Item-Rest Regressions, Item Response Functions, and the Relation Between Test Forms

Dato N. M. de Gruijter
University of Leiden

John H. A. L. de Jong
Dutch Institute for Educational Measurement (CITO)

Levine (1982) used item-rest regressions for the estimation of item parameters, and this relationship was exploited in this research in the context of vertical equating. Results from a simulation and an empirical dataset were used to demonstrate that item-rest regressions were useful in verifying the relationship between two tests obtained from item parameter estimates. It is shown that in vertical equating designs the Rasch model cannot replicate the relationship between tests at the lower score levels when guessing is present. At higher score levels, however, the correct transformation function can be estimated, irrespective of the IRT model used. *Index terms: equating, guessing parameter, item response functions, item-rest regression, Rasch model.*

In item response theory (IRT), it is postulated that the regressions of item responses on latent ability—the item response functions (IRFs)—have a special form. The item parameters that determine the form of the IRFs can be estimated by unconditional maximum likelihood (UML), marginal maximum likelihood (MML), or—in the case of the Rasch model—conditional maximum likelihood (CML).

The model IRFs can be compared with the empirical regression of item responses on estimated ability (Kingston & Dorans, 1985). Empirical item-ability regressions are related to item-test and item-rest regressions, to the extent that estimated ability correlates with total test score. In this paper, the relation between item-test regression and UML is discussed. Next, the potential use of item-rest regressions for vertical equating is considered. A simulation study is reported, and the use of item-rest regressions is demonstrated in a vertical equating study.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 1, March 1991, pp. 25-34

0146-6216/91/010025-10$1.75

UML and Item-Test Regressions

One of the IRT models is the three-parameter logistic model. For this model, the probability of a correct response to item i given ability (θ) can be written as

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-a_i(\theta - b_i)]} \,, \qquad (1)$$

where b_i is the difficulty parameter of item i, a_i is the discrimination parameter, and c_i is the lower asymptote or pseudo-guessing parameter. Other models are special cases of this model. The Rasch model is obtained when $a_i = 1$ and $c_i = 0$ for all items.

There is a direct relationship in the Rasch model between IRFs obtained with UML and item-test regressions. The estimation equations for the item parameters can be written as

$$\sum_{k=1}^{n-1} N_k[p_{ik} - P_i(\hat{\theta}_k)] = 0, \; i = 1, \ldots, n, \qquad (2)$$

with

n = the number of items,

N_k = the number of examinees with a total score of k,

p_{ik} = the proportion correct on item i given total score k,

θ_k = estimated θ for examinees with a total score of k.

The proportions p_{ik} do not depend on the θ distribution in the Rasch model. It is clear from Equation 2 that weighted differences between item-test regressions (p_{ik}) and model item-ability regressions are minimized.

For $n = 2$, Andersen (1973) discovered that the item parameter estimates obtained with Equation 2 and similar equations for person parameters have a fixed bias. Wright and Douglas (1977) considered a generalization of Andersen's result to $n > 2$. However, it is obvious from Equation 2 that the bias depends on the marginal score distribution N_k ($k = 1, \ldots, n - 1$) (Andrich, 1989; de Gruijter, 1990; Divgi, 1986, 1989). For larger values of n, the bias is small and generally inconsequential.

Due to the sufficiency of the total score for θ, item-test regressions from Equation 2 are equal to item-$\hat{\theta}$ regressions, after a transformation of the total score scale. Given the overall validity of the model, item-test regressions contain useful information.

The invariance of p_{ik} in the Rasch model can be exploited for the estimation of item parameters. This leads to the well-known unbiased CML estimation procedure. Item-test regressions p_{ik} also play a role in item bias detection, based on the Mantel-Haenszel statistic (Holland & Thayer, 1986). Use of this statistic on the basis of item-test regressions has a sound rationale, given the validity of the Rasch model.

The total score in other IRT models is not a sufficient statistic for θ, although total scores might be used for the estimation of θ (e.g., Yen, 1984). Use of nonoptimal total scores (Lord, 1980, p. 74) can be defended for other than statistical reasons.

With respect to item-test regressions, other models differ strongly from the Rasch model. The relation between item and test score is spurious, due to the fact that the item is part of the test; therefore, item-test regressions do not truthfully reflect the relation between items. Even uncorrelated items result in item-test regressions with a positive slope (Lord, 1980, p. 30).

Item-Rest Regressions

Item-test regressions provide an inherently biased picture of the strength of the relationship between the item and θ, unless the Rasch model is valid. The use of item-rest regressions instead of item-test regressions eliminates this problem. Item-rest regressions also contain information on IRFs. Lord (1970) demonstrated this in a study in which item-rest regressions were obtained, the score scales were transformed, and the regressions were compared to IRFs, based on ML estimation of parameters. The two approaches gave results that were in agreement.

Urry (1976) used item-rest regressions for the estimation of item parameters with the assumption of normally distributed θ. Levine (1982) argued that the item-rest regression for long tests is close to the regression of the item on the true score of the rest test. On the basis of item-rest regressions, item-item curves (p_{ir}, p_{jr}), where r denotes the rest score, can be constructed that reflect the invariant relation between two items contained in their IRFs. Levine constructed item-item curves from item-rest regressions using lower asymptotes $\hat{c}$ from a LOGIST (Mislevy & Stocking, 1989) analysis, and used these curves for the estimation of b parameters for the items. The estimates were close to estimates from LOGIST.

The relation between item-rest regressions and IRT suggests their usefulness in certain contexts in which IRT is used, such as equating and scaling. In vertical equating, for example, the relationship between tests unequal in difficulty must be estimated. Long and reliable tests allow for the use of item-rest regressions. Let p_{ir} be the item-rest regression of item i on rest score r, ignoring the fact that p_{ir} is based on a different rest test for each i. The relationship between two tests containing items $i = 1, \ldots, n_1$ and $j = n_1 + 1, \ldots, n_1 + n_2$, respectively, can be approximated by the set of pairs (y_r, z_r) for $r = 1, \ldots, n_1 + n_2 - 1$, with

$$y_r = \sum_{i=1}^{n_1} p_{ir}/n_1 \quad , \tag{3}$$

and

$$z_r = \sum_{i=n_1+1}^{n_1+n_2} p_{ir}/n_2 \quad . \tag{4}$$

The resulting relationship can be compared with the one obtained by an IRT analysis.

The most interesting case arises when the scaling with item-rest regressions differs from IRT scaling. A large difference is a strong indication that the IRT model is inappropriate. In an analysis with the Rasch model, for example, the resulting relation between the two subtests can differ from the one obtained with the nonparametric method (Divgi, 1981) based on item-rest regressions, especially at the lower end of the scales y and z. In this case, the conclusion is that the one-parameter model is too simple to reflect the relation between the subtests.

A Simulation

To demonstrate the utility of the item-rest regression approach in the presence of guessing, data were simulated for 2,000 hypothetical examinees using the model in Equation 1 on two non-overlapping tests: a "low-level" test and a "high-level" test. All cs were set to .25, and all as = 1. The low-level test consisted of 20 easy items with bs in the range -1 to 0, and the high-level test had 60 difficult items with bs in the range 0 to 1. The examinees were sampled from an N(0,1) distribution. Item-rest regressions were computed according to the suggestion of Levine (1982):

$$P_{ir} = N_{ir(+)}/N_{ir} \tag{5}$$

where

$N_{ir(+)}$ = the number of examinees answering item i correctly, and $2r$ or $2r - 1$ other items correctly, and

N_{ir} = the total number of examinees answering $2r$ or $2r - 1$ other items correctly.

Rest scores were then grouped: $r = 1$ corresponded to rest scores 1 and 2, $r = 2$ corresponded to rest scores 3 and 4, and so forth. The number of score groups was reduced to n_g by combining score categories until N_{ir} exceeded 20 for all i.

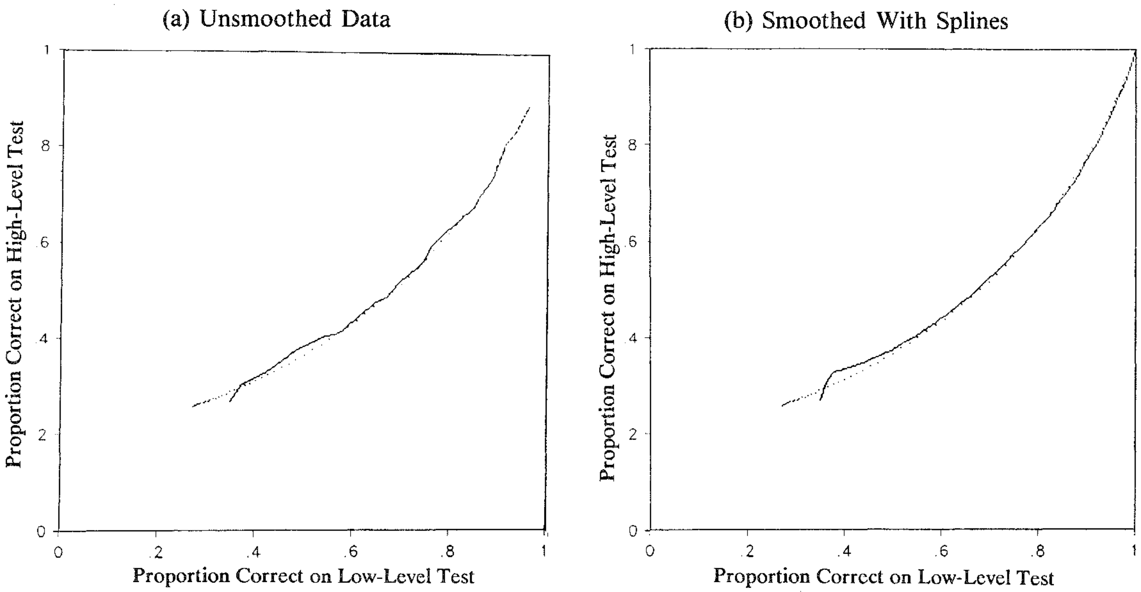
The resulting regressions exhibited fluctuations. It was assumed that the item-θ regressions were monotonically increasing, which seems reasonable with most IRT models, except possibly at low θ levels. To smooth the regressions, monotone regressions were computed. If for two consecutive score groups r and $r + 1$, $p_{ir} > p_{ir+1}$, both groups were combined with $N_{ir(+)} = N_{ir(+)} + N_{ir+1(+)}$ and $N_{ir} = N_{ir} + N_{ir+1}$, and the number of score groups was decreased by 1. This process was repeated until all reversions were eliminated. Finally, the original number of score groups was restored, with tied values p for scores r that were combined in some stage of the previous step. After the adjustments of the regressions, the proportions correct y_r and z_r ($r = 1, \ldots, n_g$) were computed.

Figure 1a displays the results for the simulation. The true relationship was recovered adequately, except at the lower end of the scales. The downward bend at the lower end of the curve reflects the fact that there were some examinees with low θ levels who performed lower than the minimum expected value c on the high-level test. The curve in Figure 1a relating both tests is rather irregular, and smoothing (Fairbank, 1987) was indicated. The empirical curve was smoothed with cubic B splines using NAG subroutines E02BAF and E02BBF (Numerical Algorithms Group, 1987) after the point (1,1) was added with a large weight. The smoothed curve in Figure 1b resulted.

Figure 1 indicates why vertical equating with the Rasch model, which has no guessing parameter, frequently gives unsatisfactory results in the case of guessing. When guessing is present, the Rasch model cannot replicate the relationship between the tests at the lower score levels. The choice of a model without a guessing parameter is of lesser consequence in horizontal equating where the relationship between tests deviates less from a straight line through the origin.

An IRT approach based on a model with a guessing parameter seems to be more adequate than a simple regression approach. A deviation like that in Figure 1b is not to be expected in a LOGIST analysis, where some information on the

Figure 1
Relationship Between Low-Level and High-Level Tests Based on
Item-Rest Regressions (Dotted Line Is True Relation, Solid Line Is Observed Relation)



(a) Unsmoothed Data

(b) Smoothed With Splines

lower asymptote is present at all θ levels. Moreover, low scoring examinees give estimation problems with short tests (Samejima, 1973) and are often removed from the UML approach in LOGIST.

In many UML analyses, $\hat{c}$s are less than the reciprocal of the number of alternatives, which suggests a negative bias (Lord, 1983a). Lord investigated the bias of item parameter estimates, and obtained small negative biases for c. However, his study was based on fixed θ parameters. The determination of the lower asymptotes was a problem in the Harris and Hoover (1987) study. Their Figure 1 showed that different equatings of two tests differed at the lower ends of the scales. Apparently, the lower asymptotes on the difficult test were lower for the younger age groups, which suggests a downward effect on cs. Harris and Hoover argued that multidimensionality might have played a role, apart from item parameter estimation problems.

The same data on which Figure 1 was based were analyzed with LOGIST using a common starting value $\hat{c} = .2$. The data in the example were so extreme with respect to LOGIST that $\hat{c}$ remained fixed at .2. As can be seen from Figure 2, the LOGIST analysis resulted in a smoothed relation between the tests that seemed quite adequate, notwithstanding the discrepancy between the true value c and the estimated value. If the IRT analysis had differed too much from the item-rest results, other starting values for item parameters might have been tried.

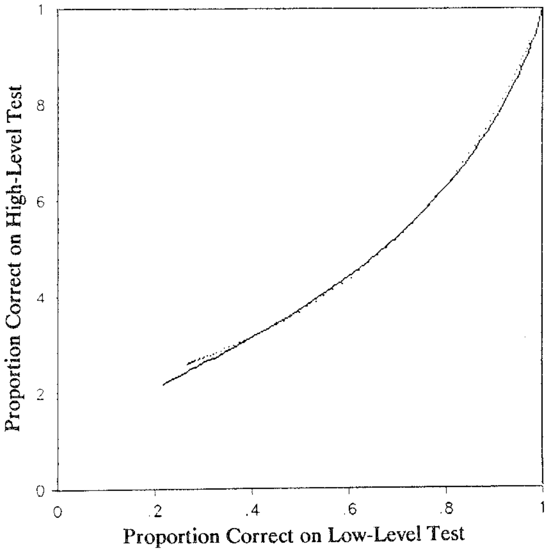
An Empirical Example

Method

To further illustrate the usefulness of the item-rest regression approach, data were used from the administration of two tests of listening comprehension of English as a foreign language on 2,261 examinees. Examinees were sampled from a population of about 80,000 students in secondary education in the Netherlands who had been taught in a mixed-ability program.

Secondary school graduation examinations in the Netherlands consist of an external and an internal part. Foreign language examinations cover

Figure 2
Relationship Between the Two Tests in Figure 1, Based on a LOGIST Run With $c = .2$ (Dotted Line Is True Relation, Solid Line Is Observed Relation)



reading and listening comprehension, writing, and speaking. Reading comprehension is assessed in the external examination, and the remaining skills are assessed by the schools. Most schools (about 95%) use tests that are provided by CITO to assess listening comprehension; therefore, the CITO listening comprehension tests function as national standardized tests.

The examination procedure provides for a school-based assessment of all students at two distinct levels, irrespective of their proficiency and their expected level of graduation. Each student receives two grades, corresponding to the two levels within the program (de Jong, 1986a, 1986b). On the basis of these grades, students may choose to take the external examination at either level.

Hence, students base the choice of the level of the external examination and their prospective graduation level for each subject separately on the results at both levels in the internal assessment procedure. Once students have made a decision, the results of the internal assessment at the level not selected are discarded: only the results at the selected level are taken into account.

Because students need to be assessed at two distinct θ levels, one simple procedure would be to produce a test with two cutoff scores—one for either level. However, measurement error at both levels must be reduced to a minimum, because students' choices of graduation level are based on their test results. A test designed for both levels would need to include a number of easy items, which would lead to ceiling effects for the higher-level students. Also, in order to discriminate effectively among higher-level students, an adequate number of difficult items would need to be included in the test, which would be expected to lead to guessing behavior among the lower-level students. Because students guess only if they feel they cannot solve an item, guessing will typically lead to higher scores for students whose expected scores are well below chance level, rather than for students who are expected to score at, or just below, the chance level.

Such nonmonotonicity (Lord, 1980, pp. 17–19) is sometimes referred to as a "dip" or "valley" in the observed item response function. Several authors have proposed models to account for this phenomenon (Choppin, 1983; Lord, 1983b; Thissen & Steinberg, 1984). However, the practical implication of such models is that some students would be penalized for scoring higher than other students, because the correct answer of the lower-ability students to some items is assumed to be due to guessing. The solution of Thissen and Steinberg (1984, p. 518) that rejects the use of such items in practical testing is the only correct solution. Therefore, two separate tests are used for the examinations at the termination of the mixed ability program. Each test is designed to yield maximum information at one of the intended θ levels.

The tests used here were the CITO tests of listening comprehension used in the 1989 examinations for English as a foreign language. The low-level test consisted of a 28-item easy subtest designed specifically for the lower ability levels, and a 17-item anchor subtest, which yielded a total test of 45 items. The high-level test consisted of a 31-item difficult subtest focused at the higher ability levels, and the 17-item anchor subtest, which yielded a total of 48 items. The common anchor was introduced primarily to reduce administration time, because all students have to take both tests.

All items were multiple-choice with either two or three options. The number of options was limited in order to avoid an undue emphasis on reading skill and/or memory factors. Naturally, the restriction of the number of response alternatives raised the chance level. However, the potentially substantial influence of guessing was avoided by designing the average item difficulty in each test to yield a probability of a correct response halfway between the chance score and the maximum score for the average student in each of the target populations.

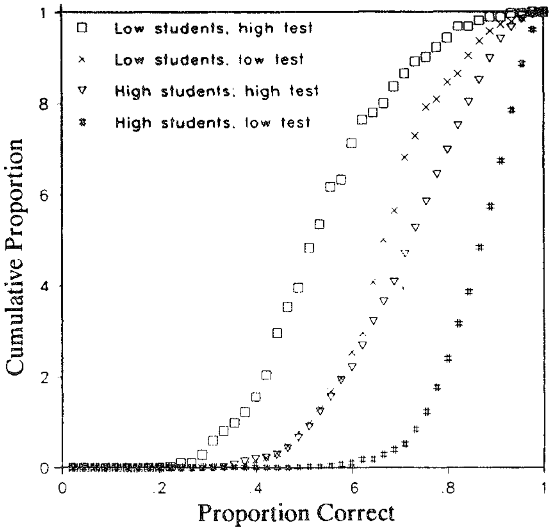
Thus the cutoff, which was between the 20th and 30th percentile in the target student distribution for both tests, was located well above chance level. The tests were pilot-tested to adjust their difficulty level and check their internal consistency. Items within each test were selected to fit a unidimensional IRT model (i.e., the one-parameter Rasch model; for a more detailed description of the item selection procedure, see de Jong, 1986b).

Results

Figure 3 shows the observed cumulative distributions of students from both ability levels on either test as a function of score level. Because the students opting for graduation at the lower level constituted only about 20% of the total population, their distributions are somewhat less smooth than those of the higher-level students. It is clear from Figure 3, however, that a substantial difference is revealed by both tests between the two ability levels, and that both tests are of similar difficulty for the respective target populations.

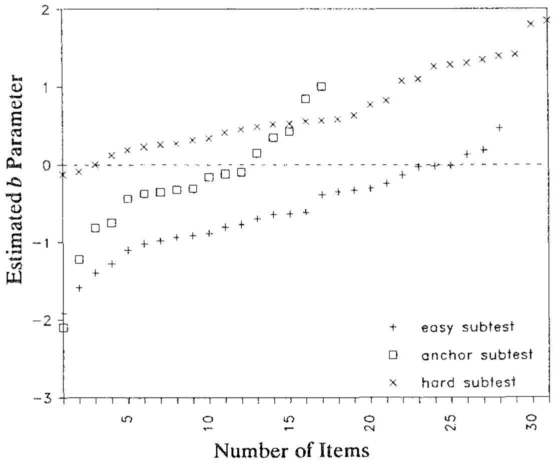
An interesting and relevant question in the context of the mixed-ability program is whether the data will allow for vertical equating of the two tests. If scores obtained on either test could be transformed reliably to scores on the other test, there would be no need to administer two separate tests.

Figure 3
Cumulative Proportion of Lower- and Higher-Ability Students on Low- and High-Level Tests as a Function of Proportion-Correct Scores



Item parameters according to the Rasch model were estimated using the complete data matrix of responses of all students to all items in both tests. Figure 4 shows the distribution of the *b* parameter estimates for the subtests. The *b*s of

Figure 4
Estimated Rasch *b* Parameters for Easy, Difficult, and Anchor Subtests for Items Arranged in Order of Difficulty



the 28 items in the easy subtest were estimated to lie in the range of –1.918 to .465, with an average of –.707 and a standard deviation (SD) of .581. For the 31 items in the difficult subtest, *b*s were estimated in the range of –.125 to 1.842, with an average of .696 and SD = .534. The estimated *b*s of the 17 items in the anchor subtest ranged from –2.100 to 1.001, with an average of –.103 and SD = .568. Adding subtests at both levels, the average *b* parameter was estimated at –.479, with SD = .646 in the low-level test, and at .413 with SD = .667 in the high-level test (see Table 1).

Table 1
Mean and Standard Deviation (SD) of Rasch Item Parameters and Three-Parameter Logistic Model Item Parameters in the Total Set and in Different Subtests (All $\hat{c}$s = .25, Except for One Item With $\hat{c}$ = .30)

Statistic	1PL $\hat{b}$	3PL $\hat{a}$	 $\hat{b}$
Total Set of Items (76 Items)			
Mean	.000	.683	-1.006
SD	.835	.210	.959
Easy Subtest (28 Items)			
Mean	-.707	.743	-1.659
SD	.581	.184	.466
Anchor Subtest (17 Items)			
Mean	-.103	.505	-1.441
SD	.568	.118	.835
Difficult Subtest (31 Items)			
Mean	.696	.728	-.177
SD	.534	.219	.708
Low-Level Test: Easy + Anchor (45 Items)			
Mean	-.479	.653	-1.577
SD	.646	.199	.640
High-Level Test: Difficult + Anchor (48 Items)			
Mean	.413	.649	-.625
SD	.667	.217	.967

Figure 5 displays the relationship between the two tests, based on item-rest regressions and on the Rasch parameter estimates. It shows that the Rasch model cannot correctly replicate the relationship between the tests at the lower score levels, and thus the findings of the simulation presented in Figure 1 are confirmed.

Based on the same data matrix, item parameters were estimated according to the three-parameter model using LOGIST. The starting values for $\hat{c}$s were set at .2. Figure 6 shows that the distribution of the estimated *b* parameters over the subtests is similar, except for a scale shift. The estimated *a*s for the 45 items

Figure 5
Relationship Between Proportion Correct on Low- and High-Level Tests Based on Item-Rest Regression (Solid Line) and on Rasch Parameter Estimates (Dotted Line)

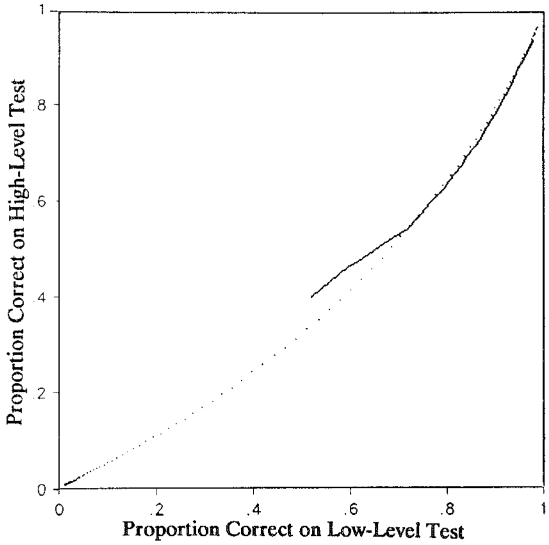
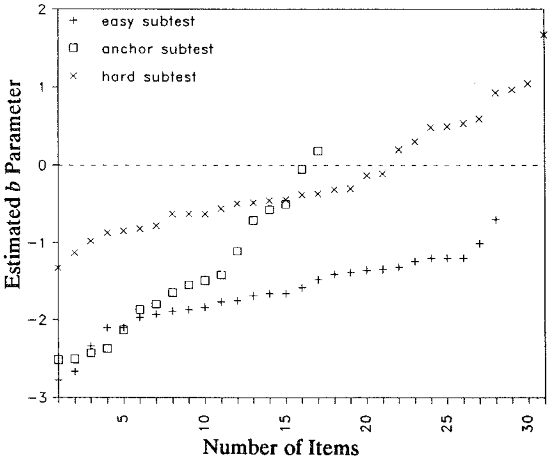


Figure 6
Estimated Three-Parameter Logistic Model *b* Parameters for Easy, Difficult, and Anchor Subtests for Items Arranged in Order of Difficulty

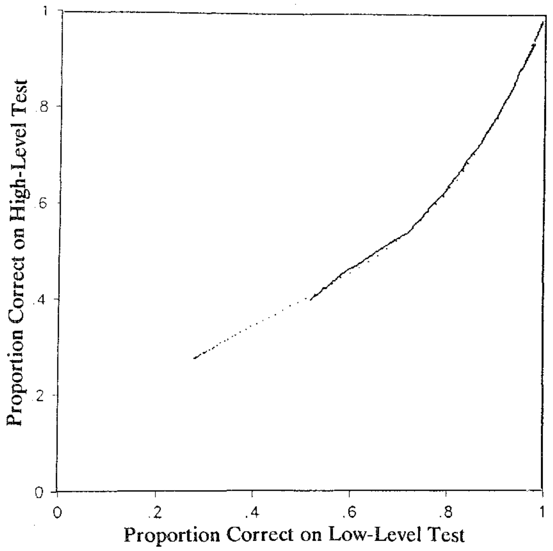


in the low-level test were in the range of .317 to 1.121, with an average of .653 and SD of .199; estimated *b* parameters ranged from –2.783 to .175, with mean of –1.577 and SD = .64. For the 48 items in the high-level test, the estimated *a* parameters ranged from .317 to 1.123, with mean of .649 and SD = .217; estimated *b* parameters ranged from –2.669 to 1.666, with mean of –.625 and SD = .967.

For all items in both the low-level and the high-level test, *c* parameters were estimated at .25, except for one item in the high-level test, which was estimated at .3. The results did not suggest a relevant difference between the two tests, with respect to the *a* and *c* parameters. The $\hat{c}$s in both tests were well below the expected chance level of .5 in the anchor subtest (two-option items) and .33 in the easy and difficult subtests (three-option items).

Figure 7 displays the relationship between the two tests based on item-rest regressions and the three-parameter logistic model item parameter estimates. As in the simulation (Figure 2), the three-parameter model provided a better replication of the relationship between the two tests than did the Rasch model, because the occurrence of some guessing behavior of lower-level students on the high-level test was reflected in the three-parameter model.

Figure 7
Relationship Between Low- and High-Level Tests Based on Item-Rest Regression (Solid Line) and on Three-Parameter Logistic Model Parameter Estimates (Dotted Line)



Discussion

Item-rest regressions contain information on the empirical relationship between a set of items and ability. Item-rest regressions of items from different subtests may be combined in order to clarify the relationship between these subtests. With total tests of appropriate length, this relationship can be recovered satisfactorily, except possibly at the lower end of the scale.

The relationship observed from item-rest regressions could be used in test-equating designs. The most appropriate approach seems to be to use both item-rest regressions and IRT estimation. The IRT estimation would allow for calibration of the tests on a common scale and provide a scaling factor, whereas the item-rest regression analysis would guide the investigator in deciding on the most appropriate IRT model to be used, in determining a starting value for a *c* parameter (if such a parameter is included), and in evaluating the outcomes of an IRT analysis.

The use of item-rest regressions was illustrated here in a context of vertical equating, in which each examinee responded to all items. In most equating designs, different tests sharing a common subtest as anchor are administered to different groups of examinees. Item-rest regressions can then be used in a similar manner by examining the relationships between total and anchor tests.

The results of the empirical example justify the use of two different tests in the context of the mixed-ability program, and they illustrate how guessing behavior at lower ability levels may jeopardize such designs. However, examinees scoring above 75% correct on the low-level test and above 50% on the high-level test would provide an excellent basis on which to equate the two tests, irrespective of the IRT model used. The finding that the scores of lower-ability students on the high-level tests cannot be recovered correctly from their scores on the low-level test using the one-parameter model demonstrates simply that tests need to be well attuned to the ability level of the target group.

The incorporation of a guessing parameter may yield a more adequate prediction of the average scores of students on one test falling in certain categories on the other, but it cannot predict those scores with acceptable accuracy at the level of the individual student. There would be no point in designing measurement instruments at different ability levels if any test would yield the same precision, irrespective of the ability of the examinees. In vertical equating designs, therefore, response data from badly-matched student-item encounters should be eliminated from the dataset. Students' ability cannot be measured adequately using inappropriate items, and items cannot be calibrated using inappropriate students.

References

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31-44.

Andrich, D. (1989). Statistical reasoning in psychometric models and educational measurement. *Journal of Educational Measurement, 26,* 81-90.

Choppin, B. (1983). *A two-parameter latent trait model* (CSE Report No. 197). Los Angeles CA: University of California, Graduate School of Education, Center for the Study of Evaluation.

de Gruijter, D. N. M. (1990). A note on the bias of UCON item parameter estimation in the Rasch model. *Journal of Educational Measurement, 27,* 285-288.

de Jong, J. H. A. L. (1986a). Achievement tests and national standards. *Studies in Educational Evaluation, 12,* 295-304.

de Jong, J. H. A. L. (1986b). Item selection from pretests in mixed ability groups. In C. W. Stansfield (Ed.), *Technology and language testing.* Washington DC: Teachers of English to Speakers of Other Languages (TESOL).

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. *Applied Psychological Measurement, 5,* 203-208.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23,* 283-298.

Divgi, D. R. (1989). Reply to Andrich and Henning. *Journal of Educational Measurement, 26,* 295-299.

Fairbank, B. A., Jr. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement, 11,* 245-262.

Harris, J. D., & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11,* 151-159.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure.* Technical Report No. 86-69, Research Report No. 86-31. Princeton NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9,* 281-288.

Levine, M. V. (1982). Fundamental measurement of the difficulty of test items. *Journal of Mathematical Psychology, 25,* 243-268.

Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form—a confrontation of Birnbaum's logistic model. *Psychometrika, 35,* 43-50.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Lord, F. M. (1983a). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika, 48,* 425-435.

Lord, F. M. (1983b). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika, 48,* 477-482.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Numerical Algorithms Group, Inc. (1987). *Fortran library manual, Mark 12, Vol. 3.* Downers Grove IL: Author.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38,* 221-233.

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

Urry, V. W. (1976). Ancillary estimators for the item parameters of mental test models. In *Computers and testing: Steps toward the inevitable conquest* (PS-76-1). Washington DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Wright, B. D., & Douglas, G. A. (1977). Conditional vs. unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement, 37,* 573-586.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21,* 93-111.

Author's Address

Send requests for reprints or further information to Dato N. M. de Gruijter, Educational Research Center, University of Leiden, Boerhaavelaan 2, 2334 EN Leiden, The Netherlands.