Psychological Bulletin
1973, Vol. 79, No. 6, 380-385

PROBLEMS IN THE MEASUREMENT OF COGNITIVE DEFICIT[1]

LOREN J. CHAPMAN[2] AND JEAN P. CHAPMAN

University of Wisconsin

Several dozen studies of differential cognitive deficit appear each year. Artifacts cloud the findings in a majority of these studies because of differences in discriminating power of tasks coupled with generalized deficit of the patients. With two tests of differing reliability, the test with the higher reliability will yield greater performance deficit for the less able subjects. We argue for the advisability, in studies of differential cognitive deficit, of matching tasks on reliability, shape of the distribution of scores, and mean, variance, and shape of the distribution of item difficulty, using normal subjects alone as a standardization group.

Most hypotheses about cognitive deficit are tested in designs in which two or more error scores are compared. We show that in such studies artifactual findings are produced by differences in test characteristics which affect discriminating power. We illustrate these principles by studies of schizophrenic thought disorder, although the principles apply equally well to hypotheses about all other cognitive pathologies such as those of brain damage, drug states, and mental deficiency.

Most investigations of cognitive deficit are concerned with specific deficits. A single score per subject is not sufficient to measure a specific deficit because subjects with cognitive pathology typically show generalized cognitive deficit, that is, low scores on almost any measure of intellectual functioning. Therefore, meaningful statements about specific deficit must be in terms of differential deficit, that is, a greater loss in one ability than in one or more other abilities.

Discriminating Power and Measures of Differential Deficit

The degree of differential deficit in ability is inferred, of course, from a comparison of the extent to which two measures discriminate pathological performance from normal performance. However, the extent of the inferiority of score of pathological subjects depends not only on their deficit in ability but also on the discriminating power of the test. "Discriminating power" refers here to the extent to which the score differentiates the more able from the less able subjects and, hence, differentiates two groups that differ in the ability measured by the test. Tests have identical discriminating power if they have identical distributions of error-free true scores. If deficient subjects are as inferior to normal subjects on one ability as on another, but the test that is used to measure one of the abilities is more discriminating than the test of the other, a greater performance deficit will be found on the more discriminating measure.

For tests of the same format and mode of scoring (e.g., free response items with dichotomous scoring), the discriminating power of a test is a function of mean item difficulty, dispersion of item difficulty, mean item covariance, and number of items. We focus first on item difficulty. In many published studies of cognitive pathology the pairs of tasks used differ sharply on item difficulty in ways which

[1] Preparation of this article was supported by a research grant (MH-18354) and by a Research Scientist Award (K05-MH-05198) to the first author; both grants are from the National Institute of Mental Health.

The authors are indebted to Anne Cleary, Jum Nunnally, and Lyle Jones for criticisms of an earlier draft of this manuscript, and to Randall Daut and Sonja Farley for assistance in gathering the data. They are also indebted to Janet Rafferty for the data on school children; to Joseph Cocke, Richard Cameron, and Thomas Pritchett of San Antonio State Hospital for permission to test schizophrenic subjects; and to Elmer Cady, Raymond Anderson, and Thomas Biever of Wisconsin State Prison for permission to test prison inmates.

[2] Requests for reprints should be sent to Loren J. Chapman, Psychology Department, University of Wisconsin, Madison, Wisconsin 53706.

could easily have produced the results of the study artifactually.

Other things being equal, items closer to the 50% level of difficulty for all subjects tested have the highest discriminating power on a dichotomously scored free response task composed of items on which guessing is not likely to produce the correct answer (Lord, 1952b). On multiple-choice tasks, the highest discriminating power is expected for accuracy levels slightly higher than midway between chance and 100% accuracy (Lord, 1952a).

The effect of item difficulty on discrimination of groups of differing ability can be illustrated by comparing two age groups of children on vocabulary items that differ in difficulty. Figure 1 shows the percentage accuracy of a group of 34 third-grade children and a group of 39 fifth-grade children on the vocabulary items of the Stanford-Binet Intelligence Scale. The abscissa of Figure 1 presents six sequential groups of vocabulary items, each representing a different level of difficulty (as measured by percentage accuracy for third and fifth grades combined). Each point on the abscissa represents several vocabulary items of a given level of difficulty. Against this abscissa are plotted curves for the percentage of third-grade children and the percentage of fifth-grade children passing each set of items. (If the two grades were combined, the resultant curve would represent a variable plotted against itself, except for the slightly different scale conventions of the abscissa and ordinate.) The curves for the two grades tend to join at the extreme ends of the dimension of difficulty, and they attain greatest separation for items at about the 50% level of difficulty for the two groups combined. This relationship between difficulty and discriminating power may be generalized to any two groups that differ in accuracy, including pathological and normal subjects.

To illustrate how differences in difficulty can be mistaken for differential deficit, let us examine a study by Truscott (1970). Truscott compared schizophrenic and normal subjects on four tasks, including recall of random word strings and recall of normal sentences. The groups differed less on recall of random word strings than on recall of normal sentences. Truscott concluded that schizophrenics show a lack of facilitation of recall by normal syntactic constraints.

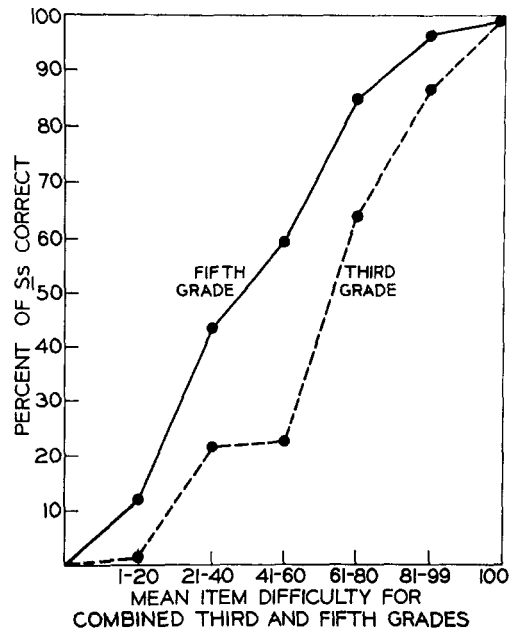

FIG. 1. Accuracy of third- and fifth-grade children on Stanford-Binet vocabulary items of gradated difficulty.

However, her results may instead have been at least in part an artifact of the levels of difficulty (and, hence, discriminating power) of her two tasks. Truscott's tasks were very difficult (corresponding to the left-hand side of Figure 1). Recall of random word strings was at about 20% accuracy for the two groups combined and recall of sentences was much closer to 50% accuracy. If Truscott had used shorter word sequences for both tasks, she would presumably have sampled items at the easy end of the difficulty dimension (corresponding to the right-hand side of Figure 1). Because random word strings are often more difficult than sentences, they would then have been closer than sentences to the middle range of difficulty. Hence, Truscott would probably have found the opposite result, that is, a greater deficit in recalling random word strings than in recalling sentences.

In many studies one cannot judge the direction of the probable artifact because it is not clear which task is more discriminating for

normal subjects. If the items of both tasks are very easy, most normal subjects make almost no errors on either task, but a less able group that makes more errors might find one task more difficult than the other. Still other studies use tasks for which adequacy of performance is defined by criteria other than number of errors, such as by speed of response, as in reaction time tasks, or trials to criterion, as in verbal learning. If the tasks are not matched on these criteria of difficulty, the direction of the probable artifact is ambiguous because the most discriminating level of performance is unknown. One may nevertheless be sure that differences in difficulty produce differential discrimination of groups.

How to Match Tests on Discriminating Power

In conventional psychometric theory, reliability is the index of discriminating power in that the square root of the reliability is the correlation of obtained scores with error-free true scores. Yet equivalent reliability is not sufficient by itself for matching tests in studies of differential deficit because of differences in general ability level between the standardization sample and the experimental subjects. The rationale follows.

The proper standardization sample consists of normal subjects only. Tests differ in discriminating power to the extent that they distinguish the more able from the less able subjects among all subjects tested, that is, the normal and pathological subjects treated as a single group. Yet, one would not wish to match tasks for the two groups combined because reliability of the tests is affected by any true differential deficit due to pathology. The task with the greater schizophrenic deficit will have the larger variance for the two groups combined, and hence, the higher reliability. Matching on reliability for the two groups combined would eliminate some or all of this evidence of differential deficit in ability.

Tasks should instead be matched using normal subjects of a wide range of ability. If the tasks are developed before the experiment proper, the groups in the experiment will probably have a different range of ability than the standardization sample. As a result, mean item difficulty of the tests will be different than in the standardization sample, so that tests that had been matched only on reliability and not on item difficulty would probably become unmatched on reliability for the new groups. However, if tests are also initially matched on item difficulty, they should usually remain matched on reliability for a normal group of a different mean ability level. The match should be on mean, variance, and shape of the distribution of item difficulty. Comparable shapes of the distributions of test scores are also necessary if scores from the two tests are to be meaningfully compared. The matching of the two tests using normal subjects of a wide range of ability will usually result in the two tests retaining equivalent reliabilities for normal subgroups of different levels of ability. This equivalence must be verified for any particular group of pathological subjects by recomputing test characteristics for a normal sample which scores at the level of the pathological group. With such equivalence any difference between the two test scores for pathological subjects can be attributed to genuine differential deficit in ability.

Despite fairly close matching on the above variables, tests may differ slightly on mean or variance for normal subjects. Tests may also differ slightly on number of items. If, in such cases, the two distributions of scores have similar shapes, one might reasonably convert all scores to standard scores on the basis of the performance of the normal standardization sample. This conversion corrects for differences between the tests in mean and variance and facilitates comparison. (It does not, of course, correct for differences between tests in discriminating power.)

Illustrative Example

The need for equivalent discriminating power may be illustrated by some data recently gathered by the present writers. A 114-item multiple-choice analogies test with items of gradated difficulty was given to 49 severely disturbed schizophrenics and 206 normal subjects. This test was used as a pool of items from which various subsets of items were drawn to compare schizophrenic and normal performance for purposes of this paper. The

114 items were arbitrarily divided to yield two forms. Form A consisted of the 62 items for which the first word of the analogy began with a letter that occurs in the alphabet between A and L. Form M consisted of the remaining 52 items for which the first letter of the word fell between M and Z. Normal subjects scored about the same on the two forms, and the schizophrenics showed about the same performance deficit on one form as on the other.

The effects of difference in reliability on performance deficit may be illustrated by drawing subtests from the two forms with identical means but different reliabilities for normal subjects. Coefficient alpha (Kuder-Richardson Formula 20) was our measure of reliability. As shown in Table 1, a pair of subtests was drawn with a coefficient alpha of .80, one subtest from each form. A second pair of subtests was drawn with lower coefficient alpha values, .47 for the subtest from Form A and .49 for the subtest from Form M. These four subtests were closely matched on mean accuracy. The more reliable subtests had the lower variances of item difficulty and the higher test variances, as might be expected. Table 1 shows the mean accuracy of the schizophrenic and normal subjects for each subtest. The schizophrenics showed, as predicted, a greater preformance deficit on the subtests with the higher reliabilities, whether from Form A or M. Thus, a comparison of the schizophrenic performance on the *less* reliable Form M subtest with the *more* reliable Form A subtest yields poorer schizophrenic accuracy on Form A than on Form M ($t = 4.89$, $p < .001$). On the other hand, if one should choose the *more* reliable Form M subtest and the *less* reliable Form A subtest, one would obtain the opposite finding, with the schizophrenics scoring lower on Form M than on Form A ($t = 4.32$, $p < .001$). These results show how easily differences in reliability may artifactually yield an apparent differential deficit in ability. If the two forms were designed to measure different abilities, as in the usual study of schizophrenic cognition, the investigator could easily conclude from performance on either one of these pairs of subtests that he had demonstrated a differential deficit in ability.

Let us now examine data that show how differential difficulty of items can produce differences in discriminating power of tests despite equivalent reliabilities (both difficulty and reliability again measured by normal performance alone). These data are from a free response vocabulary test that was given to 52 schizophrenics and 56 normal control subjects, some of the items being affect laden (emotional) and others being affectively neutral. Table 2 shows the performance of schizophrenic and normal subjects on two 10-item emotional and two 10-item neutral subtests with the coefficient alphas around .70. Two subtests, one emotional and one neutral, were chosen from items of medium difficulty and two others, one emotional and one neutral, from very difficult items. The items of medium difficulty were of about 65% accuracy for the normal subjects alone and about 51% for the two groups combined. The very difficult items were of less than 10% accuracy for

TABLE 1

RELATIONSHIP OF SCHIZOPHRENICS' PERFORMANCE DEFICIT TO TEST RELIABILITY

Subtest	Number of items	Coefficient alpha	Variance of item difficulty	Accuracy			
				Normals		Schizophrenics	
				$\bar{X}$	*SD*	$\bar{X}$	*SD*
High reliability							
Form A	10	.80	.02	6.3	2.8	3.0	2.8
Form M	10	.80	.03	6.4	2.7	3.0	2.6
Low reliability							
Form A	10	.47	.08	6.5	1.6	4.2	1.9
Form M	10	.49	.08	6.3	1.7	4.6	2.1

TABLE 2

RELATIONSHIP OF SCHIZOPHRENICS' PERFORMANCE DEFICIT TO TEST DIFFICULTY

Degree of difficulty	Number of items	Coefficient alpha	Variance of item difficulty	Accuracy			
				Normals		Schizophrenics	
				$\bar{X}$	SD	$\bar{X}$	SD
Medium							
Emotional	10	.68	.04	6.4	2.3	3.8	3.2
Neutral	10	.68	.04	6.6	2.2	3.8	2.8
High							
Emotional	10	.73	.003	.7	1.3	.8	1.7
Neutral	10	.71	.003	.7	1.3	.5	1.2

each group. The schizophrenics, as expected, did not differ from the normal subjects on the very difficult subtests ($t = .32$ for emotional and .72 for neutral), but showed a larger deficit on the easier subtests ($t = 4.76$ for emotional and 5.84 for neutral; $p < .001$ for both). The investigator who compared groups on one emotional and one neutral subtest might conclude that he had found a greater schizophrenic deficit in emotional than in neutral vocabulary, or vice versa, if he chose subtests of differing levels of difficulty.

ALTERNATIVE WAYS TO RULE OUT DIFFERENTIAL DISCRIMINATING POWER AS AN ARTIFACT

In some situations, tasks unmatched on discriminating power may give legitimate evidence of differential deficit in ability. If the control task is a more discriminating measure of nonpathological differences in ability than the experimental task, and if, despite this disparity, the experimental task yields the greater difference between pathological and normal performance, one must attribute the differential performance to a true differential deficit rather than to a generalized deficit coupled with a difference between the tasks on discriminating power. For example, the more reliable subtests of standard intelligence tests are usually more discriminating than the less carefully constructed experimental tasks that are used in most studies of pathological deficit.

A less adequate approach that is sometimes used is to select atypically poor-scoring control subjects so as to match pathological and normal groups on accuracy on the control task. This method is often complicated by the problems of statistical regression. One must select atypically good-scoring pathological subjects and atypically poor-scoring normal subjects on the control task to achieve the matching, and these atypically scoring subgroups usually regress toward the means of the larger groups on their scores on the experimental task. Matching subgroups may also produce systematic unmatching on other variables, as recently discussed by Meehl (1971). The usefulness of findings from a study using such matching ranges from fair to zero depending on how extremely atypical the subjects must be to achieve the matching.

STATISTICAL SOLUTIONS

Several investigators have attempted statistical solutions to the problem of demonstrating differential deficit with tests of differing reliabilities. These statistical solutions include (*a*) analysis of covariance to covary out the scores on the control task from scores on the experimental task, or regression analysis to estimate scores on the experimental task from score on the control task, and then assign a residualized error score to each subject, and (*b*) estimation of error-free true scores by correcting obtained scores using measures of reliability. Neither of these methods is adequate. In each case the more discriminating test will still yield a greater performance deficit than the less discriminating test when given to a group of subjects with lower ability

than the control group. A statistical solution to the problem may be possible, but we are not aware that one has yet been developed.

REFERENCES

LORD, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika,* 1952, **17**, 181–194. (a)

LORD, F. M. A theory of test scores. *Psychometric Monographs,* 1952, No. 7. (b)

MEEHL, P. E. High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology,* 1971, **77**, 143–148.

TRUSCOTT, I. P. Contextual constraint and schizophrenic language. *Journal of Consulting and Clinical Psychology,* 1970, **35**, 189–194.

(Received May 22, 1972)