THE MEASUREMENT OF DIFFERENTIAL DEFICIT

LOREN J. CHAPMAN and JEAN P. CHAPMAN
University of Wisconsin, Madison, Wisconsin 53706, U.S.A.

DISTURBED schizophrenics show a generalized performance deficit. They perform less well than normal subjects on almost any task that requires a voluntary response. Therefore, a lower than normal score on any single task cannot be interpreted as indicating a deficit of special importance in schizophrenia. Recognizing this, most investigators of cognitive deficit in schizophrenia have been concerned with differential deficit, that is, a greater deficit in one ability than in at least one other ability. Differential deficit in ability is measured using two or more tasks. A differential deficit in ability is inferred customarily from a greater performance deficit on one task than on at least one other task.

On previous occasions we have pointed out that a differential deficit in performance does not necessarily indicate a differential deficit in ability. It may, instead, merely reflect the schizophrenics' generalized performance deficit coupled with the fact that one of the two tasks measures generalized deficit better than the other. We have pointed out that the higher are the reliability and the variance of a task, the greater will be the difference in mean score between the more competent and the less competent subjects. Since schizophrenics usually perform less well than normal subjects, they will show a greater performance deficit on the one of the two tasks that yields the greater dispersion of scores. We have advocated that to measure differential deficit in ability, one must match tasks on several psychometric characteristics.

We explore this problem again now in order to offer what we believe to be a clearer statement of the problem. In addition, we now propose a solution with slightly less cumbersome requirements for adequately matched tests. (We use the term 'tests' here in the broad sense to include any task that is used to measure cognitive deficit).

The central problem is how to move from statements about differential deficit in performance on specific tests to statements about differential deficit in ability. We will present here some implications of classic mental test theory for the solution of this problem.

Some principles of psychometric theory

When parallel tests are given on successive occasions, any given subject will usually not earn identical scores on the two testings. He will score differently even when he does not change in his knowledge of the material and shows no practice effects as a result of his taking the earlier tests. These fluctuations in test score are attributable to changes in such variables as the subject's motivation, fatigue, emotional state, alertness, luck in guessing, and to

*Preparation of this article was supported by a research grant (MH-18354) from the National Institute of Mental Health, United States Public Health Service.
changes in conditions of testing. Fluctuations in test score that occur for such reasons are random and are usually described as reflecting errors of measurement.

It is customary to divide test scores into two components, true score and error of measurement.

\[ X_0 = X_t + X_e \]

where \( X_0 \) = observed score; \( X_t \) = true score; \( X_e \) = errors of measurement.

Errors of measurement are as often positive as negative. If a series of parallel tests were to be administered, the observed test scores for a single subject would be expected to form a normal distribution. The mean of this normal distribution would be the subject's true score, and the difference between the subject's observed score on a given testing and his true score would be attributable to the errors of measurement for the individual on that testing.

The term "true score" may be misleading. It does not refer to the ability which the subject truly has. Instead, it refers to the portion of the score which is replicable, that is, reliable.

It is assumed that true score and errors of measurement are uncorrelated. It follows that the variance of observed scores is the sum of the variances of true score and of errors of measurement.

\[ S_0^2 = S_t^2 + S_e^2 \]

where \( S_0^2 \) = variance of observed scores; \( S_t^2 \) = variance of true scores; \( S_e^2 \) = variance of errors of measurement.

This means that when we observe the distribution of scores of a sample of subjects, the distribution reflects two component distributions, one being the distribution of true scores and the other the distribution of errors of measurement. Thus, the variance of true scores is smaller than the variance of observed scores (except in the unrealistic case of a perfectly reliable test). The relationship of the variance of observed scores and the variance of true scores is expressed as follows:

\[ S_t^2 = r S_0^2 \]

where \( S_t^2 \) = variance of true scores; \( r \) = reliability of the test; \( S_0^2 \) = variance of observed scores.

Some implications of these principles

Let us consider the problem of inferring differential deficit in ability from performance on two tests which differ in true-score variance. We will call these Test 1 and Test 2. Figure 1 presents frequency distributions of true scores for subjects who have taken both tests.

![Fig. 1](image-url)
Figure 1 is drawn so as to make the difference between the tests in true score variance a very large one. Test 1 has the larger variance. In the particular case represented here, the variance of Test 2 is small because of the use of very easy items. Now, consider the result if you compare the mean score of the more able subjects with the mean score of the less able subjects on each of these two tests. For example, one might compare for each test the mean of the high scoring half of the subjects with the mean of the low scoring half of the subjects. Test 1 yields a relatively large mean difference in true score between subjects of high ability and subjects of low ability \( (X_{HI} - X_{LO}) \). For Test 2, the comparable mean difference, \( (X_{HI} - X_{LO}) \), is smaller. Test 1 is intended, of course, as a measure of Ability 1, and Test 2 is intended as a measure of Ability 2. For example, a vocabulary test asking for definitions of nouns might be used as a measure of the ability to define nouns and a vocabulary test asking for definitions of adjectives might be used as a measure of the ability to define adjectives. However, the larger difference in true score for Test 1 does not mean that the more able and less able subjects differ more on Ability 1 than on Ability 2, but instead, reflects only the fact that the test items were chosen so that Test 1 has a greater true-score variance than Test 2. The differential deficit in performance merely reflects the particular choice of items for the two tests. In most comparisons of schizophrenic and normal subjects, the situation is very much like the example represented in Fig. 1. The schizophrenics are the lower scoring subjects and the normal subjects are the better scoring subjects, except, of course, that the two groups overlap somewhat in their accuracy scores. The mean of the observed scores for schizophrenics or for normal subjects is identical to the mean of their true scores; it follows that the schizophrenics will show a greater performance deficit on a test which, like Test 1, has a greater true-score variance than on a test with smaller true-score variance. Differential performance deficit obtained on pairs of tests which differ on true-score variance does not indicate anything about differential deficit in the abilities measured by the tests. One could obtain the opposite differential performance deficit, that is, a greater performance deficit on the test for Ability 2 than on the test for Ability 1, simply by choosing items which yield differing true-score variances for the tests. It follows that to measure differential deficit in ability, one must match tests on true-score variance.

Let us now turn to the problem of just how to go about matching on true-score variance. Since true-score variance is the product of reliability and observed-score variance, true-score variance will be greater when observed score variance and reliability are greater. Both the observed-score variance and the reliability are greater when the items are more numerous, when the items more often fall in the middle range of difficulty, and when the items tend to measure the same ability. One must also match on the shape of the distribution of item difficulty. Otherwise the tests will become unmatched on variance and on reliability for groups of differing mean ability level.

But, one might ask, why not just match on observed-score variance instead of true-score variance? The answer is that observed-score variance includes variance due to errors of measurement, that is, variance which is not replicable. A less reliable test will not reflect the effects of generalized incompetence as well as a more reliable test. One might match two tests of differing reliability on observed-score variance and then give the tests to disturbed schizophrenics and to normal subjects. The schizophrenics would be expected to show less
deficit on the less reliable test than on the more reliable test even if the schizophrenics have
equal deficit in the two abilities. This principle is most intuitively obvious in the extreme case
of a test with a reliability of zero. An example is the task of obtaining 'heads' in coin tossing.
If each of a group of subjects tosses 10 coins and is given a score of number of heads
obtained, the resultant distribution of obtained scores has a variance just as any other
distribution of scores. The reliability, however, is zero and, therefore, the true-score variance
is also zero. If this task were to be paired with a second task that has the same obtained score
variance but has a non-zero reliability, only the second task could distinguish groups that
differ on generalized competence.

The measure of reliability. Often the most convenient estimate of reliability for matching
purposes is coefficient alpha (equivalent to Kuder–Richardson Formula 20 when items are
dichotomously scored) which is the mean of all possible split-half reliability values. (See
Nunnally9 or Lord and Novick4.) Coefficient alpha estimates parallel-form reliability but
requires only a single setting.

The role of difficulty of items
Both test reliability and observed-score variance are heavily influenced by difficulty level.
In the extreme case in which the test is so easy that all subjects earn a perfect score or so
difficult that all subjects earn a zero score, both reliability and variance of observed scores
are zero. Both reliability and variance rise with departures from zero difficulty and 100%
difficulty. Items that are at a 50% level of difficulty for the entire group tested are most
reliable and yield the highest observed-score variance on a dichotomously scored free-
response test on which subjects can seldom use sheer guessing to obtain a correct answer.3
On multiple-choice tests, for which guessing raises accuracy, the highest reliability and
observed-score variance are expected for accuracy levels somewhat higher than midway
between chance and 100% accuracy.6

It follows that the choice of difficulty level of items for two tests has a powerful influence
on which one of the two tests yields the larger difference between schizophrenic and normal
subjects. Let us assume that schizophrenics are as deficient in their ability to define adjectives
as they are in their ability to define nouns. As mentioned earlier, the tests for these two
abilities may differ on true-score variance. To return to our earlier example, investigator A
puts together a noun definition test with items in the middle range of difficulty and an
adjective definition test that is considerably easier. In this case, the distribution of scores for
the noun definition test will resemble the distribution labeled as Test 1 in Fig. 1 and the
distribution of scores for the adjective definition test will resemble the distribution labeled
Test 2. As in the example above, there will be a greater mean difference between the more
able (normal) and the less able (schizophrenic) subjects on the test with the greater true-score
variance—in this case the noun definition test. So investigator A concludes that schizoph-
renics show a greater deficit (as judged by comparison with normal subjects) in defining
nouns than in defining adjectives. Now investigator B sets out to replicate this finding except
that he or she chooses adjectives that are in the middle range of difficulty (say about 0-50)
and chooses nouns that are easier (say about 0-75 accuracy). This time the distribution of test
scores for defining nouns will look like Test 2 and the distribution of scores for defining
adjectives will look like Test 1. The finding will be a greater mean difference between
schizophrenics and normals in defining adjectives than in defining nouns. Investigator B might conclude that schizophrenics show a greater deficit in the ability to define adjectives than in the ability to define nouns. Obviously, both investigators can not be correct. Both findings may be artifacts of the difference in true-score variance on the tests which the investigator chose to test his hypothesis.

This is not just a ceiling or a floor effect. One cannot avoid the effects of difficulty on difference in mean score of groups by limiting one's manipulation of a variable to a middle range of difficulty, using, for example, tasks that produce an error rate between 25% and 75%. Differences in difficulty within this middle range have a substantial effect on the true-score variance and, therefore, on the difference between the more able and the less able subjects. Lord and Novick (p. 329) show how the variance of observed scores may be computed from the item difficulties. For a free-response test of uniform inter-item correlation of 0.20, the observed-score variance is approximately 500 for a 100-item test with uniform item difficulty of 0.50, but drops to approximately 375 for a uniform item difficulty of 0.75. The difference in true score variance will be even larger than indicated by these values because the middle range of item difficulty also yields the higher reliability as well as the higher observed score variance. Hence, a test of items at the 0.75 level of difficulty will result in a smaller difference in mean scores between the more able and less able subjects than will a test of 0.50 difficulty.

We have thus far discussed the case in which differences in difficulty are imposed by the investigator's arbitrary choice of items for the two tests. In many situations, the difference is instead imposed by the hypothesis. An example is the hypothesis that schizophrenics' intellectual accuracy is impaired more than that of normal subjects by distraction. Several investigators have tested this hypothesis by giving a task both with and without distraction to schizophrenic and normal subjects. The usual finding has been that both normal and schizophrenic subjects show poorer performance with distraction, but that schizophrenics show the greater decrement with distraction. This finding might well be an artifact of a greater true-score variance on the task with distraction. To investigate whether distraction has a more specific relationship to schizophrenic deficit one can obtain both a variable that produces distraction and a second variable which is also a source of lowered accuracy for normal subjects. The investigator should match these two error producing variables on the increase in number of errors that they produce for normal subjects. One can then test the hypothesis that distraction produces a greater impairment of schizophrenic performance than does the second source of lowered accuracy. This method is illustrated below in a study of associative intrusions in verbal behavior and in a study of response to contextual constraint.

Choice of standardization group for matching tasks. Normal subjects of a wide range of general competence are the appropriate standardization group. One should not match tests for schizophrenics alone or for combined groups of normal and schizophrenic subjects because the reliabilities and observed-score variances would be affected by any schizophrenic differential deficit in ability. Matching in either of these two ways would require choice of the less reliable items (as judged by normal performance) for the ability in which schizophrenic subjects have the greater deficit. This is because schizophrenic deficit raises true-score variance on a test which measures it. Matching of tests on true-scores variance for
The best nonpsychotic group for matching includes subjects who score at about the same level as the schizophrenics on the two tests combined. Tests which are matched on true-score variance for groups which score at one level may become unmatched for groups which score at a different level because different items contribute to matching at the different performance levels. For example, on a test with a wide range of item difficulties, very competent people will answer all of the easier items correctly and differences among them will be measured only by the more difficult items. On the same test, less competent people will get all of the very difficult items wrong and differences among them will be measured only by the easier items. In effect, the highly competent and less competent people will be taking different tests. The solution is to obtain tests of a wide range of item difficulty, matched, item by item, for subjects of a wide range of ability.

Practical steps to achieve matched tasks

One might suspect that the job of matching on true-score variance is prohibitively laborious. Actually it is not. Matching on all of the test variables can be readily achieved for most pairs of tests by giving large, and preferably equal, numbers of items of the two types to a group of normal subjects of a wide range of ability and then selecting pairs of items of the same difficulty and item-scale correlation. Gulliksen showed that such matching yields tests of equivalent variances and reliabilities. It also yields the same distribution of true scores. The reason for the equal size of the two original pools of items is to reduce the probability that the matching will be achieved by capitalizing on errors of measurement more on one task than on the other. In any case, the matching should be cross validated on a second sample of normal subjects. At least 100 subjects (preferably more) should be used in the initial matching to reduce errors of estimating the true-score variance.

Applications of these principles

Several studies of schizophrenic deficit have been completed using matched tasks. In some of these studies, a traditional schizophrenic deficit was confirmed, and in others, disconfirmed.

Associative intrusions in verbal behavior. Rattan and Chapman tested the hypothesis that schizophrenics are impaired in their use of words by an inability to avoid responding to verbal associative connections of one word to another. The investigators developed two multiple-choice vocabulary subtests. The items in one subtest contained as an incorrect alternative an associate to the stimulus word. An example is the following item:

POOL means the same as
1. puddle (correct)
2. notebook (irrelevant)
3. swim (associate)
4. none of the above

The items in the other subtest contained a second irrelevant word in place of an associative distractor. Using a standardization sample of normal subjects with average and below
average IQ, the subtests were matched on psychometric characteristics as recommended in this paper. Since normal subjects are to some extent distracted by the presence of associates, the matching was achieved by using words in the No-Associates Subtest that would have been more difficult than the words of the With-Associates Subtest if neither subtest contained associative distractors. Thus the differential deficit measured was a greater inability to resist distractors than to define hard words. Chronic schizophrenics, unlike normal subjects, showed a significantly greater deficit on the With-Associates Subtest than on the No-Associates Subtest. The increased schizophrenic errors were chiefly attributable to the patients' choosing associative alternatives. These findings demonstrate schizophrenic heightened susceptibility to associative distraction without the possible artifact of generalized deficit coupled with differential discriminating power of tests.

Schizophrenic response to affect-laden verbal material. Clinicians commonly suggest that schizophrenics show greater thought disorder when dealing with affect-laden material than when dealing with affectively neutral material. Two dozen or more investigators have tested this hypothesis using an emotional and a neutral task which were not matched on psychometric properties. About two-thirds of these studies reported results which support the hypothesis. (The present writers have reviewed this literature.) Two more recent studies have remedied the defect of unmatched tasks.

Chapman et al. gave affect-laden and affectively neutral free-response vocabulary items to drug-free schizophrenics (N = 120) and nonpsychotic subjects (N = 167). The two subtests were matched on the relevant psychometric characteristics. Both chronic and newly admitted schizophrenics were no less accurate on the affective subtest than on the neutral subtest.

One might object that vocabulary is a highly practiced, "overlearned" skill. Chapman and Chapman used a task that requires solution of new intellectual problems by manipulating affect-laden ideas. They gave drug-free schizophrenics (N = 74) and nonpsychotic subjects (N = 206) a test of affect-laden and affectively neutral multiple-choice analogy items. The two subtests were again matched on the relevant psychometric characteristics. Neither newly admitted schizophrenics nor long-term chronic schizophrenics performed differently on the affective subtest than on the neutral subtest. The many published findings of a cognitive deficit in schizophrenia in response to affect-laden stimuli can probably be attributed to the inappropriate use of unmatched tests.

We believe that the popularity of the clinical inference that schizophrenics are disrupted by affective materials is due to the fact that the more severely disordered is a schizophrenic, the more likely he is to talk freely about affect-laden topics that nonpsychotics keep more private. The clinician hears his more disordered patients talk about affect-laden topics and mistakenly infers that the affective content exacerbates the cognitive disturbance.

Schizophrenic cognitive deficit as a function of scoring standards. Schizophrenics are usually found to be less impaired on vocabulary score on standard intelligence tests than on other parts of intelligence tests. Yacozynski suggested that this relative preservation of vocabulary is an artifact of the scoring method. He suggested that schizophrenics give marginally correct answers on all kinds of tests and that such answers are, by convention, scored as correct on vocabulary tests but not on other tests.
CHAPMAN and CHAPMAN tested this hypothesis. They scored schizophrenic and normal subjects' responses to the Stanford–Binet Vocabulary items, the WAIS Vocabulary items, and the WAIS Similarities items by two methods, one relatively strict and the other relatively lenient. For each of the three kinds of items, subtests of strictly and leniently scored items were matched for normal subjects on psychometric characteristics which determine power of the test to distinguish the more able from the less able subjects. A greater deficit on the strictly scored than on the leniently scored items was found for chronic schizophrenics on the Stanford–Binet Vocabulary, for newly admitted schizophrenics but not for chronic schizophrenics on the WAIS Vocabulary, and for neither group on the WAIS Similarities. The results thus gave partial support to the Yacorzynski hypothesis.

Schizophrenics' response to contextual constraint. Contextual constraint in language refers to the fact that the presence of any word in a meaningful sentence or phrase limits the variety of other words that may reasonably precede or follow it. Ordinary English discourse has higher contextual constraint than random word strings and for this reason is easier to learn and remember. Six previous studies compared the effects of contextual constraint on recall by schizophrenic and normal subjects, and four of the studies found greater schizophrenic deficit for learning passages with high constraint than for learning passages with low constraint. All these studies used unmatched tasks.

RAULIN and CHAPMAN remedied this defect. They compared the effects of contextual constraint and list length on short-term recall of word lists by 24 chronic schizophrenics and 120 normal subjects. One subtest of relatively short lists of low constraint and one subtest of longer lists of high constraint were matched on the relevant psychometric characteristics. These matched subtests were used to compute a difference score of accuracy on low constraint lists minus accuracy on high constraint lists. On this difference score, schizophrenics scored lower than normal subjects of the same ability level \( t(47) = 2.33, p < 0.05 \). The schizophrenic deficit in recall was increased less by an increase in contextual constraint than by a shortening of word lists. The direction of this difference is opposite to that which would be expected from previous studies of the effects of contextual constraint on recall by schizophrenics. The findings of the previous studies were probably artifacts of the use of unmatched tasks.

Conclusions

The research literature on differential deficit is rather a mess. Many findings have not been replicable, and researchers have not been able to use measures of differential deficit to identify subgroups of schizophrenics who differ in variety of thought disorder. If we are to do better in the future in the investigation of cognitive deficit, we must use matched tasks.

REFERENCES