DOCUMENT RESUME

ED 292 857 TM 011 239

AUTHOR Razel, Micha; Eylon, Bat-Sheva
TITLE Validating Alternative Modes of Scoring for Coloured Progressive Matrices.
PUB DATE Aug 87
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Intelligence Tests; Measurement Techniques; *Scoring; Test Validity; *Weighted Scores; *Young Children
IDENTIFIERS *Coloured Progressive Matrices

ABSTRACT

Conventional scoring of the Coloured Progressive Matrices (CPM) was compared with three methods of multiple weight scoring. The methods include: (1) theoretical weighting in which the weights were based on a theory of cognitive processing; (2) judged weighting in which the weights were given by a group of nine adult expert judges; and (3) empirical weighting in which the weights were a function of the test scores of the examinees who chose each response. The study is based on data from a group of children, aged four to six years. Validity of the CPM with different scoring modes was measured by Pearson product moment correlations between the scores on each administration of the CPM and the scores on other tests of general intelligence. Results indicate that multiple weight scoring of the CPM is superior to conventional scoring in that it increases the test's reliability and validity. Empirical weighting was the most efficient scoring method. Three tables, one figure, and one graph are presented. (TJH)

Validating Alternative Modes of Scoring for the Coloured Progressive Matrices

Micha Razel and Bat-Sheva Eylon

Science Teaching Department

The Weizmann Institute of Science

Rehovot, ISRAEL

August 1987

This paper was presented at the AERA Convention, Washington DC, April 1987. The au ›rs express their gratitude to Yetti Varon and Michal Rapson for helping with the statistical analyses.

Abstract

Conventional scoring of the CPM was compared with three methods of multiple weight scoring: (a) theoretical weighting, in which the weights were based on a theory of cognitive processing; (b) judged weighting, where the weights were given by a group of expert judges; and (d) empirical weighting, where the weights were a function of the test scores of the examinees who chose each response. The results, based on data from a group of children 4 to 6 years old, indicate that multiple weight scoring of the CPM is superior to conventional scoring in that it increases the test's reliability and validity. Empirical weighting was the most efficient scoring method.

Alternative Scoring Modes for the Coloured Progressive Matrices

Administering the Coloured Progressive Matrices (CPM) (Raven, 1977) to four-year-olds in the framework of a curriculum evaluation and development study (Eylon & Razel, 1986, Razel & Eylon, 1986), it was noted that as the test goes on and the items get more difficult, the tester becomes increasingly frustrated with the child's inability to point to the correct answers. At the same time, the child continues contently to work through the test choosing these incorrect alternatives very calmly. This observation could be explained by the assumption that the child thinks that he solves the items correctly even if, by the examiner's and the test's standards, he does not.

Standard rights-only scoring of the CPM gives the examinee one point for each correct item and no points for choosing any of the incorrect distractors. This scoring method can be justified only by the presupposition that no information concerning the examinee's intelligence can be obtained from his particular choice of incorrect alternatives and, in other words, that he chooses randomly between incorrect alternatives if he does not know the correct answer. This assumption was originally challenged by Sigel (1963), who argued that there is much information to be gained from the analysis of the incorrect responses. He did not. however, point to a systematic way to do this. Raven (1977), though admitting that responses for difficult problems are not random, claimed that "erroneous responses cannot be used satisfactorily for the quantitative assessment of intellectual" ability (p. 4).

One simple and systematic way of integrating information contained in the choice of distractors with the information contained in the choice of correct responses is multiple weight scoring in which all choices receive different scores, or weights. Thissen (1976) found that multiple weight scoring of the CPM yielded from one third more to nearly twice the information obtained by conventional scoring for the lower half of ability range among 561 junior high school students. For the upper half of the ability range no information increase was obtained. Thissen presented curves that showed

clearly that the probability of choosing different incorrect response alternatives varied differently with ability, indicating that different incorrect items are favored in different ability levels (for similar findings with other tests see Levine & Drasgow, 1983, and Thissen & Sternberg, 1984). Jacobs and Vandeventer (1970) showed that choice of certain "better" distractors in the CPM is systematically related to superior overall performance on the test among young children and low-ability subjects but not among older, higher ability examinees.

Several investigators have compared reliability and validity indices of tests scored conventionally and scored with weights for all choices. Davis and Fifer (1959) did this with a multiple choice arithmetic test, Hendrickson (1971)--with subtests of the SAT, Reilly and Jackson (1973) used GRE tests, and Kansup and Hakstian (1975) employed verbal and arithmetic reasoning tests. These investigators found that multiple weighting usually resulted in substantial increases in reliability but in no change, or decrease, in validity of the tests. Raffeld (1975), however, obtained increases in both reliability and validity.

This study is aimed at applying the technique of multiple weighting to the CPM when used with young children. As indicated above, several researchers obtained evidence for a relationship between choice of incorrect options in the CPM and mental ability. However, no comparison of reliability and validity under alternative scoring methods was made using the CPM.

In the present study, conventional scoring was compared with three methods of multiple weighting. The first, *empirical weighting*, used in most of the studies reviewed above, was based on the average, conventionally scored, CPM score of all subjects who had chosen a particular distractor. The second, *judged weighting* (referred to as "a-priori" and "logical" by Davis & Fifer, 1959, and Kansup & Hakstian, 1975, respectively), was based on merits of the choices as judged by a group of adults. The third, *theoretical weighting*, consisted of an evaluation of the response alternatives in light of a theoretical model of cognitive mental development.

The model distinguishes 4 levels of cognitive processing. (a) *Wholistic processing* is the lowest

level and consists of choosing a distractor, such as choice 1 of item AB9 in Figure 1, that globally

---

Insert Figure 1 about here

---

gives the same impression as the matrix. It is based on a matching response but it reflects an inability to make detailed and exact comparisons (e.g., pay close attention to size), and a disregard for the formal requirements of the task. (b) *Matching processing* is a correct response in certain items of the CPM, that is based on matching the pattern in the answer to the pattern in the matrix, such as answer 6 of item A5. (c) *Single dimension processing* consists of choosing a distractor that is correct as far as one dimension, horizontal or vertical, of the matrix is concerned, e.g., answers 1 and 6 of item A8. It is also based on a matching response. but there is a prior isolation of a single dimension along which the match is made. (d) *Two dimension processing* consists of choosing a correct response in certain CPM items that is correct on both the horizontal and vertical dimensions of the matrix, such as answer 2 of item A8. Our theoretical weighting of the CPM's response alternatives consisted of giving weights in accordance with the level of processing hypothesized to have been used by the child to reach his response.

This simple model of processing yields as an immediate result some intuitive conclusions that cannot be accounted for by conventional scoring. For example, that the choice of an erroneous distractor on some items may reflect a higher cognitive level than a correct response on simpler items. Hence, according to the model, choosing an incorrect alternative in certain items based on single dimension processing is considered superior and gives the subject more credit than choosing the correct answer based on matching processing in other items. Another possibility revealed by the model is that lower levels of processing may, through a chance effect, result in the choice of superior responses. For instance, answer 2 is the only correct response for item A8 for a person who reached the two dimension processing level. But a child who operates on the single dimension processing level may choose either distractor 1 or 6, based on single dimension processing, as well as the correct answer, 2, based on either vertical or horizontal processing.

Method. About 200 preschoolers, 4 to 6.5 years old, were tested twice with the CPM, once at the beginning and once at the end of a 1.5-year-long experiment (Eylon & Razel, 1986; Razel & Eylon, 1986). Most of these children were also given the Harris-Goodenough Draw-a-Man intelligence test twice during the same testing periods. The Draw-a-Woman test was administered only once, during the first testing period. A complete WPPSI (Lieblich, 1969) was individually administered to a subsample of the children during both testing periods. To retain the full range of variability of the scores in our sample, raw scores were used as the basis for all analyses rather than the transformed IQ scores.

The average standardized total, conventionally scored, CPM scores of all subjects who chose a particular response alternative were used as weights for the empirical weighting method. The method was essentially identical to what was described by other researchers, e.g. Reilly and Jackson (1973) with the one difference that these researchers computed the total score on the *remaining* items of the test, while we used all the test's items for calculating the total score in order not to discard any information concerning the examinee's intelligence.

Nine adults working in the field of science education were used as judges for the judged weighting method. The CPM was introduced to them as an intelligence test and they were asked to rate the response alternatives on a scale from 1 to 6 from the "poorest" to the "best" answer. These ratings were averaged and used as weights for judged weighting.

For theoretical weighting, weights 1, 2, 3, and 4 were given to the response alternatives that were based on the processing levels described above as *a*, *b*, *c*, and *d* respectively. All other responses were given a weight of 0. Responses that could be reached through more than one processing level were given the appropriate averaged weight.

Results and discussion.

Table 1 gives the internal-consistency coefficients, α, for the different scoring methods. The data

---

Insert Table 1 about here

---

indicate a sizeable increase in reliability going from conventional scoring to empirical weighting. Table 1 also provides the k values calculated by the Spearman-Brown formula (e.g., Reilly & Jackson, 1973) which give the estimated number of times the original test was effectively increased, i.e., the increase in test length that would be necessary, given conventional scoring, in order to achieve the obtained increase in reliability. The table shows that to achieve the increase in reliability obtained through empirical weighting while using conventional scoring, one would have to increase the CPM 2 5 times, i.e., give the children 90 items instead of the present 36. Theoretical and judged weighting also yielded effective test length increases but to more moderate extents.

Validity of the CPM with different scoring modes was measured by Pearson product moment correlations between the scores on each administration of the CPM on the one hand, and scores on the other tests of general intelligence on the other hand. The data are given in Table 2. Using

---

Insert Table 2 about here

---

multiple weight instead of conventional scoring, the average correlation between the CPM score and scores on the criterion tests (calculated after performing Fisher's r to Z transformation) rose from .40 up to .45. Conventional scoring was compared to the three methods of multiple weighting as to which yielded higher validity coefficients based on the data given in Table 1. In 23 cases multiple weighting was superior to conventional scoring, in 6 cases the reverse obtained and there was 1 tie. A sign test (Hays, 1963, p. 625) yielded $z = 2.97$, indicating a statistically significant advantage of multiple weighting over conventional scoring. To compare the individual methods of multiple weighting with each other and with conventional scoring, all pairwise sign test comparisons were performed, the z values of which are given in Table 3. The table indicates that conventional scoring,

---

Insert Table 3 about here

---

theoretical, judged, and empirical weighting constitute a series of scoring methods that increasingly improve the validity of the CPM. Of the six comparisons given in Table 3 only two reached a one-tailed significance level of .05 - the superiority of empirical weighting over conventional scoring and over theoretical weighting.

The results indicate that conventional scoring is the poorest form of scoring in terms of reliability and validity. Our explanation is that making use of the information contained in the child's choice among incorrect responses gives the other three scoring methods an edge in reliability and validity. Theoretical weighting seems to be inferior to judged weighting probably because of the simplicity of the model of cognitive processing on which it was based relative to the complexity of the CPM. For example, the theory does not apply at all to ten items of the CPM where the distractors differ from the correct response on such dimensions as direction, color, number, size etc. For such variations there does not seem to be an a priori principle by which they could be ordered in terms of difficulty or levels of mental processing, and they were therefore not included in the model. The judges whose ratings were used in the judged weighting seem to have used implicit cognitive theories that were more complex and that provided a closer approximation to the true processes. They were thus also able to order the distractors that our cognitive processing model was unable to order. One possible reason why empirical weighting was superior to judged weighting is that the adult judges were not completely able to identify with the children who took the CPM and judge corectly what was easy and difficult for the young examinees.

Why did empirical weighting result in superior validity and reliability in this study while this was not always found in other studies? One reason may have been the difficulty of the CPM for the young subjects in this study. Levine and Drasgow (1983), Thissen (1976) and Thissen and Sternberg (1984) pointed out that the information gain resulting from multiple weights lies in the lower ability

half. The CPM was intended by its author to be used by almost the whole range of human development: from age 3 to 60 (Raven, 1977). Of necessity, this makes the 36-item test extremely difficult for the youngest ages, the very ages included in our analysis. To see whether the effect of multiple weighting is age-related, the sample was divided in two. children who were between 4 and 5 years old when taking the CPM and those who were between 5.5 and 6.5 years old. Only correlations between tests taken within one year were considered. The average correlation between the children's scores on the CPM and their scores on the WPPSI or Draw-a-Man/Woman are given in Figure 2. The average slope is steeper for the younger group which seems to show that multiple

---

Insert Figure 2 about here

---

weighting was relatively more effective for the younger age group than for the older group.

One explanation for the finding in the above cited studies that multiple weight scoring yielded a greater information increase for low- than for high-ability subjects and for the present finding that multiple weighting yielded greater improvement in validity for younger than for the older subjects may be the smaller number of items answered correctly by low-ability and young children. This relatively small number of correct responses leaves a relatively large number of items that are not used as a source of information by conventional scoring but are so used by multiple weight scoring. For example, the average number of items solved correctly in the groups of 4- and 6.5-year-olds was 12.5 and 17.4 respectively. Thus, for our subjects, multiple weighting made it possible to derive information concerning the children's intelligence from either an additional two thirds or an additional half of the test's 36 items depending on the child's age.

A second reason why empirical weighting resulted in superior validity and reliability in this study and not in others may be related to the test, the CPM. It may be that the distractors of the CPM are particularly constructed so as to appear correct to different lower levels of cognitive development while the distractors of other tests may be constructed according to very different principles, e.g., to

be very similar to the correct response. Raffeld (1975), for example, called for test "writers to deliberately attempt to write distractors that appeal to examinees of differing ability levels" (p. 184). But the failure of tests that were not so written to yield increased validity and reliability for multiple weight scoring should not be surprising. Summarizing the above considerations, the finding of increased reliability and validity with multiple weight scoring may be test- and age-specific.

References

Davis, F. B., & Fifer, G. (1959). The effect of test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement, 19,* 159-170.

Eylon, B., & Razel, M. (1986). The acquisition of some intuitive geometrical notions in the ages of 3 - 7: Cognitive gains acquired through the Agam Method. In L. Burton, & C. Hoyles (Eds.), *Proceedings of the tenth International Conference for the Psychology of Mathematics Education,* pp. 87-92. London: University of London Institute of Education.

Harris, D.B. (1963) *Drawing as a measure of intellectual maturity: A revision and extension of the Goodenough Draw-a-Man Test.* New York: Harcourt, Brace & World.

Hays, W. L. (1963). *Statistics for psychologists.* New York: Holt, Rinehart and Winston.

Hendrickson, G. (1971). The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement, 8,* 291-296.

Jacobs, P. I, & Vandeventer, M (1970). Information in wrong responses. *Psychological Reports, 26,* 311-315.

Kansup, W., & Hakstian, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement, 12,* 219-230.

Levine, M V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43,* 675-685.

Lieblich, A. (1969). [*WPPSI: A Hebrew manual for the* Wechsler Preschool and Primary Scale of

Intelligence (2nd ed.). Jerusalem: Hebrew University and Ministry of Education and Culture.

Raffeld, P. (1975). The effect of Guttman weights on the reliability predictive validity of objective tests when omissions are not differentially weighted. *Journal of Educational Measurement, 12,* 179-185.

Raven, J. C, Court, J. H., & Raven, J. (1977). *Manual for Raven's Progressive Matrices and Vocabulary Scales: The Coloured Progressive Matrices.* London: Lewis.

Razel, M., & Eylon, B. (1968). Developing visual language skills: The Agam Program. *Journal of Visual and Verbal Languaging, 6(1),* 49-54.

Reilly, R. R., & Jackson, R. (1973). Effects of empirical option weighting on reliability and validity on an academic aptitude test. *Journal of Educational Measurement, 10,* 185-194.

Sigel, I. E. (1963). How intelligence tests limit understanding of intelligence. *Merril-Palmer Quarterly, 9,* 39-56.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13,* 201-214.

Thissen, D., & Sternberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501-519.

Table 1

Internal-Consistency Coefficients

for Four Scoring Methods

	α	k
Conventional Scoring	.63	--
Theoretical Weighting	.66	1.14
Judged Weighting	.71	1.44
Empirical Weighting	.81	2.50

Table 2

Correlations between Scores on Each of Two Administrations of the CPM and Scores on Other Intelligence Tests

Test 1	Man1		Man1		Wom1		WPSSI1		WPPSI2		Mean
Test 2	CPM1	CPM2	CPMi	CPM2	CPM1	CPM2	CPM1	CPM2	CPM1	CPM2	
n	219	97	176	178	219	96	121	76	79	79	
Conventional Scoring	.38	.30	.18	.09	.37	.34	.64	.63	.47	.56	.40
Theoretical Weighting	.39	.36	.13	.15	.38	.39	.63	.64	.51	.55	.41
Judged Weighting	.44	.32	.14	.15	.44	.31	.68	.62	.56	.56	.42
Empirical Weighting	.43	.37	.19	.15	.45	.37	.69	.68	.51	.61	.45

Note. Man = Draw-a-Man, Wom = Draw-a-Woman, 1 suffixed to test name = test administered during pretesting, 2 suffixed to test name = test administered during post-testing.

Table 3

Z-test Scores for Pairwise Sign Test Comparisons for Four Scoring Methods Based on Validity Correlation Coefficients

	Theoretical Weighting	Judged Weighting	Empirical Weighting
Conventional Scoring	.95	.67	2.85*
Theoretical Weighting		.67	1.77*
Judged Weighting			1.33

Note. A positive z indicates the superiority of the scoring method given in the column over that given in the row.

* $p < .05$

Figure 1

Three sample items from the CPM

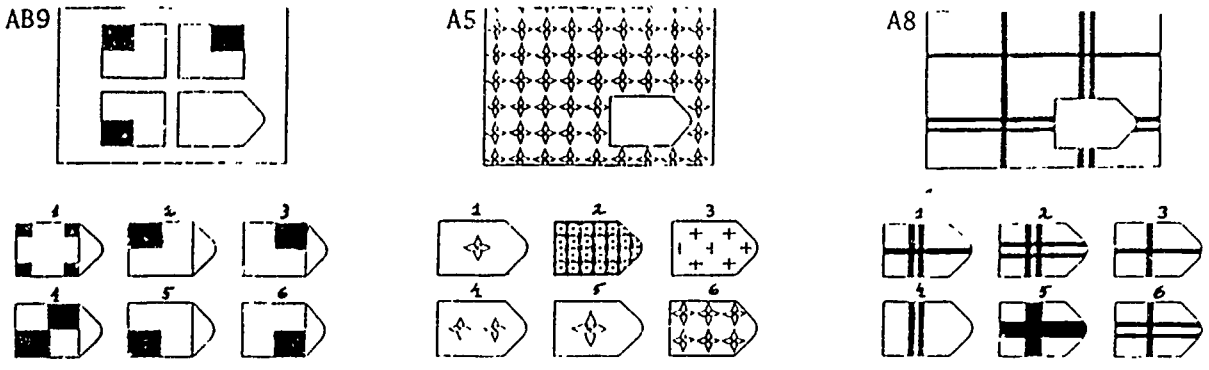

Figure 2

Average correlation between CPM and WPPSI or

Draw-a-Man/Woman by age and scoring method