Behavior Research Methods
2007, 39 (3), 570-573

Using the Rasch model to quantify the causal effect of test instructions

Ana R. Delgado
Universidad de Salamanca, Salamanca, Spain

This experimental study was designed to quantify, by means of the Rasch model (RM), the effects of three instruction/scoring conditions on student measures and on the reliability of an achievement multiple-choice test in a field context. Examinees performed the test in one of three conditions which differed only in the instructions provided. Predictions regarding performance indicators were fulfilled, and the expected differences in reliability favoring omission-inducing instructions did appear. This difference in reliability was found for both Rasch and raw data and thus it can be concluded that the fact that results from previous studies failed to corroborate this prediction must have been due to the lack of important consequences of test scores for the students. The RM has served to neatly quantify the differences between instructions promoting guessing and instructions promoting omission under uncertainty, showing that the recommendation to omit is not only educationally but also psychometrically sound.

A test score is a summary of the evidence contained in the answers to the items of a test that are related to the construct being measured. In addition to the number-correct sum, there are various algorithms to compute scores, such as punishing errors by discounting what is expected from random guessing or offering a small reward for omissions; they all share the goal of optimizing the use of the evidence provided by the item answers to infer the construct level (Thissen & Wainer, 2001).

Number right (S_1) and the formula that tries to counterbalance the effects of random guessing by imposing a penalty of $1/(k-1)$ for each error, where k is the number of response options (S_2), are the commonly used procedures in the psychometric context. A different strategy to discourage guessing is to promote omissiveness by giving a bonus of $1/k$ for each omission (S_3) (Traub, Hambleton & Singh, 1969). Both formula-scores are linearly related by $S_3 = [N + (k - 1)S_2]/k$, where S means score and N is the number of items composing the test. Traub and Hambleton (1972) carried out a between-subjects experiment in which examinees under S_1 were informed that their scores would be the number of correct answers and that they should try to answer all questions, guessing when necessary, examinees under S_2 were informed that they should try to answer all questions, guessing when necessary, but were threatened with a small penalty for wrong answers, and examinees under S_3 were discouraged from guessing, being promised a small reward for omitted questions. Results indicated that the effect of instructions on omissiveness was large, with those associated to S_3 being more effective than those associated to S_2, but that, contrary to expected, mathematical reasoning scores under S_1 instructions were the most reliable. From a theoretical point of view this is an anomalous result—reliability is inversely related to random error and thus guessing with impunity under S_1 should be decremental. Neither did Prieto and Delgado (1999), in an experiment carried out 25 years later, find the expected increment in score reliability under S_3 even though their predictions regarding some performance indicators—right, wrong and omitted answers—were mostly fulfilled.

Budescu and Bar-Hillel (1993) compared S_1, S_2, and S_3 in three dimensions: strategic, (how should a rational decision device respond to maximize its score), psychological, (the actual way examinees respond), and psychometric (concerning score reliability and validity). From the strategic point of view, answering is never too detrimental for the examinees, even when they guess at random. Psychometricians have tried to discourage guessing by using formula-score rules, mainly S_2 (although, in theory, S_3, with the least variance due to random guessing, should give more reliable scores). However, from a psychological point of view, examinees do not always properly understand some implications of the instructions. Budescu and Bar-Hillel (1993) concluded that S_1, by encouraging examinees to always answer would remove the advantage of the risk-taking ones (Budescu & Bar-Hillel, 1993).

From a technical point of view, test scoring is a statistical enterprise (Thissen & Wainer, 2001). But is it just that? It cannot be so, given the educational implications of some instructions necessarily associated with the use of the scoring algorithms. Prieto and Delgado (1999) pointed to a fourth dimension in which S_1, S_2, and S_3 could be compared: the educational one. From an educational point of view, both S_1 and S_2 can be accused of promoting guessing in examinations—S_1 encourages

A. R. Delgado, adelgado@usal.es

all sorts of guessing, and S_2 discourages only random guessing—which no doubt can be seen as an undesirable habit (Thorndike, 1971). In addition, S_2 spreads the false *belief in the law of small numbers* (Tversky & Kahneman, 1971) by implicitly suggesting that, when guessing at random, right and wrong answers will cancel each other out which may be true for large data matrices, but is uncertain for the three or four item answers from one single student. Under S_3, examinees can know exactly what bonus they will receive from the number of questions omitted, no risks are taken, and thus Prieto and Delgado (1999) proposed, adding to Traub and Hambleton (1972), to favor S_3, even though its expected advantage in reliability failed to appear.

Should the hypothesis of a better reliability for scores under S_3 be considered falsified? There are two methodological aspects that could be improved before answering this question. First, it might well be that the lack of important consequences for the examinees in the above experiments would have decreased the effect of instructions on answers. Second, test scores such as those described have only ordinal justification, which may distort the causal effect of instructions. Both aspects were taken into account in this study by carrying out a field experiment and analyzing its results by means of the Rasch model (Rasch, 1960, 1968; Wright, & Stone, 1979).

The Rasch model (RM) for binary data gives the probability that person n passes item i, P_{ni}, as follows: $P_{ni} = \exp(\beta_n - \delta_i)/(1 + \exp[\beta_n - \delta_i])$, where β_n means the person construct level and δ_i is the item location (Rasch, 1960, 1968).

Logits are the units defined by the RM. A person's ability in logits is the natural log odds for succeeding on items of the kind chosen to define the scale zero (which is conventionally set to the difficulty mean). An item's difficulty in logits is the natural log odds for failure on that item by persons with abilities at the scale zero. Data fit to the Rasch model can be evaluated by summing the standardized square of residuals after fitting the model over persons or items to form approximate chi-square-distributed variables (*infit*, weighted by item Fisher information, and *outfit*, unweighted and thus very sensible to outliers). The Rasch model is statistically strong. Its parameter estimators are sufficient, consistent, efficient, and unbiased (Andersen, 1970, 1973; Rasch, 1968).

In testing practice, the RM is typically used as a mathematical model to make post hoc statistical adjustments in test scores for item difficulties when matching or randomization is impossible (van der Linden & Hambleton, 1997). However, because of its desirable metric properties, the RM can also be fruitfully used in the context of other psychological research designs, such as when directly modeling experimental data (Verguts, de Boeck, & Ruts, 1998) or, as it is proposed here, to quantify the results of a between-subjects experiment in which control techniques such as manipulation and elimination are also present. It would seem that measurement has not been a common worry among experimental psychologists, even though metric considerations are in the base of some relevant methodological problems in current research (Embretson, 2006).

This study was designed to quantify the causal effect of instructions associated with scoring procedures S_1, S_2, and S_3 on a measure with consequences in real life (an examination). In order to control for any other possible source of differences between groups, experimental and statistical control strategies were included.

METHOD

Participants

A total of 534 psychology students took the Research Methods multiple-choice test described below as part of their course examinations. In 2004, 2005, and 2006, students took the test under S_3 (n = 170), S_1 (n = 188), and S_2 (n = 176), respectively. Only students getting the test for the first time have been included in this sample. Due to ethical considerations, experimental groups were not formed by random assignment, but rather consisted of natural groups in successive years, which can be thought of as random samples from the same population. In each of the groups median age was 20 years (arithmetic mean: 20.6, 20.7, and 20.9 years), and proportion of female students was .87, .87, and .82, respectively.

Instruments

The measurement instrument was an achievement multiple-choice test composed of forty-five items assessing basic-level knowledge of Research Methods in Psychology that the students had to pass in order to get the course credits. Following recent advice on nonfunctioning options, three-option items were constructed. Empirical deleting of the worst option has been shown not to affect the reliability of test scores in a significant way (Cizek & O'Day, 1994), and three-option item tests have been shown to function similarly to four-option item ones (Delgado & Prieto, 1998; Rogers & Harley, 1999). Haladyna and Downing (1993) have even suggested that three options per item may be, under many circumstances, a *natural limit* for multiple-choice item writers.

Procedure

Under S_3, examinees were informed that, in order to discourage guessing, there would be a reward for omissions of 1/3; the second group, under S_1, was advised to guess given that there would be no penalty for errors; and under S_2, examinees were informed that there would be a penalty for errors of 1/2 in order to discourage random guessing. A standardized test application followed with a one-hour time limit. Control over the questionnaires was strict so that all of them were given back to the professor at the end of the exam together with the answer sheet.

Tests were marked by means of an optical reading device and the dichotomous responses of each separate group were Rasch-modeled with Quest 2.1 for Macintosh (Adams & Khoo, 1996). (Students were not marked according to the Rasch-modeled scores, but according to the scoring rule associated to the instruction under which they got the test.)

RESULTS

First of all, it had to be corroborated that the instructions had been properly understood by the examinees. Table 1 shows clear differences between experimental conditions in the average number of omitted, wrong and right answers: under instructions promoting guessing (S_1) there were fewer omissions, and more right and wrong answers, than under instructions which do not promote random guessing (S_2 and S_3). No significant differences were found between S_2 and S_3 in omissions, right and wrong answers (Scheffé F tests were .55, .01, and .15, respectively; $p < .01$).

Table 1
Means and Standard Deviations for Test Performance Indicators by Instructions, and *F* Tests for Instructions

	S_1 (188)		S_2 (176)		S_3 (170)		
	M	*SD*	*M*	*SD*	*M*	*SD*	*F*(2,531)
Omitted	0.42	1.54	7.39	3.88	7.76	4.12	32.00**
Wrong	13.50	5.26	8.41	4.45	8.13	4.59	72.70**
Right	31.07	5.28	29.20	6.25	29.10	6.43	6.27**

Note—A posteriori comparisons indicated that mean differences between S_2 and S_3 are not significant. $^{**}p < .01$.

Results in Table 1 replicated previous results and corroborated that examinees had followed the instructions. The next step was to model the answers of the examinees in the three groups to the same set of 45 items by means of the RM, which allows the measure of both items and persons in the same logit continuum. A summary of item estimates and fit indices under the three conditions can be seen in Table 2. A repeated measures ANOVA on the parameter item estimates under S_1, S_2, and S_3 yields an $F(2,88) = 0.0001$, $p = .9999$. No item got zero or perfect scores and item fit was adequate considering both *infit* and *outfit* values. The absence of differences in item properties under different experimental conditions was expected both from the mathematical properties of the RM (item parameter estimation is independent from the calibration sample and the scale zero is set at the mean difficulty level so that scale origin is the same for the three groups) and from the experimental design (the same items were applied to samples that can be considered as random samples from the same population). Items were very reliably estimated and thus *separation*—the ratio of adjusted item *SD* to average item standard error, analogous to the Fisher discriminant ratio (Fisher, 1936)—was large, allowing at least seven difficulty *strata* to be differentiated in the test, that is to say, seven difficulty levels separated by three errors of measurement (Linacre, 2002).

Case estimates and fit indices can be seen in Table 3. No person got zero or perfect scores and mean case fit was also adequate. Given the difference in right answers between instructions, case level in the construct should be higher under S_1, which it is: The difference between S_1 and S_3 means is 0.24 logits, and between S_1 and S_2 is 0.23 logits. The difference in logits between S_2 and S_3 is too small to be considered. This is, in a field context and in nonarbitrary units, the causal effect of instructions on Rasch-modeled measures of knowledge of research methods. If preferred, it can as well be translated to conventional language (e.g., the standardized difference between S_1 and S_3 is $d = 0.35$).

Table 2
Rasch-Modeled Items—Summary by Instructions

	S_1 (188)	S_2 (176)	S_3 (170)
SD	0.99	1.19	1.27
Adj. *SD*	0.97	1.17	1.25
Infit Mnsq	1.00	1.00	1.00
Infit *SD*	0.07	0.09	0.10
Outfit Mnsq	0.97	1.00	0.97
Outfit *SD*	0.16	0.22	0.21
Reliability of estimate	0.96	0.97	0.97
Separation	5.02	5.83	5.84

Note—Conventionally, item mean has been set to zero.

As to the causal effect of instructions on reliability of estimates: Is reliability higher under S_3 than under S_1? Results indicate that it actually is ($z = 1.87$, $p < .05$, one-tailed). No differences can be found between S_2 and S_3. This implies that person separation (the ratio of adjusted person *SD* to average case standard error) is also significantly larger under S_3, which is desirable if between-person comparisons were to be made. In addition to the model reliability estimates, Quest 2.1 provides (Cronbach's alpha) internal consistency reliability estimates for raw data: .71, .80, and .81 under S_1, S_2, and S_3, respectively. It can be seen that these values are nearly identical to the model reliability estimates, and thus the difference in reliability of raw scores under S_3 and S_1 is again significant, favoring S_3 ($z = 2.24$, $p < .05$, one-tailed).

DISCUSSION

Under instructions promoting guessing (S_1) there was hardly any omission, contrary to what happened under instructions that do not promote random guessing (S_2 and S_3), where omissions were on average about one sixth of the test. Right and wrong answers were in the range of what would be predicted if guessing were really at random: rounding down, under S_1 examinees get 2 more right and 5 more wrong answers than examinees under S_2 and S_3, adding up to 7 answers, which is the mean number of omissions under S_2 and S_3. No significant differences were found between S_2 and S_3. It seems that Psychology students do understand the implications of the instructions, and thus their behavior is close to that of a rational device (the strategic point of view). This adds to the internal validity of the experiment, although it is possible that it affects external validity, given that not every examinee behaves in such a rational way (psychological and strategic dimensions do not always coincide). In any case, it is clear that, at least when consequences are important, students informed of the scoring rules can rationally follow instructions.

Results concerning item parameter estimates were as expected from the RM as well as from the experimental design. Results concerning right answers were neatly reflected in the case Rasch-modeled measures. Examinees' mean level in the construct was higher under S_1 than under S_2 and S_3 (about one third of a student standard deviation), and the difference in logits between S_2 and S_3 was practically nought. It seems that both S_2 and S_3 get closely equivalent measures in a field context, and therefore, given the already mentioned educational advantages of S_3 over S_2, it seems more interesting to focus on the comparison between person measures under S_1 and S_3.

This study did find a significant effect of instructions on reliability, which was larger under S_3 than under S_1, as theoretically expected if all guessing would have been at random. Even though three-option item tests have been shown to function similarly to four-option item ones (Delgado & Prieto, 1998; Rogers & Harley, 1999), thus far the

Table 3
Rasch-Modeled Cases—Summary by Instructions

	S_1 (188)	S_2 (176)	S_3 (170)
Mean	1.04	0.81	0.80
SD	0.72	0.83	0.85
Adj. *SD*	0.62	0.74	0.77
Infit Mnsq	1.00	0.99	0.99
Infit *SD*	0.13	0.14	0.14
Outfit Mnsq	0.97	1.00	0.97
Outfit *SD*	0.25	0.35	0.33
Reliability of estimate	0.73	0.80	0.81
Separation	1.64	1.98	2.04

effect of guess-minimizing instructions on reliability has only been demonstrated for three-option items. This difference in reliability was found for both Rasch and raw data and thus it can be concluded that the fact that results from previous studies failed to corroborate this prediction must have been due to the lack of important consequences of test scores for the students.

Finally, the RM has served to neatly quantify the differences between instructions promoting guessing and instructions promoting omission, showing that the recommendation to omit is not only educationally but also psychometrically sound. It must be noted that even though parametric model-based methods are typically used in the social and behavioral sciences, most variables in these sciences do not have interval justification—raw scores are not measures. The RM satisfies the ordinal conditions of Luce and Tukey (1964) for conjoint measurement, i.e., consistent ordering of subjects by items and consistent ordering of items by subjects. Sufficient fit of the RM to item responses indicates that the probability of a response can be expressed as an additive function of a person parameter and an item parameter. Thus the logistic transformation serves as a basis for an interval scale on which persons (and simultaneously items) can be located. Subjects are assigned interval-scale scores on the latent variable that are monotonically related to raw scores, yet have more desirable properties, such as sample independence, extreme score unbiasedness, and linearity (Long, Feng, & Cliff, 2003).

AUTHOR NOTE

I thank John H. Krantz and an anonymous reviewer for their precise and useful suggestions. Correspondence concerning this article should be sent to A. R. Delgado, Departamento de Psicología Básica, Psicobiología y Metodología, Avenida de la Merced, 109-131, 37005 Salamanca, Spain (e-mail: adelgado@usal.es).

REFERENCES

Adams, R. J., & Khoo, S. T. (1996). *Quest-2.1. The interactive test analysis system* [Computer software]. Camberwell: ACER.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society*, **32**, 283-301.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, **38**, 123-140.

Budescu, D., & Bar-Hillel, M. (1993.) To guess or not to guess: A decision theoretic view of formula scoring. *Journal of Educational Measurement*, **30**, 277-291.

Cizek, G. J., & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational & Psychological Measurement*, **54**, 861-872.

Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, **14**, 197-201.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist*, **61**, 50-55.

Fisher R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational & Psychological Measurement*, **53**, 999-1010.

Linacre, J. M. (2002). Number of person or item strata. *Rasch Measurement Transactions*, **16**, 888.

Long, J. D., Feng, D., & Cliff, N. (2003). Ordinal analysis of behavioral data. In J. A. Schinka & W. F. Velicer (Vol. Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 635-661). Hoboken, NJ: Wiley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, **1**, 1-27.

Prieto, G., & Delgado, A. R. (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment*, **15**, 143-150.

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

Rasch, G. (1968). *A mathematical theory of objectivity and its consequences for model construction*. Report from the European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational & Psychological Measurement*, **59**, 234-247.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

Thorndike, R. L. (Ed.) (1971). *Educational measurement*. Washington, DC: American Council on Education.

Traub, R. E., & Hambleton, R. K. (1972). The effect of scoring instructions and degree of speedness on the validity and reliability of multiple-choice tests. *Educational & Psychological Measurement*, **32**, 737-758.

Traub, R. E., Hambleton, R. K., & Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational & Psychological Measurement*, **29**, 847-861.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **76**, 105-110.

van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer.

Verguts, T., de Boeck, P., & Ruts, W. (1998). Analyzing experimental data using the Rasch model. *Behavior Research Methods, Instruments, & Computers*, **30**, 501-505.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

(Manuscript received March 20, 2006;
revision accepted for publication June 23, 2006.)