Test-Free Person Measurement with the Rasch Simple Logistic Model

Howard E. A. Tinsley
Southern Illinois University at Carbondale

René V. Dawis
University of Minnesota

This research investigated the use of the Rasch simple logistic model in obtaining test-free ability estimates. Two tests each of word, picture, symbol, and number analogies were administered to college and high school students. Differences between scores on each pair of tests were analyzed to determine whether the ability estimates were independent of the tests employed. The results indicate that raw-score ability estimates are influenced by the difficulty of the items used in measurement but that Rasch ability estimates are relatively independent of the difficulty of these items. The need is discussed for additional research in which an individualized item-presentation procedure is used with the Rasch model.

Whenever an individual completes two tests measuring the same ability in a short period of time, the difference in the obtained ability estimates is expected to be close to zero. It is not uncommon, however, for ability estimates to differ substantially under such circumstances. In addition to errors of measurement, the difficulty levels of the tests contribute to discrepancies in ability estimates.

Rasch (1960) has proposed a simple logistic model for ability tests which, theoretically, allows the independent estimation of item difficulty and person ability. Ability estimates expressed on a common scale can be obtained from any set of calibrated items which fit the Rasch model, even when the same items are not administered to all subjects. This makes possible the use of "tailored" or "adaptive" tests in which only those items appropriate for examinees of a given ability level are administered. In short, the Rasch simple logistic model makes possible what Wright (1968) has called *test-free person measurement.* If these claims are substantiated, tests based on the Rasch model would represent a marked improvement over tests based on classical psychometric theory.

Several investigators have studied the use of the model for item calibration (Anderson, Kearney, & Everett, 1968; Brooks, 1965; Rasch, 1960; Tinsley & Dawis, 1975; Wright, 1968). The work of Wright (1968) is, however, the only research the present authors were able to find which investigates whether the model leads to test-free person measurement. Wright's research was based on the responses of 976 first-year law students to 48 reading-comprehension items in the Law School Admission Test. He divided the test into subtests containing the 24 easiest items and the 24 hardest items. Next, he calculated the raw score and Rasch ability estimate for each subject on the two tests. Finally, he calculated the difference between the two raw scores and the difference between the two Rasch ability estimates.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 4 Fall 1977 pp. 483–487

Wright (1968) used two criteria in evaluating the scoring procedures. The first criterion stated that the distribution of differences in the ability estimates should be unbiased (i.e., the mean difference should be approximately zero). The distribution of differences for raw scores ($\overline{X} = 6.78$, $S = 3.30$) was almost entirely above zero, while the distribution of differences in Rasch ability estimates ($\overline{X} = .061$, $S = .749$) was centered around zero (Wright, 1968). It appears, therefore, that raw scores are biased estimates of ability but that Rasch ability estimates are unbiased estimates which are not influenced by the level of test difficulty.

The second criterion used by Wright was that the distribution of differences in the ability estimates should have a small standard error. For each individual, Wright divided the difference between the two Rasch ability estimates by the measurement error of this difference. This produced what he called the distribution of standardized differences with a mean of .003 and a standard deviation of 1.014. Wright concluded from these data that the only variation observed in Rasch ability estimates was of the same magnitude as that expected from the standard error of measurement in the test and that these data support the claim that the Rasch simple logistic model allows the measurement of a person with any set of calibrated items.

The research of Wright (1968) has apparently not been replicated. Thus, the generalizability of these findings remains in question. The purpose of this research was to investigate the invariance of Rasch ability estimates across tests of differing difficulty levels when tests measuring a different ability domain and employing a different item format are used.

Method

Selection of Item Format

Spearman's *g* or general mental ability seems to be represented in almost all the major intelligence tests in use today. Helmstadter (1964) points out that tests dealing with abstract relationships (such as verbal, numerical, or symbolic analogies) come closest to representing what is meant by *g*. For this reason, the analogy format was selected for study here. Guilford (1959) suggests that there are several different methods of asking analogy questions (i.e., figurally, symbolically, semantically, and behaviorally) depending upon the type of material used to present the question. To make the present research as general as possible, figural (picture), symbolic (number and symbol), and semantic (word) test items were used. Two types of symbolic material were used because of intrinsic differences in the two and because Guilford (1966) has reported the discovery of more than one factor in some cells in his Structure-of-Intellect.

Instruments

Two analogy test booklets were utilized in this study. Each test booklet began with one standard page of instructions. The first booklet contained a 60-item, word-analogy test followed by a 50-item, symbol-analogy test. The second booklet contained a 60-item, number-analogy test followed by a 50-item, picture-analogy test. All items were multiple-choice with five response alternatives. Although the Rasch model assumes that guessing is not a factor influencing the probability of a correct response, the applicability of the Rasch model to multiple-choice items has been previously demonstrated (Tinsley & Dawis, 1975). None of the tests employed time limits. (For a discussion of the test construction process, see Tinsley, 1972.)

Subjects

Data were obtained from college students enrolled in an introductory psychology class and from students enrolled in two suburban high schools. The college students were volunteers who participated in the research to earn additional points toward their grade. The high school students were enrolled in the classes of teachers who volunteered to participate in the study. Each high school student completed only one test booklet and was allowed only one 50-minute class period in which to do so. College students were given the option of completing one or both booklets and were under no time restraints. The data from both groups of students were combined for analysis.

Analysis

Before formal analysis was begun, the data were edited to eliminate careless or slow examinees and those examinees for whom guessing may have been a factor in determining their ability estimate. The former was accomplished by eliminating any examinee who left several consecutive items blank, left the last few items in a test blank, or left more than five items in the entire test booklet blank. Almost all of the subjects eliminated for this reason were high school students who had been unable to complete the test booklet in the 50 minutes allowed. The exclusion of subjects for whom guessing may have been a factor was accomplished for any given test by eliminating those respondents whose "total score" was equal to or lower than the following index recommended by Panchapakesan (1969, p. 115):

$$r = k/m + 2\,(k(m-1)/m^2)^{1/2}$$

where k = the number of items

m = the number of response alternatives

Note that k/m is the expected score when responding is purely on the basis of chance and $k(m-1)/m^2$ is the variance of the chance score. Therefore, subjects were eliminated from this study if the probability was greater than .023 that their score could have been obtained by chance.

Since the question under investigation requires that two tests of the same ability be administered to each subject, each test was divided into two subtests with one subtest containing the difficult items and the other the easy items. For each of the four pairs of subtests, those respondents who received a perfect raw score on either subtest were eliminated because of indeterminant ability estimates. The total number of word-, picture-, number-, and symbol-analogy tests dropped from the analysis for the above reasons was 87, 69, 79, and 118, respectively.

The data from each of the four pairs of subtests were then analyzed in the following manner: First, Rasch item-easiness estimates were calculated using a common scale for the items in both subtests. Next, the subtest raw scores and Rasch ability estimates were computed for each subject (see Panchapakesan, 1969, and Wright & Panchapakesan, 1969, for a detailed discussion of the procedure used in calculating the Rasch ability estimates). For each individual, the difference between the two subtest scores was then calculated. Finally, the mean and standard deviation of the difference scores were computed for the raw scores and Rasch ability estimates.

Results

Table 1 shows the mean and standard deviation of the differences in subtest scores. The

Table 1

Mean and Standard Deviation of Differences in Subtest Scores

Ability Estimate	Analogy Test Word (N = 865)	Picture (N = 590)	Number (N = 580)	Symbol (N = 834)
Raw Scores	10.14 ± 3.357	10.42 ± 3.001	12.55 ± 3.760	9.25 ± 2.951
Rasch	.047 ± .696	.094 ± .733	.197 ± .901	.038 ± .916

mean difference in raw scores ranged from 9.25 for symbol analogies to 12.55 for number analogies with the mean varying between 3.0 and 3.5 standard deviations above zero. The mean difference in Rasch ability estimates was .047 and .036 for word and symbol analogies and .094 and .197 for picture and number analogies.

Discussion

One of the most promising features of the Rasch model is that it permits individualization of measurement. Once an item pool has been developed and calibrated on a common scale, individuals need complete only a subset of items appropriate to their ability level. Their scores can then be converted to ability estimates on a common scale. This means that the scores of two individuals can be compared, even if the tests they completed did not have one single item in common. It was with this aspect of the Rasch model that this research was concerned.

This research investigated the hypothesis that raw scores and Rasch ability estimates are invariant with respect to the items used in measurement. The results indicate that raw-score ability estimates are biased in that they are influenced by the difficulty of the items used in measurement. The Rasch ability estimates, however, appear to be relatively independent of the difficulty of the items used in measurement. Accordingly, these data are in agreement with the theoretically derived position that an individual's Rasch ability estimate is independent of the difficulty of the items used in measurement and of the ability of the individuals used to calibrate the test.

These data represent the strongest evidence to date of the robustness and generality of the Rasch model. The robustness of this model for the assumption that guessing is not a factor influencing the probability of a correct response is demonstrated by the fact that the model performs as predicted when applied to multiple-choice items which allow the respondent to guess the correct answer. Although the use of Panchapakesan's (1969) *r* index eliminated persons from the study if the probability that their score was obtained entirely by chance was greater than .023, it seems very likely that the score of most persons retained in the study was influenced to some degree by chance. The generality of the Rasch model is demonstrated by the fact that semantic, symbolic, and figural tests of analogical reasoning were investigated with largely uniform results. The Rasch ability estimates appeared to be relatively independent of the difficulty of the items in each instance.

These data are consistent with the conclusion that the Rasch model yields relatively test-free person measurement. Clearly, however, further research on this issue is desirable. A more precise test of this feature of the model would require a procedure for the individualized administration of items. The goal of the Rasch model is to measure each individual as accurately as possible. Measurement precision depends on the number of items administered and the appropriateness of these items to the ability level of the examinee (Panchapakesan, 1969). For the Rasch model to maximize the precision of measurement, each examinee should be administered only those items appropriate to his/her ability level. Thus, the Rasch scoring procedure requires some mechanism for "tailoring" the test to the individual. Under such circumstances, the advantages of the Rasch procedure over ability estimates derived from classical psychometric theory would be maximized. Research dealing with these issues would be a logical follow-up of this investigation.

References

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. *The British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 231–238.

Brooks, R. D. *An empirical investigation of the Rasch ratio-scale model for item difficulty indexes.* (Doctoral dissertation, University of Iowa) Ann Arbor, MI: University Microfilms, 1965, No. 65–434.

Guilford, J. P. Three faces of intellect. *American Psychologist*, 1959, *14*, 469–479.

Guilford, J. P. Intelligence: 1965 model. *American Psychologist*, 1966, *21*, 20–26.

Helmstadter, G. C. *Principles of psychological measurement.* New York: Appleton-Century-Crofts, 1964.

Panchapakesan, N. *The simple logistic model and mental measurement.* Unpublished doctoral dissertation, University of Chicago, 1969.

Rasch, G. *Probability models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research, 1960.

Tinsley, H. E. A. *An investigation of the Rasch simple logistic model for tests of intelligence or attainment.* (Doctoral dissertation, University of Minnesota) Ann Arbor, MI: University Microfilms, 1972, No. 72–14387.

Tinsley, H. E. A. & Dawis, R. V. An investigation of the Rasch simple logistic model: Sample-free item and test calibration. *Educational and Psychological Measurement,* 1975, *35,* 325–339.

Wright, B. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service, 1968, 85–101.

Wright, B., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 1969, *29,* 23–48.

Acknowledgements

This research is based on Tinsley's doctoral dissertation (for which Dawis was principal advisor) and was supported by grant number N00014-68-A-0141-0003 from the Office of Naval Research to the Center for the Study of Organizational Performance and Human Effectiveness and by grants from the Minnesota Research Coordinating Unit for Vocational Education and the University of Minnesota Department of Psychology.

Author's Address

Howard E. A. Tinsley, Department of Psychology, Southern Illinois University, Carbondale, IL 62901.