EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1982, 42

A MULTITRAIT-MULTIMETHOD VALIDITY STUDY OF A TEST OF FORMAL REASONING

PATRICIA KENNEDY ARLIN
University of British Columbia

A paper-and-pencil test of seven of Inhelder and Piaget's formal operational schemes was developed. This objective assessment of formal operational thought was cross-validated clinically. The multitrait-multimethod procedures indicated that the objective test is a valid and reliable measure of formal operational thought. It has the advantage of being administered in large groups, of being easily scored, and of not requiring special training for its administration.

THE application of Piaget's model of cognitive development to instructional decisions, to the adaptation of curriculum, and to educational policy and planning has been hampered by the lack of assessment tools for large group testing. The requirement of clinical procedures to determine both a child's operational level and the child's acquisition of specific concrete and formal concepts limits the usefulness and availability of such information. Despite this limitation some relationships have been clinically established between operational competence and school achievement (Arlin, 198lb; Gallagher, 1979; Tomlinson-Keasey, Eisert, Kahle, Hardy-Brown, and Keasey, 1979).

The renewed interest in cognitive levels matching (Epstein, 1981; Fusco, 1981; Shayer, 1979; Shayer and Wylam, 1978) has also emphasized the need for large group assessment procedures. Through cognitive levels matching, curricular tasks are analyzed in terms of their assumed cognitive prerequisites and attempts are

Requests for reprints may be sent to Patricia Kennedy Arlin, Department of Educational Psychology, 2125 Main Mall, University Campus, Vancouver, B. C. V6T 1W5.

made to match the demands of these tasks with the cognitive levels and capacities of the learners. The simplest definition of cognitive level is in terms of the Piagetian stages, i.e., whether one is concrete operational or formal operational. This interest in "matching" emphasizes the need for a paper-and-pencil means of assessing both cognitive levels and operational competence. This interest in cognitive levels matching is related to a growing awareness on the part of educators of the research on brain growth (Epstein, 1978, 1981; Epstein and Toepfer, 1978) and of the possible linkage between the stages of brain growth and the Piagetian stages of the development of logical thinking (Inhelder and Piaget, 1958).

Recently the investigator has attempted to develop a comprehensive objective measure of the classical formal operational tasks. This attempt is in keeping with Neimark's (1979) observation:

> there is mounting insistence upon tightening up of procedures to insure greater comparibility across studies and shifting toward use of a variety of objective measures of task performance in place of the traditional gross qualitative ratings (p. 62).

Objective measures are a necessity for any possibility of the large scale use of cognitive developmental information in educational settings. The construction of an objective test to assess formal operational thought in terms of the schemata could control for the performance factors of vocabulary level, specificity of instructions, and the varying complexity of the materials used in the traditional assessments.

If a standardized paper-and-pencil assessment can provide a valid assessment of whether or not groups of students are functioning at the formal operational level, the advantages of such an approach over the clinical method would be substantial. Included among these advantages are (a) economy of time in testing large groups rather than individuals one-at-a-time; (b) a comprehensive assessment across seven of the formal schemata rather than a restriction to select one scheme for the individually administered task sessions; (c) the waiving of a requirement of trained clinicians, and (d) use of an objective rather than an subjective scoring scheme.

The purpose of this study was to determine the validity of a paper-and-pencil test of formal reasoning to ascertain whether it could be substituted for the more time-consuming clinical procedures for obtaining formal operational information and whether it could be used in applied settings without the need for trained clinical interviewers.

Method

Subjects

The subjects were 38 military (non-Army) recruits who were tested at one of two recruit basic training centers within six weeks time. They were sampled from 217 recruits tested at the West Coast center and 177 tested at an East Coast training center. The sample was predominantly white male. There was an ethnic mixture (though not in the same percentages as the general population) as well as a variation in socio-economic status. Approximately 15% were female.

For purposes of the present study, all females, all individuals beyond the age of 20.0 years, and persons unavailable for retesting (primarily because of attrition) were eliminated from the subject pool. This series of constraints left a male sample pool of 244 with an average age of 18.7 years (range being 17 years, 9 months to 19 years, 11 months).

Materials: Objective Test

The Test of Formal Reasoning was constructed by the author (Arlin, 1982b). It consists of 48 multiple-choice items. These items are organized into nine subtests: (a) classification; (b) volume; (c) combinations; (d) isolation of variables; (e) proportions; (f) probability; (g) correlation; (h) mechanical equilibrium; and (i) coordination of two frames of reference. Seven of these nine subtests were composed of at least four items each. The probability subtest had three items. The volume and mechanical equilibrium subtests each had two items.

Performance on seven of the eight formal schemata (Inhelder and Piaget, 1958), was assessed. The only scheme not appearing in the test was "forms of conservation beyond direct verification." This task was omitted because of the author's inability to develop an appropriate multiple-choice format representation of this type of problem. This scheme will be included in a revision of the test currently under development. These schemata are defined by Inhelder and Piaget (1958) as "the concepts which the subject potentially can organize from the beginning of the formal level when faced with certain kinds of data, but which are not manifest ouside these conditions" (p. 308).

Two other types of tasks were included to provide a basis for

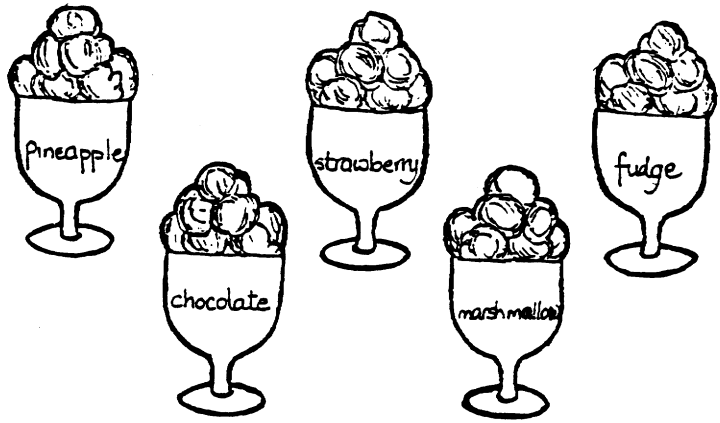
instrument validation and to extend the lower range of possible performance responses. These tasks are the subtests of classification and volume, the latter being an example of the scheme of multiplicative compensation. The classification items were typical of the hierarchical and multiple-classification tasks that can be represented in matrix or inter-sectional form. The conservation of volume at equal concentrations of matter was the second task. The classification tasks provided a reference point from the concrete stage for assessing progress in the acquisition of the formal schemes. The conservation of volume task was included because of the unique role accorded the conservation of volume in the transition from concrete operational to formal operational thought.

All items which are in a multiple choice format are made up of a sentence stem and four alternatives. The answers are recorded by the subjects on a standard "trans-optic" answer sheet. The readibility for all directions and items was kept at or below the level of seventh grade. The test does not require rigid timing, as it is not one of speed of performance. Field testing indicated that one hour is a sufficient time for most to compete the 48 items. The test booklet is made up of 22 pages: (a) a front page, (b) a general instructions page, and (c) 20 pages of items. Forty percent of each test page is allocated to a line drawing that represents the problem in graphic form. This drawing is followed by multiple-choice items which relate to that drawing as shown in the example in Figure 1.

Whenever applicable, a basic problem is posed in the first item and an answer is elicited. Then the second item requires the selection of a category of explanations that the subject most closely associates with his/her answer. These responses were derived from the explanations subjects offered when presented with the original Inhelder and Piaget tasks in a clinical situation.

Scoring of the Objective Test

The scoring of the total test, the scoring of the subtests, and the item analysis were done by computer. The subtest scores were refined by applying the rule that a score of "1" would be assigned if, and only if, correct responses occurred on both items of an item pair which represented the solution to the presented problem and the choice of an explanation for that solution. If either item in the pair was incorrect, a score of "0" was assigned. This procedure was followed for the purpose of studying developmental trends in the individual's acquisition of the 7 formal schemata.



A local ice cream shop features a Do-It-Yourself-Sundae-Bar with five choices of toppings. The five toppings are: chocolate, fudge, strawberry, marshmallow and pineapple.

25. If you wanted to make a sundae using 3 different toppings, how many different kinds of sundaes could you prepare?

a) five types of sundaes;
b) eight types of sundaes;
c) ten types of sundaes;
d) fifteen types of sundaes.

26. How many different types of sundaes, each with a unique combination of toppings could you make using at least one topping on each sundae?

a) 10 different types;
b) 15 different types;
c) 31 different types;
d) 50 different types.

27. I chose my answer to question 26 for the following reason:

a) there are 10 possible pairs of toppings that I could use to make the sundae.
b) there are 10 possible pairs of toppings that I could use to make the sundae and then I could use each of the toppings by themselves for an additional five.
c) there are 31 different combinations that I can make if I have to use at least one topping each time.
d) There are many different combinations that I could make — at least fifty and so I chose this answer.

Figure 1. Sample items based on the combinations scheme.

Clinical Assessment

Six tasks were selected for use in the clinical interviews: (a) conservation of volume (at equal concentrations of matter) (Piaget and Inhelder, 1974, p. 48); (b) isolation of variables, the pendulum

(Inhelder and Piaget, 1958, p.62); (c) the quantification of probabilities (Piaget, 1971, p. 127f.); (d) combinatorial thinking, the colorless liquids problem (Inhelder and Piaget, 1958, p. 98), and a five-button electronic analogue of this problem adapted from Sills and Herron (1976); (e) proportional reasoning, the balance-beam problem (Inhelder and Piaget, 1958, p. 144); and "Mr. Big/Mr. Small," Form A (Karplus and Peterson, 1970); and (f) the coordination of two frames of reference, the snail and path problem (Piaget, 1971, p. 115). The protocols followed as closely as possible the descriptions of the original tasks (Arlin, 1982b).

Scoring of Clinical Tasks

Performance criteria were inferred from the descriptions of Inhelder and Piaget (1958) with respect to the behavioral characteristics of each substage in each task with the modification introduced by Martorano (1977) that "Stage IA" and "Stage IB" were combined into one level. The author scored the written protocols (transcripts from audio-tapes of the interviews) of each task for all subjects prior to any analysis of the objective test scores. A second person, who was well versed in Piaget's theory, independently scored the interviews. There was 71% agreement on exact substage placement, and all disagreements were within one substage of each other. The interrater reliability coefficient for the interview total scores was .86.

Procedures: Objective Test

The 38 subjects along with the other 356 recruits were tested in a familiar group testing situation by employing the standard procedures that were in use at each basic training center. The recruits were told that the test was experimental and that it would not be used in their general classification for positions after their basic training. It was stressed that it was important for them to make an effort on the examination, as the intent was to use the items as a part of a revised testing program for the general classification of future recruits.

The recruits were retested by using the same group test during the eight and final week of training. The first testing covered a six-week period. Retesting followed the same schedule. All testing sessions were supervised by the author.

Clinical Test

The recruits were told that names would be selected at random from the roster for follow-up interviews. These interviews would involve tasks similar to those on the test but in the interview situation they would be able to work directly with the materials of the problem. The recruits were free not to participate in either the group testing situation or in the follow-up interview. One recruit, selected for an interview, chose not to participate. This person was replaced by another whose name was taken at random from the roster.

These subjects were tested individually within one day of sitting for the group examination. There were 49 subjects who were interviewed. Thirty-eight were used for this study based on the selection criteria previously outlined. The rights of all of the individuals who took the group examination and who participated in the interviews were protected.

The interviews followed as closely as possible Inhelder and Piaget's protocols. The order of task presentation was varied randomly to control for order effects.

Data Analysis Methods

The psychometric characteristics of the group test and the developmental information which the test scores reveal have been reported elsewhere (Arlin, 1981a, 1982a). The analysis in this study is concerned with construct validation. An adaptation of the multitrait-multimethod procedure (Campbell and Fiske, 1959) was used. The clinical and paper-and-pencil forms of the formal operational tasks were regarded as two distinct methods. The correlation matrix included six traits (volume conservation; isolation of variables; probability; proportions; combinations and systems or frames of reference) and two methods (clinical, paper-and-pencil).

The two methods employed created some difficult problems in their comparison and in the construction of the reliability diagnosis. The paper-and-pencil test yields interval-type data for which Pearson correlations are appropriate. Hoyt estimates of reliability as well as test/retest reliabilities are also readily computed. The clinical tasks yield rank-order data which are based on judges' ranking of the subject's performance in terms of five levels from a trial-and-error description of performance to that of a high formal performance. Hoyt estimates of reliability, split-half reliabilities, and test/retest

reliabilities are basically inappropriate in this context. Therefore, the judgment was made to use inter-rater reliabilities in the reliability diagonal for the clinical data and Hoyt estimates of reliability in the diagonal for the objective test data. Finally, the subtest scores of the objective test were assigned to the same five levels as represented in the clinical data.

Findings

The major statistical outcomes are set forth in three tables. Thus, Table 1 summarizes the means and the standard deviations for the six traits by both methods.

Table 2 shows the multitrait-multimethod matrix.

The reliability diagonals in parentheses (Hoyt or Kendall as described earlier) show the same trait measured by the same method. The validity diagonal containing six validity coefficients (Kendall tau's) in brackets shows the degree of relationship beteen the six traits (formal schemata) measured by both methods. Convergent validity is supported by the size of the correlations. Discriminant validity is suggested because the values in the validity diagonal are higher than the values in the heterotrait-heteromethod triangles (formed by the dotted lines adjacent to the diagonal). Further the validity diagonal values are higher than those found in the heterotrait-monomethod triangles (indicated by solid lines). This outcome suggests that the trait variance is larger than the method variance—a finding that strengthens the case for discriminant validity.

TABLE 1
Means and Standard Deviations for Formal Operations Measures

Measures	Mean	(N = 38)	S.D.
(Objective Test)			
1. Volume Conservation	2.59		.61
2. Combinations	2.77		.91
3. Isolation of Variables	3.14		.82
4. Proportions	3.55		1.04
5. Probability	3.14		.93
6. Systems of Reference	2.68		1.57
(Clinical Tasks)			
1. Volume Conservation	2.95		.67
2. Combinations	3.18		1.04
3. Isolation of Variables	3.72		.92
4. Proportions	3.55		1.33
5. Probability	3.82		1.08
6. Systems of Reference	3.05		.94

TABLE 2
Multitrait-Multimethod Matrix of Formal Operations Measures

Traits	Method 1 (Paper-and-Pencil) 1.	2.	3.	4.	5.	6.	Method 2 (Clinical) 1.	2.	3.	4.	5.	6.
(Method 1)												
1. Volume Conservation	(.72)											
2. Combinations	.19	(.61)										
3. Isolation of Variables	.11	.23	(.85)									
4. Proportions	.22	.40	.26	(.68)								
5. Probability	.28	.34	.27	.46	(.62)							
6. Systems of Reference	.23	.35	.29	.39	.47	(.52)						
(Method 2)												
1. Volume Conservation	[.55]	.43	.29	.40	.28	.08	(.74)					
2. Combinations	.26	[.65]	.37	.55	.59	.33	.39	(.81)				
3. Isolation of Variables	.08	.22	[.64]	.43	.26	.27	.02	.31	(.97)			
4. Proportions	.30	.41	.46	[.74]	.55	.49	.10	.53	.55	(.85)		
5. Probability	.12	.43	.51	.50	[.60]	.29	.04	.54	.55	.51	(.94)	
6. Systems of Reference	.39	.30	.35	.39	.58	[.62]	.18	.28	.43	.37	.50	(.66)

TABLE 3
Validities of Traits in the Study of Formal Operations Assessments as Judged by the Heteromethod Comparisons

Trait	Validity	Highest Heterotrait Value	Number Higher
1. Volume	.55	.43	0
2. Combinations	.65	.59	0
3. Isolation of Variables	.64	.51	0
4. Proportions	.74	.55	0
5. Probability	.60	.59	0
6. Systems of Reference	.62	.58	0

To summarize the validation picture with respect to comparisons of validity values with other heteromethod values in each block, Table 3 has been prepared.

For each trait (formal scheme) and for each of the two methods, Table 3 presents the values of the validity diagonal coefficients between clinical and paper-pencil test measures for the same trait or scheme, the highest heterotrait value involving that trait, and the number out of the 16 heterotrait values which exceed the validity diagonal in magnitude. Inspection of Table 3 indicates a highly significant general level of validity. There is one trivial exception to the requirement that the validity diagonal exceed all others in its heteromethod block—that is, the correlation between probability and proportional reasoning.

Discussion

Although it might be desirable to have all the validity and reliability coefficients above .80 the multitrait-multimethod matrix indicates the possibility of developing an objective measure of Inhelder and Piaget's formal operational schemata that can validily and reliably parallel an individual's performance in the clinical setting. The matrix also provides a basis for test revision to improve specific subtests. One of the problems in constructing such an instrument is that the "traits," although treated as relatively independent of each other, are, in fact intercorrelated. The Piagetian model requires intercorrelations when it proposes a "structure d'ensemble" for each stage by which the various concepts, schemata, and operations are integrated and consolidated.

The Test of Formal Reasoning furnishes a gross assessment of levels, and its subtests afford diagnostic tools and measures for

curriculum development projects. Its value is in the provision of reliable and valid information about the cognitive levels of groups of individuals so that information about the cognitive levels and characteristics of large groups of students can be utilized for instructional and curricular decisions.

REFERENCES

Arlin, P. K. Performance time lags in the development of formal operations. Paper presented at the Annual Meeting of the American Education Association, Los Angeles, April, 1981. (a)

Arlin, P. K. Piagetian operations as predictors of reading and mathematics readiness in grade K-1 children. *Journal of Educational Psychology*, 1981, 73, 712–721. (b)

Arlin, P. K. Adolescent intellectual development: A revision of the Piagetian model and its instructional implications. *Journal of Applied Developmental Psychology*, 1982, in press. (a)

Arlin, P. K. An objective test of formal reasoning. Preliminary version (Unpublished), 1982. (b)

Campbell; D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81–105.

Epstein, H. T. and Toepfer, C. F. A neuroscience basis for reorganizing middle grades education. *Educational Leadership*, 1978, 35, 656–8, 660.

Epstein, H. T. Growth spurts during brain development: Implications for educational policy. In J. Chall (Ed.) *Education and the brain*, Chicago: University of Chicago Press, 1978.

Epstein, H. T. Correlated brain and intelligence development in humans. In M. Hahn (Ed.) *Development and evolution of brain size: behavioral implications*. New York: Academic Press, 1979.

Epstein, H. T. Learning to learn: Matching instruction to cognitive levels, *Principal*, 1981, 12, 25–30.

Fusco, E. Matching curriculum to students cognitive levels, *Educational Leadership*, 1981, 39, 47.

Gallagher, J. M. Problems in applying Piaget to reading. *Journal of Education: Boston University School of Education*, 1979, 161, 72–86.

Inhelder, B. and Piaget, J. *The growth of logical thinking from childhood to adolescence*. New York: Basic Books, 1958.

Karplus, R. and Peterson, R. W. Intellectual development beyond elementary school II: Ratio, a survey. *School Science and Mathematics*, 1970, 70, 398–406.

Martorano, S. C. A developmental analysis of performance on Piaget's formal operations tasks. *Developmental Psychology*, 1977, 13, 666–672.

Neimark, E. D. Current status of formal operations research. *Human Development*, 1979, 22, 60–67.

Piaget, J. *The child's conception of movement and speed.* New York, Ballantine Books, 1971.

Piaget, J. and Inhelder, B. *The Child's construction of quantities.* London: Routledge and Keegan Paul, 1974.

Shayer, M. and Wylam, H. The distribution of Piagetian stages of thinking in British middle and secondary school children: 14–16 year-olds and sex differentials, *British Journal of Educational Psychology*, 1978, 48, 62–70.

Shayer, M. Has Piaget's construct of formal operational thinking any utility? *British Journal of Educational Psychology*, 1979, 49, 265–276.

Sills, T. W. and Herron, J. D. Study of an elecronic analogue to the combinations of chemical bodies Piagetian task. *Journal of Genetic Psychology,* 1976, 129, 267–272.

Tomlinson-Keasey, C., Eisert, D. C., Kahle, L. R., Hardy-Brown, K. and Keasey, B. The structure of concrete operational thought. *Child Development*, 1979, 50, 1153–1163.