ED 358 138 TM 019 924

AUTHOR	Reckase, Mark D.
TITLE	Portfolio Assessment: A Theoretical Prediction of Measurement Properties.
PUB DATE	Apr 93
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Educational Assessment; *Estimation (Mathematics); Grade 12; High Schools; High School Students; Literature Reviews; *Portfolios (Background Materials); Predictive Measurement; Reliability; Scores; *Scoring; Statistical Distributions; Statistics; Student Evaluation; *Theories; Validity; *Writing Evaluation
IDENTIFIERS	Composite Scores; Large Scale Programs; *Model Development; *Performance Based Evaluation

ABSTRACT

In this non-experimental study, a model was developed for portfolio assessment based on definitions and applications in the assessment literature. This model describes portfolio components, scores to be computed, and uses to be made of the scores. The literature was then reviewed to find examples of actual applications that would provide realistic estimates of statistical characteristics of assessment results. Estimates of statistical characteristics were also obtained from performance assessments similar to portfolios. The model and estimates were then used to estimate characteristics of an operational large-scale portfolio assessment program. Estimates were obtained of reliability of results, score distribution characteristics, and the validity of the procedure for writing evaluations for 12th graders. The analysis suggests that a well-structured and carefully scored portfolio assessment has the potential to provide scores that meet standards of reliability required for use with individual students. By developing a score that is the sum of the scores on the various portfolio entries, useful discriminations can be made among students. A factor that has not been considered is the cost of obtaining the expected results, which would undoubtedly be high. One table presents analysis results, and one graph shows composite score reliabilities. (SLD)

Portfolio Assessment: A Theoretical Prediction of Measurement Properties[1]

Mark D. Reckase
ACT

Large scale assessment of educational attainment have been used since the about 1915 (for an example of an early assessment see Ayres (1915)), but it was only with the stimulus of the development of the Army Alpha (Yerkes, 1921) that the current mode of large scale assessment, the standardized multiple-choice test, began to evolve. The educational environment in the 1920s led to the development of assessment instruments that could (1) test large numbers of students at the same time, (2) be scored quickly and accurately, (3) sample broad domains of content, and, because of the educational philosophy of the times, (4) be based on indicators of performance rather than the performance itself.

This last point bears special mention because it is critical to understanding the current desire to revise the assessment system that is the result of a long evolutionary process. To make the point especially clearly, an example based on temperature will be presented first, and then the implications for educational assessment will be discussed. For most practical applications, such as turning the heat on and off in a house, temperature is not measured by assessing the energy in moving molecules of matter (the formal definition of temperature), but rather, is measured using either the height of a column of liquid or the expansion of pieces of metal. These measures are indicators of temperature, rather than direct measures of temperature itself. For most practical applications such measurement is acceptable, although in my house, since there is only one thermostat, some rooms are cooler than others. There are also times when the sun heats the wall by the thermostat, raising the temperature in that locale, and the rest of the house gets pretty cold. Thus, although using an indicator works most of the time, a single indicator has definite disadvantages. However, single indicators are often used because the alternative, many measures of molecular energy, is more costly than is merited by the practical application, controlling the temperature in the house. After all, for this application, if you get cold, you can simply change the temperature setting and turn on the heat. The entire house gets warmer. No one would recommend simply heating a thermometer without heating the rest of the house.

As with the thermometer, most educational assessments are indicators of educational attainment. They are not direct measures of educational attainment

[1] Paper presented at the meeting of the American Educational Research Association, Atlanta, GA, April 1993.

themselves. The goal of education is not to answer multiple-choice questions. However, these measuring instruments continue to be used because they provide useful information about the general state of learning. This is not to say that these measures cannot at times give misleading results. Similar to the thermostat in the sun, concentrated effort on the skills measured by a particular assessment can raise those skills while neglecting the other goals of education. In other words, the educational "house" can get cold even though the measuring instrument indicates that the house is warm. When this happens the solution is not to add more instruction related to what the test measures, but to heat the whole house -- improve instruction for all goals of education. Of course, as with the measurement of temperature, it is also possible to measure the goals of assessment directly, with performance assessments, but such assessments will be more complex and costly. The real question is whether multiple direct measures of performance are worth the cost.

Over the last few years, the educational environment has changed, forcing the evolution of new kinds of assessment instruments. Whereas previous measurement instruments were selected on the basis of their reliability, efficiency and domain coverage, the current educational environment has stimulated selection of assessment devices on the basis of instructional relevance. That is, the measuring instruments are required to be good models for instruction rather than just good indicators of attainment. Multiple-choice measures of writing skills have been determined to be poor models for instruction and assessments using samples of writing are selected as replacements because they are seen as better models of what students should be doing in class (Resnick & Resnick, 1992). Likewise, performance assessments in other areas are being developed and, in some cases, implemented (Berlack, Newman, Adams, Archbald, Burgess, Raven & Romberg, 1992).

Among the assessment procedures that are currently gaining favor because of their perceived instructional relevance is portfolio assessment. Although several definitions of portfolio assessment are available, the definition proposed by Meyer, Schuman & Angello (1990) seems to encompass all of the major requirements. The short form of their definition is as follows:

A portfolio is a purposeful collection of student work that exhibits to the student (and/or others) the student's efforts, progress or achievement in (a) given area(s). This collection must include:

- student participation in selection of portfolio content;
- the criteria for selection;
- the criteria for judging merit; and
- evidence of student self-reflection.

Portfolio assessments based on models that are consistent with this definition are beginning to be implemented on a large scale (e.g., Koretz, McCaffrey, Klein, Bell & Stecher, 1992). The reasons for the enthusiasm for portfolio assessment are that such assessments are (1) seen as using real examples of students' classroom work rather than surrogates (molecular energy rather than height of a column of liquid); (2) the assessments cover an extended period of time rather than a few hours; and (3) the portfolios are thought to stimulate good instructional practices. The first two of these reasons do have the potential to improve the content validity of educational assessment in the same way that measures of temperature from many locations in a house based on molecular motion or energy will give a much more thorough assessment of the temperature of a house. The question is whether the potential of these procedures can ever be achieved.

Given the projected use of porttolio assessment procedures for large scale assessment, the important question becomes how well can the portfolio assessment methodology be expected to serve the traditional functions of educational measurement. For the purposes of this paper, those functions will be assumed to be (1) the reporting of scores on individual students that can be used for (2) selection into special programs, (3) monitoring progress over time, and (4) evaluation of programs through aggregation of individual student results. To support these functions it is expected that the results of the portfolio assessment will have to meet the accepted standards for test reliability, that score distributions will have to exhibit characteristics that support the uses (e.g., sufficient spread in scores), and that the results will demonstrate both content and criterion related validity as required by each use. The remainder of this paper will consider the evidence that exists for each of these psychometric criteria.

Methodology

The methodology for this study was non-experimental. First, a model for portfolio assessment was developed based on the portfolio definition and applications that exist in the assessment literature. This model describes the component parts of the portfolio, the scores that would be computed, and the uses to be made of the scores. The model serves as the basis for the further evaluation of the assessment methodology.

Following the development of the model, the literature on portfolio assessment was reviewed to identify examples of actual applications of the methodology that could provide realistic estimates of the statistical characteristics of assessment results. Where empirical data could not be obtained from actual portfolios, estimates of statistical characteristics were obtained from other performance assessment procedures that were similar to portfolio assessments.

The portfolio model and the estimates of statistical characteristics were then used to estimate the characteristics of an operational, large scale portfolio assessment program. Where possible, ranges of values were reported to reflect the uncertainty of the estimates. In particular, estimates were obtained of the reliability of results, the score distribution characteristics, and the validity of the procedure for common applications of test scores.

The Portfolio Assessment Model

For the purposes of this paper, the portfolio assessment that will be considered will be in the area of writing assessment for twelfth grade students. This academic area has been selected because more work has been done to apply portfolio assessment methodology in writing than in any other content area, and twelfth grade assessments have many applications including program evaluation, career guidance, certification of competency, and college admissions and placement. Since the twelfth grade curriculum is very flexible, defining the model at this level will directly confront the issues of selection of materials from across the curriculum.

The portfolio assessment will be assumed to be a part of a statewide assessment system and the results of the portfolio will be aggregated and reported at the school, district, and state level. In addition, results will be reported to individual students and their parents. If the student desires the portfolio can be submitted to a university for use in placement into entry level college writing courses (e.g., the Miami University placement program, Black, Daiker, Sommers & Stygall, 1992).

To fulfill the promise of portfolio assessment as a methodology based on multiple measures, high instructional relevance, and high content validity, the portfolio will be composed of materials that have been selected jointly by the student and teachers to reflect the students work over the entire twelfth grade. All work, except for a self-reflective cover letter, will be taken directly from classroom activities. The portfolio will include not only final versions of the activities, but all of the drafts as well. All materials will be organized into the portfolio according to a specified table of contents. Students must provide one and only one work sample for each category. However, up until the point that portfolio is submitted, the selections can be changed.

in addition to the contents of the portfolio, a teacher who is familiar with the student's work will be asked to submit a written verification that the work was done as part of actual classroom activities. The table of contents for the portfolio is given below.

Table of Contents

I. A reflective letter to the readers of the portfolio telling why the particular materials were selected for inclusion.

II. A narrative or descriptive piece communicating a significant experience.

III. An explanatory, exploratory, or persuasive essay.

IV. A research paper.

V. An interpretive or evaluative response to a written text.

VI. The teacher's verification statement.

VII. Appendices with all previous drafts of parts II through V.

The materials for the portfolio may come from any course as long as the materials match the Table of Contents. The total number of pages included in parts I. through V. of the portfolio cannot exceed 25. However, the number of pages of text included in any section is up to the student. All work in parts I. through V. must be typed double spaced. To help the student and the teacher select materials for the portfolio, both a student guide and a teacher's guide will be available complete with examples of papers taken from previous portfolios. The guides will also include descriptions of the criteria used to score the portfolios.

Up to this point, the procedures used to score the portfolios have not been presented. A discussion of the scoring procedures and the expected results of the scoring will be the focus of the next section of this paper.

Analysis of Possible Portfolio Scoring Methods

The portfolio model that has been presented has five scoreable parts (I. to V. in the above list). This structure makes it possible to score the portfolio in a number of ways. Some existing portfolio assessment procedures score all of the pieces together to provide a single report of the results (Black et al, 1992; Moss, Beck, Ebbs, Matson, Muchmore, Steele & Taylor, 1992). Others scoring models involve scoring each piece, sometimes in a variety of ways, and then either combining the scores into a single total score (Nystrand, Cohen & Dowling, 1993), or reporting a profile (Koretz et al, 1992). For the hypothetical portfolio assessment presented here, the goal is to report as reliable score as possible so the results can be used at the individual level for placement and other important educational decisions.

Reliability

The literature on portfolio assessments does not provide much guidance on the selection of the scoring methods because so little has been published on the reliability of portfolio scores. Nystrand et al (1993) reported internal consistency reliabilities of portfolio scores in the mid .50s for total scores based on three papers and two readers. This study used portfolios from third year college students. Koretz et al (1992) reported average reliabilities of .43 the scores on five areas on the eighth grade writing portfolio from Vermont. If the scores were summed, the reliability of the total was expected to be about .58. In both of these cases, the rating process used four-point scales.

If portfolios are scored holistically as a complete entity, it is unlikely that the score reliability will be much higher than those reported in these two studies. Since students are not responding to common prompts, and because there can not be scoring guides designed for a particular piece of writing, the sources of task variation are bound to be large. This is the clear message communicated by such research studies Dunbar, Koretz & Hoover (1991) and Ruiz-Primo, Baxter & Shavelson (1993). However, if each piece in the portfolio is scored and the scores are summed to form a composite score, it may be possible to achieve levels of reliability that would support the use of the scores to inform decisions at the student level.

To determine how realistic it would be to expect composite scores on portfolios to reach acceptable levels for use with individual students -- say .80 -- the reliabilities of hypothetical composites were estimated using the formula for the reliability of battery composites given in Feldt & Brennan (1989). This formula gives the reliability of the composite score given the average reliability of the subtest scores and the average intercorrelation between the subtests. For the hypothetical portfolio being considered here, reliability values of .43 and .55 were considered and correlations of .16 and .28 were used. These values are consistent with those found in Koretz et al. (1992) and Nystrand et al. (1993). Figure 1 shows the expected level of the composite score reliability for portfolios with from one to ten scoreable entries.

Insert Figure 1 about here

The results presented in the figure show that to even approximate a reliability of .80 for a five entry portfolio like that hypothesized here will require at least a single entry internal consistency measure of at least .55 and a correlation between entries of .28. Given the single paper reliabilities summarized in Dunbar et al (1991), these values are higher than most that are obtained, but they are not unreasonable. A correlation of .28 was reported between the persuasive and summary entries in the portfolio analyzed by Nystrand et al (1993), but that was the highest correlation observed.

The level of correlation between entries that is required to achieve a reasonable composite reliability implies that the portfolio entries cannot be too disparate. Another way of saying this is that the entries need to measure the same thing. A goal of the portfolio assessment should be to get good domain coverage, but the domain should be well enough defined that reporting a single score makes sense. This is the same internal consistency assumption made for other kinds of assessment methods.

Score Scale

The number of score points used to rate the individual entries in the portfolio will determine the total number of points available for the composite score scale. Of course, the number of points actually used will depend on the ability of the readers of the portfolio entries to differentiate among levels of performance. Some simple statistical estimates of the possible score distribution for the hypothetical portfolio used in this paper will help make this issue more concrete. First some basic assumptions need to be specified.

Linn (1991) summarized the types of score scales used to score writing assessments used by a number of state testing programs in the United States. That summary indicates that most states use four to six point scales. Since the goal of the hypothetical portfolio is to provide information that can be used in a number of ways, a higher level of discrimination will be needed rather than a lower level. Therefore, a six point scoring scale will be assumed. It will also be assumed that each entry in the portfolio is read by two individuals and that the inter-rater correlation is .7. This is consistent with the inter-rater reliabilities summarized in Dunbar et al (1991).

In our experience at ACT, reasonable values for the mean and standard deviation of scores on a six point scale for a writing sample are 3.5 and .8, respectively. Similar values were obtained from the operational scoring of a writing sample for eighth grade students. Given these values as a starting point and the assumptions given above, what characteristics can be expected for the score distribution on the composite score for the five entry portfolio? Predictions of the characteristics can be obtained from the standard equations for the mean and variance of the sum of correlated variables.

First, the sum of the ratings of the two judges can be expected to have a distribution that has a mean of 7.0 (3.5 + 3.5) and a variance to 2.176. If the characteristics of the score distributions of the ratings of the five separate entries in the portfolio are assumed to be the same, the composite score can be defined as the sum of the scores on each of the entries, each of those being the sum of the scores of two judges. To compute the mean and variance of the composite score distribution, the scores for each of the entries will be assumed to be correlated .28 with each other, as was assumed above for the reliability estimation. The result of the estimation of the score distribution is a mean of 35 (basically, 2 x 5 ratings with means of 3.5) and a variance of 23.0656 (standard deviation of 4.8). Assuming the distribution of scores is approximately normal, this means that virtually all of the observed scores on the composite will fall between 21 and 49 even though the possible range of scores is 10 to 60. Of course, either higher inter-rater reliability, higher inter-entry correlations, or greater variance of initial judgements would increase the range of scores that would be expected to be obtained.

To put this value in perspective, the range from 21 to 49 is 28 score points. Assuming a standard deviation of scores of 4.8 and a reliability of .8, the standard error of measurement is 2.15. The 28 point range is roughly 13 standard error units. To put this in context, the 75 item ACT Assessment English test has a mean of 20.41 and standard deviation of 5.03 on the standard score scale, for the October, 1992 test date. The comparable score range would be 5 to 35, or 30 points. The standard error of that test, which has a reliability for that test date of .92, is 1.42. Thus, the 30 points is roughly 21 standard error units. Thus, the portfolio composite score would not allow the same level of differentiation of student performance as a high quality multiple choice test. This is not to imply that the two procedures are measuring the same thing, or that one is not preferred over the other for educational reasons, but only that the hypothetical portfolio presented in this paper will not likely allow the same level of distinctions to be made as one of the commonly used multiple-choice tests.

To show the sensitivity of the composite score distribution to the inter-rater reliability and the correlation between portfolio entries, the range defined by the mean plus or minus three standard deviations is shown in Table 1 for three values of inter-rater reliability (.6, .7, .8) and three levels of inter-entry correlation (.16, .28, .40). It is clear form the small variation in the numbers that the score distributions will not change dramatically with either greater inter-rater reliability, or greater homogeneity. However, obtaining a greater spread of scores from the original judges can have a quite dramatic effect. If the standard deviation of the original ratings increased from .8 to 1.0, the range for the case with .7 inter-rater reliability and .28 correlation would increase from 21-49 to 17-53. This implies that providing a scoring guide and training that results in a reasonable spread of initial ratings will be important if a composite score that makes full use of the score scale is desired.

Insert Table 1 about here

Predictive Validity

It is difficult to derive estimates of the expected predictive validity of the hypothetical portfolio for a number of reasons. First, no actual validity studies were found in the literature, so there is no way to determine reasonable values for validity coefficients. Second, the predictive validity of a test is specific to a particular use of a test. For example, the correlation between ACT Assessment English scores and college course grades range from the .10s to about .5 depending on the characteristics of the English course and the sample of students. Third, the magnitude of the validity coefficient is strongly related to the characteristics of the criterion measure, and many different measures can be selected. Yet, some of the literature on the use of writing samples for predicting performance in courses can give some hints concerning what might reasonably be expected.

One useful source of information is the work done to determine the usefulness of an essay as part of the revision to the SAT (Bridgeman, Hale, Lewis, Pollack & Wang, 1992). In their validity studies they found that scores on a twenty minute writing sample correlated in the .20s with English course grades. The correlations were only slightly lower than those obtained using multiple-choice scores. Given that the portfolio will be based on a more extensive sample of the writing content domain and that the materials being scored will be more similar to actual in-class activities for college courses, it is likely that the predictive validity will be somewhat higher than that observed in the SAT studies. It is also likely that the reliability of the five entry portfolio composite score will be higher than the score on the SAT Essay. Since the conclusion of the Bridgeman et al (1992) study was that the essay would increase the predictive power of the SAT, it seems reasonable that a score on a writing portfolio would also be a valuable addition to the information available for predicting course performance.

Of course, portfolio assessment is already being used for course placement at the University of Miami (Black, Daiker, Sommers & Stygall, 1992) with apparent satisfaction. And, many would argue that the more important impact of portfolio assessment is on the practice of instruction (Moss et al, 1992). It is clear that much more research is needed before any clear conclusions can be drawn about the level of validity of portfolio assessments.

Discussion and Conclusions

The purposes of this paper have been to review what is currently known about portfolio assessment, to present one hypothetical model for a portfolio assessment in writing, and to analyze that portfolio assessment model to determine what measurement characteristics can be expected from the procedure. From the analysis presented here, it seems that a well structured and carefully scored portfolio assessment has the potential to provide scores that meet the standards for reliability required for use with individual students. Further, by developing a score that is the sum of the scores on the various entries in the portfolio, useful discriminations can be made among students. The scores developed in this way are likely to have sufficient predictive validity that they will at least be a useful adjunct to more traditional measures. In all of these regards, the results of this analysis has been fairly positive. However, one factor that has not yet been considered is the cost of obtaining the expected results.

The model that was presented in this paper required that each of the five entries in the portfolio be read by two individuals. Typically, there would be a third reading if the initial readings did not result in scores differed by more than a point. The limit on the number of pages in the portfolio was specified as 25. If the reading rate for the readers is optimistically estimated as a page per minute, it would take one reader 25 minutes to read all of the materials. Two readings would take about 50 person minutes. It is likely that with all of the shuffling of papers, the scoring time for the portfolio should be estimated as a person hour. Since the persons reading the portfolios will likely have to be well educated and well trained, it is not likely that such persons will be willing to work for minimum wage. Adding the cost of shipping and data processing, a total cost of scoring of $10.00 does not seem unreasonable. Note that the cost of scoring the University of Miami portfolio is about $17.50 per portfolio. Is it likely that someone will be willing to pay $10.00 to have a portfolio scored? I don't know the answer to this question, but without careful scoring, the measurement properties that have been described will not likely be achieved.

Alternatively, portfolio assessment can be reserved for formative evaluation in the classroom, with emphasis on the instructional uses. High levels of reliability would not be attained, but that may not be important for that application. This has been suggested by Moss et al (1992). Or, portfolio assessment might be used in place of traditional standardized tests. Then, all costs for the tests could be put into the scoring. It should be noted, however, that the cost of $10.00 per portfolio mentioned above is only for one content area. If portfolios were used to assess all content areas, that cost would be multiplied by the number of areas to be assessed.

The analyses in this paper suggest that it may be possible to produce a portfolio assessment procedure that meets the current standards of psychometric quality. However, if such a procedure were to be implemented in place of current procedures for high stakes assessment, it would be a very expensive alternative.

References

Ayres, L. P. (1915). *A measuring scale for ability in spelling.* New York, NY: Russell Sage Foundation.

Berlack, H., Newman, F. M., Adams, E., Archbald, D. A., Burgess, T., Raven, J. & Romberg, T. A. (1992). *Toward a new science of educational testing & assessment.* Albany, NY: State University of New York Press.

Black, L., Daiker, D. A., Sommers, J. & Stygall, G. (1992). *Handbook of Writing Portfolio Assessment: A Program for College Placement.* Oxford, OH: Miami University.

Bridgeman, B., Hale, G. A., Lewis, C., Pollack, J. & Wang, M. (1992, May). *Placement validity of a prototype SAT with an essay (RR-92-28).* Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4(4),* 289-303.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.) *Educational Measurement (3rd Ed.).* New York: American Council on Education/Macmillan.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992, December). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program: Interim report.* ???, CA: RAND Institute on Education and Training.

Linn, R. L. (1991). *Cross-state comparability of judgements of student writing: results from the New Standards Project (CSE Technical Report 335).* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Meyer, C., Schuman, S., & Angello, N. (1990, September). *NWEA White Paper on Aggregating Portfolio Data.* Lake Oswego, OR: Northwest Evaluation Association.

Moss, P. A., Beck, J. S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11(3),* 12-21.

Nystrand, M., Cohen, A. S. & Dowling, N. M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1(1),* 53-70.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: new tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.) *Changing assessments: alternative views of aptitude, achievement and instruction.* Boston: Kluwer.

Ruiz-Primo, M. A., Baxter, G. P. & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30(1),* 41-53.

Yerkes, R. M. (1921). Psychological examining in the United States Army. *Memoirs of the National Academy of Sciences, 15.*

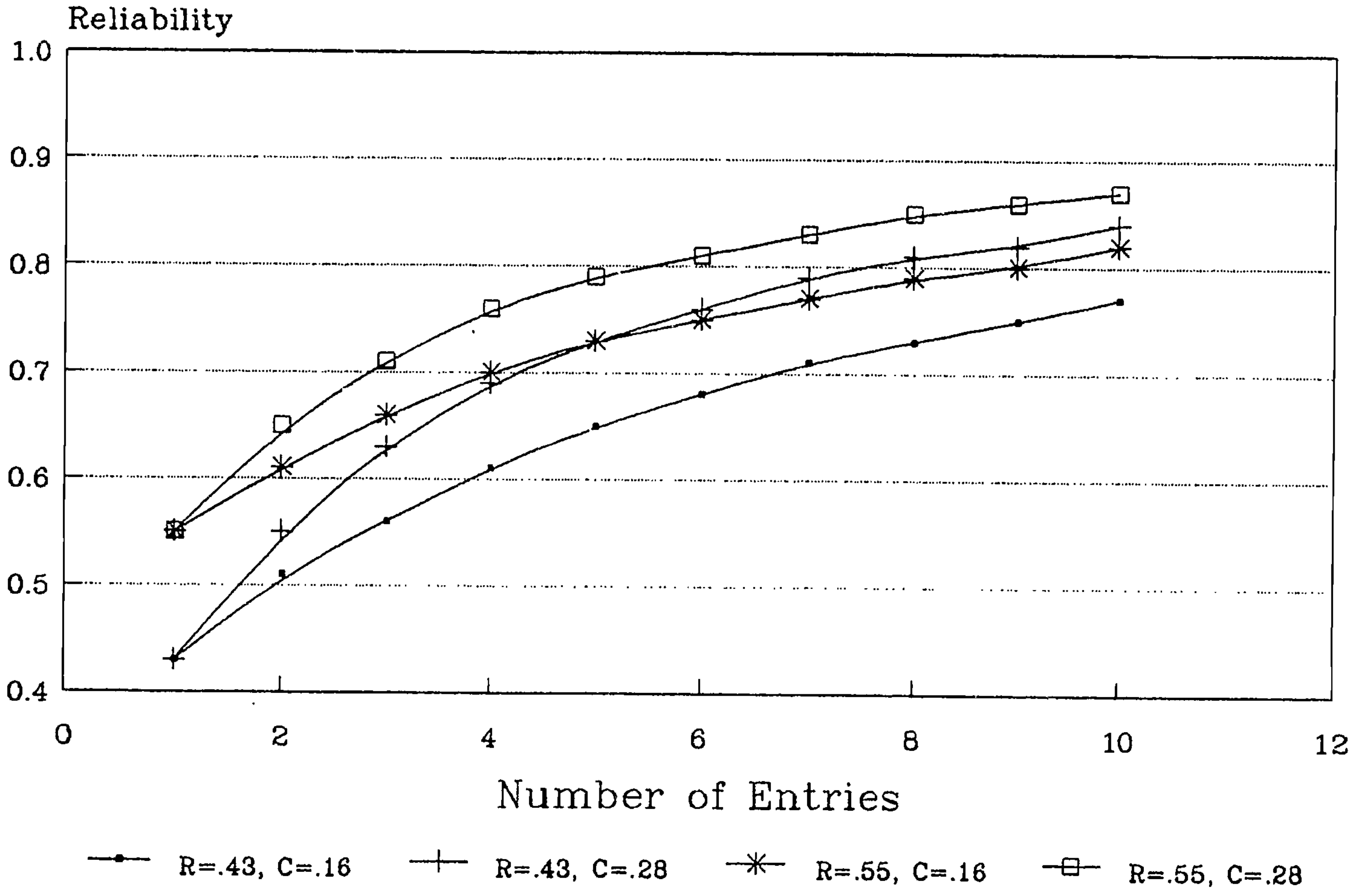
Table 1

Composite Score Range Given by Mean ± 3SD for Combination of Inter-rater Reliability and Inter-entry Correlations

Inter-rater Reliability	Correlation		
	.16	**.28**	**.40**
.5	23-47	22-48	22-48
.7	21-49	21-49	20-50
.9	20-50	19-51	18-52

Composite Score Reliability

Given Task Reliability and Correlation


Reliability
Number of Entries
R=.43, C=.16
R=.43, C=.28
R=.55, C=.16
R=.55, C=.28