suppressor consists of items which unavoidably depend to some slight degree upon clinical abnormality, but to a greater extent upon the objectionable factors in the first set. By cumulating responses to the second set of items, we hope to get an indication of the strength of these other factors, which information is then used to correct for their undesired contribution to a score attained on the first. The impurity of the suppressor itself, however, sets limits to the efficiency of such a process. Thus, a subject may obtain a high depression score because he is a plus-getter. The strength of his plus-getting tendency is assessed by items such as those of K. However, a sufficiently great degree of depression will yield considerable deviations on K, since the K items themselves are not pure for the plus-getting tendency but are also slightly loaded with clinical abnormality. In such cases K operates against us. It is interesting to note that the K scale, itself a suppressor, also *contains* a suppressor in the form of the eight "psychotic" items—but here also the effort to suppress the unwanted components of the suppressor can only be imperfectly carried out. No refinements of statistical technique enable us to escape the basic psychological fact that our smallest behavior units, the responses made to single items, are inherently of this multiphasic character.

VI. Relation of K to Age, Intelligence, and Socio-Economic Status

In the study of the correction scale N it had been observed that college students (actually, high school graduates tested at the University Counseling Bureau prior to actual matriculation) showed a distinct elevation in the "lie" direction, averaging about one sigma above the general population mean. It was also found that the younger

age group (16–25) showed a similar although smaller deviation, which was accounted for by the presence of a considerable number of medical students in that group. Furthermore, college graduates who had been some ten years out of college showed a mean T-score of about 60 on the N-scale. A similar trend is discernible in the case of K. The mean T-score of a group of 84 medical students is at 62, a deviation which is significant at the 1 per cent level. Both male and female pre-college cases average a T of 57 on K. This tendency falls in line with the fact that the mean MMPI curve for several college and pre-college groups, including some obtained elsewhere than at Minnesota, is a curve with a slight but consistent elevation on Hy, in spite of having an Hs below the mean. This indicates, as usual, a tendency to respond in the hysteroid fashion which elevates Hy-subtle enough to more than counteract the tendency to answer the somatic items on Hy in a non-hypochondriacal fashion. We are not prepared on present evidence to give an interpretation of this phenomenon. That it is not primarily a reflection of intelligence differences is suggested by a correlation of only .04 between K and ACE score among the pre-college cases, which, even taking their relative homogeneity into account, should be higher if intellect as such is the reason for the difference. If the factor at work here is not intelligence, nor the mere fact of being in college when tested, two other possibilities are socio-economic status and chronological age. A group of W.P.A. workers in the young age group 16–25 showed no elevation on K whatsoever, which would favor the socio-economic interpretation. The mean K of a group of 50 normals aged 16–25, excluding college graduates and persons in college, was 13.5 (T = 52). These figures would seem to eliminate mere chronological age as the chief basis of differentiation. We are left with socio-economic status as the most plausible remaining variable. What is needed is study of a group of persons in the upper socio-economic group who are not college students and have never been college educated. Unfortunately, we do not have a large enough sample of such persons to enable us to draw conclusions with certainty. The mean raw score on K for a group of 18 normal subjects classified in Groups I and II in the Goodenough classification, who were not, however, college graduates or attending college, was 18.50, which corresponds to a T of 61. In spite of the small N, this difference is great enough so that a t comparing their mean with that of 156 un-selected normals from the other economic classes was highly significant ($t = 6.055$, $P < .01$). It seems plausible that the college, pre-college and college-educated elevation is reflecting chiefly a difference in socio-economic status, although further evidence on this topic should be collected. If this is confirmed by subsequent investigation, it will be interesting to speculate upon the possible ways in which membership in the upper classes generates the particular kind of defensiveness involved.

VII. Summary and Conclusions

The general problem of test-taking attitudes in their effect upon scores obtained on structured personality inventories is discussed. The literature on the subject is briefly surveyed, and a discussion given of the various approaches which have been taken in an effort to solve this problem. The final result of many efforts to derive special scales for measuring various attitudes in the taking of the Minnesota Multiphasic Inventory is presented, with some indication of its validity. The relationship of this scale, called K, to other variables is used as a basis for discussing certain general problems in the theory of personality measurement. Conclusions are as follows:

1. The conscious or unconscious tendency of subjects to present a certain picture of themselves in taking a personality inventory has a considerable influence upon their scores.
2. We may distinguish two directions in this test-taking attitude: the tendency to be defensive or to put oneself in a too favorable light, and the opposed tendency to be overly honest and self-critical (plus-getting). The extremes of these tendencies are deliberate, conscious efforts to fake bad or lie good.
3. The defensive tendency appears to be related to the clinical picture of hysteria, whereas plus-getting is related to the picture of psychasthenia.
4. The MMPI scales L and F, while relatively effective in detecting extreme distortion, do not seem to be sufficiently subtle to detect the more common and often unconscious varieties of defensiveness or plus-getting. It has been found convenient to begin interpretation of L in the range of T-scores 55 or 60; whereas F does not clearly establish invalidity even up to T-score 80 (raw score about 16).
5. By contrasting item frequencies of abnormal persons showing normal MMPI profiles and elevated L scores, with the records of unselected normals, an empirical key called K has been derived which is relatively successful in detecting the influence of disturbing test-taking attitudes and can be used to improve the discrimination between normals and abnormals.
6. In studying the intercorrelations among a group of scales derived by various means but all functioning with some effectiveness to detect such attitudes, it was found that one common factor is sufficient to account for all of the intercorrelations. The scale (G) which has the largest factor loading was derived by a method of internal consistency and without recourse to any external criterion. Since K is the scale being used to measure this factor, the factor in question has been called K-factor.
7. On the basis of these findings and study of the relationship of MMPI to certain of the Guilford-Martin scales, it is suggested that perhaps the construction of personality inventories by means of item-correlation and factor analytic methods leads to the development of

tests which are excessively loaded with such test-taking attitudes. The procedure of internal consistency in its various forms is called into question as a profitable method for the construction of personality inventories.

Received July 9, 1946.

References

Adams, C. R. (1941). A new measure of personality. *J. appl. Psychol.*, 25, 141-151.

Allport, G. W. (1928). A test for ascendance-submission. *J. abn. Psychol.*, 23, 118-136.

Allport, G. W. (1937). *Personality.* New York: Henry Holt and Co.

Allport, G. W. (1942). The use of personal documents in psychological science. *Soc. Sci. Res. Council Bull.*, No. 49.

Arnold, D. A. (1942). The clinical validity of the Humm-Wadsworth temperament scale in psychiatric diagnosis. Unpublished Ph.D. Thesis, University of Minnesota.

Benton, A. L. (1935). The interpretation of questionnaire items in a personality inventory. *Arch. Psychol.*, No. 190.

Bernreuter, R. G. (1933a). Theory and construction of the personality inventory. *J. soc. Psychol.*, 4, 387-405.

Bernreuter, R. G. (1933b). Validity of the personality inventory. *Person. J.*, 11, 383-386.

Bernreuter, R. G. (1940). The present status of personality trait tests. *Educ. Rec. Supp.*, 21, 160-171.

Bills, Marion. (1941). Selection of casualty and life insurance agents. *J. appl. Psychol.*, 25, 6-10.

Bordin, E. S. (1943). A theory of vocational interests as dynamic phenomena. *Educ. and Psych. Meas.*, 3, 49-65.

Burt, C. (1941). *The factors of the mind.* New York: Macmillan.

Cady, V. M. (1923). The estimation of juvenile incorrigibility. *J. Delinqu. Monogr.*, No. 2.

Eisenberg, P. (1941). Individual interpretation of psychoneurotic inventory items. *J. gen. Psychol.*, 25, 19-40.

Eisenberg, P., & Wesman, A. (1941). Consistency in response and logical interpretation of psychoneurotic inventory items. *J. educ. Psychol.*, 32, 321-338.

Frenkel-Brunswik, E. (1939). Mechanisms of self-deception. *J. soc. Psychol.*, 10, 409-420.

Guilford, J. P. (1936). *Psychometric methods.* New York: McGraw-Hill.

Guilford, J. P., & Guilford, R. B. (1936). Personality factors S, E, and M, and their measurement. *J. Psychol.*, 2, 109-127.

Guilford, J. P., & Guilford, R. B. (1939). Personality factors D, R, T, and A. *J. abn. soc. Psychol.*, 34, 21-36.

Guilford, J. P., & Martin, H. G. (1940). *An inventory of factors STDCR.* Beverly Hills: Sheridan Supply Co.

Guilford, J. P., & Martin, H. G. (1943a). *The Guilford-Martin Personnel Inventory.* Beverly Hills: Sheridan Supply Co.

Guilford, J. P., & Martin, H. G. (1943b). *The Guilford-Martin inventory of factors GAMIN.* Beverly Hills: Sheridan Supply Co.

Hartshorne, H., & May, M. A. (1928). *Studies in deceit.* New York: Macmillan.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule: I. Construction of the schedule. *J. Psychol.*, 10, 249-254.

Hathaway, S. R., & McKinley, J. C. (1942). A multiphasic personality schedule: III. The measurement of symptomatic depression. *J. Psychol.*, 14, 73-84.

Hathaway, S. R., & McKinley, J. C. (1943). *Manual for the Minnesota Multiphasic Personality Inventory*. New York: The Psychological Corporation.

Hendrickson, G. (1932). Attitudes and interests of teachers and prospective teachers. Paper given before Section Q, AAAS, Atlantic City, Dec. 27, 1932 (unpublished).

Horst, P. (1941). The prediction of personal adjustment. *Soc. sci. res. coun. Bull.*, No. 48.

Humm, D. G., & Wadsworth, G. W. (1935). The Humm-Wadsworth temperament scale. *Amer. J. Psychiat.*, 92, 163-200.

Humm, D. G., Storment, R. C, & Iorns, M. E. (1939). Combination scores for the Humm-Wadsworth temperament scale. *J. Psychol.*, 7, 227-253.

Humm, D. G., & Humm, K. A. (1944). Validity of the Humm-Wadsworth temperament scale: with consideration of the effects of subjects' response-bias. *J. Psychol.*, 18, 55-64.

Kelly, E. L., Miles, C. C., & Terman, L. M. (1936). Ability to influence one's score on a typical pencil and paper test of personality. *Character and Pers.*, 4, 206-215.

Laird, D. A. (1925). Detecting abnormal behavior. *Jour. abn. Psychol.*, 20,128-141.

Landis, C., & Katz, S. E. (1934). The validity of certain questions which purport to measure neurotic tendencies. *J. appl. Psychol.*, 18, 343-356.

Landis, C., Zubin, J., & Katz, S. E. (1935). Empirical evaluation of three personality adjustment inventories. *J. educ. Psychol.*, 26, 321-330.

Loth, N. N. (1945). Correlation between the Guilford-Martin Inventory of Factors STDCR and the Minnesota Multiphasic Personality Inventory at the college level. Unpublished Master's thesis, Univ. Minn.

Ludolph, M. (1944). The Guilford-Martin Inventory of Factors GAMIN and its relation to the Minnesota Multiphasic Personality Inventory. Unpublished paper, Univ. Minn.

MacKinnon, D. W. (1944). The structure of personality. In J. McV. Hunt (Ed.), *Personality and the behavior disorders.* New York: Ronald Press.

Maller, J. B. (1930). The effect of signing one's name. *Sch. and Soc.*, 31, 882-884.

Maller, J. B. (1932). *Character sketches.* New York: Bureau of Publications, Teachers College, Columbia University.

Maller, J. B. (1944). Personality tests. In J. McV. Hunt (Ed.), *Personality and the behavior disorders.* New York: Ronald Press, 170-213.

McKinley, J. C., & Hathaway, S. R. (1940). A multiphasic personality schedule: II. A differential study of hypochondriasis. *J. Psychol.*, 10, 255-268.

McKinley, J. C., & Hathaway, S. R. (1942). A multiphasic personality schedule: IV. Psychasthenia. *J. appl. Psychol.*, 26, 614-624.

McKinley, J. C., & Hathaway, S. R. (1944). The Minnesota Multiphasic Personality Inventory: V. Hysteria, hypomania, and psychopathic deviate. *J. appl Psychol.*, 28, 153-174.

McNemar, Q. (1945). The mode of operation of suppressant variables. *Amer. J. Psychol.*, 58, 554-555.

Meehl, P. E. (1945a). The dynamics of structured personality tests. *J. clin. Psychol.*, 1, 296-303.

Meehl, P. E. (1945b). An investigation of a general normality or control factor in personality testing. *Psychol Monogr.*, 59, No. 4.

Meehl, P. E. (1945c). A simple algebraic development of Horst's suppressor variables. *Amer. J. Psychol.*, 58, 550-554.

Metfessel, M. (1935). Personality factors in motion picture writing. *J. soc. abn. Psychol.*, 30, 333-347.

Mosier, C. I. (1936). A note on item analysis and the criterion of internal consistency. *Psychometrika*, 1, 275-282.

Olson, W. C. (1936). The waiver of signature in personal reports. *J. appl. Psychol.*, 20, 442-450.

Page, J., Landis, C, & Katz, S. E. (1934). Schizophrenic traits in the functional psychoses and in normal individuals. *Amer. J. Psychiat.*, 13, 1213-1225.

Rosenzweig, S. (1934). A suggestion for making verbal personality tests more. valid. *Psychol. Rev.*, 41, 400-401.

Rosenzweig, S. (1938). A basis for the improvement of personality tests with special reference to the M-F battery. *J. abn. soc. Psychol.*, 33, 476-488.

Ruch, F. L. [1942]. A technique for detecting attempts to fake performance on a self-inventory type of personality test. In Q. McNemar and M. A. Merrill, *Studies in personality.* New York: McGraw-Hill, pp. 229-234.

Spencer, D. (1938). Frankness of subjects on personality measures. *J. educ. Psychol.*, 29, 26-35.

Steinmetz, H. C. (1932). Measuring ability to fake occupational interest. *J. appl. Psychol.*, 16, 123-130.

Strong, E. K. (1943). *Vocational interests of men and women.* Stanford: Stanford University Press.

Symonds, P. M. (1932). *Diagnosing personality and conduct.* New York: Appleton-Century.

Thurstone, L. L., & Thurstone, T. G. (1930). A neurotic inventory. *J. soc. Psychol.*, 1, 3-30.

Torrens, J. K. (1944). An investigation and evaluation of the Guilford Inventory of factors STDCR with special reference to the Minnesota Multiphasic Personality Inventory. Unpublished paper, Univ. Minn.

Vernon, P. E. (1934). The attitude of the subject in personality testing. *J. appl. Psychol.*, 18, 165-177.

Washburne, J. N. (1935). A test of social adjustment. *J. appl. Psychol.*, 19, 125-144.

Wesley, Elaine. (1945). Correlations between the Guilford-Martin Personality Factors O, Ag, Co and the Minnesota Multiphasic Personality Inventory at the college level. Unpublished Master's thesis, Univ. Minn.

Willoughby, R. R. (1935). The concept of reliability. *Psychol. Rev.*, 42, 153-165.

Willoughby, R. R., & Morse, M. E. (1936). Spontaneous reactions to a personality inventory. *Amer. J. Orthopsychiat.*, 6, 562-575.

Zubin, J. (1934). The method of internal consistency for selecting test items. *J. educ. Psychol.*, 25, 345-356.