— ABSTRACTS/ALGORITHMS —

Assessing the fit of dichotomous item response theory models

R. J. DE AYALA

University of Maryland, College Park, Maryland

The advantages of item response theory (IRT) over classical test theory have been well documented. They include, for example, sample-free person and item parameters, as well as construction of tests with known properties (see Hambleton & Swaminathan, 1985; Wright, 1968). These benefits may be realized only when the assumptions of the IRT model used (e.g., unidimensionality, a specific shape of the item response function) are met by the data. The fitting of an IRT model to data requires the verification of the tenability of the model's assumptions. Without this verification, the validity of the model's use is questionable.

Various fit indices have been developed to aid in assessing a model's fit; typically, they are item-oriented (i.e., the model's fit is assessed to each item). Some indices—as, for example, Yen's chi-square (Yen, 1981) or Bock's chi-square (Bock, 1972)—may be applied to any of the commonly used IRT models such as the one-, two-, or three-parameter logistic models (1PL, 2PL, 3PL, respectively); others, such as the Rasch model's item fit statistic (Smith, 1988; Smith & Wright, 1987), are model-specific indices. These fit statistics are or are based on chi-square (the Rasch model's t statistics are transformations of chi-square) and differ from one another with respect to the number intervals and the computation of the predicted score. Because the power of a test is in part influenced by sample size, these indices tend to be very sensitive to lack of fit with large samples. Therefore, there is a tradeoff between the large sample sizes that are required and desirable for stable and accurate parameter estimates and the use of these indices.

In addition, these fit statistics are summary in nature. For instance, a model may fit an item for most of the ability continuum, but it may underestimate the probability of a correct response for low-ability examinees. Conversely, the model may provide poor fit throughout the ability continuum. In both situations, a fit statistic may identify the item misfit yet be unable to identify the cause of the misfit. However, this identification may aid in model selection. In the former case, for example, the use

of a model that models guessing may provide better fit than one that does not.

Graphical methods may aid in the identification of misfit and its cause. Ludlow (1986) presented a series of techniques for the analysis of residuals. In this approach, residuals are displayed by various graphical methods, such as quantile-quantile plots, sum-difference graphs, density-difference plots, box and whiskers graphs, and dot graphs. Using the combination of these techniques, Ludlow was able to identify various sources of measurement error. In contrast, Kingston and Doran (1985) compared the regression of the observed proportion of examinees answering an item correctly on estimated ability ($\hat{\theta}$) with the item response function (the "item characteristic curve," or ICC) based on the estimated item parameters. Kingston and Doran divided the ability continuum into 15 intervals at every 0.4 logit and assigned examinees to one of the intervals on the basis of their θ s. The proportion (P_i) of examinees answering the item correctly (corrected for omits) and the probability of a correct response (p_i) for each interval, given an IRT model, were computed. The observed P_i s as well as the expected p_i s were plotted against θ . The difference between the two logistic curves was an indication of item misfit.

The comparison of the empirical (observed) and predicted regressions (OCC and ICC, respectively) can be very useful with or without the calculation of a misfit statistic. For instance, given that IRT models assume a specific shape of the item response function, the use of OCC/ICC comparisons will identify the shape of the data's item response function and thereby facilitate the selection of the appropriate IRT model. The Kingston and Doran implementation of this approach may be problematic to some researchers. Specifically, the classification of examinees to intervals according to their θ s assumes that $\hat{\theta}$ is an accurate estimate. However, the existence of misfitted items (and potentially the use of an inappropriate model) will affect the estimation of θ and thereby influence the classification of examinees to intervals as well as the calculation of P_i . This problem is shared by estimation programs that provide an option to create the predicted and empirical regression plots (e.g., BILOG by Mislevy & Bock, 1982). One of the most commonly used estimation programs, LOGIST (Wingersky, Barton, & Lord, 1982), does not provide this option for the user.

An additional potential concern for researchers using the Kingston and Doran (1985) approach is that most likely the number of individuals in each interval differs across the OCC. Therefore, the investigator needs not only to compare the two regression curves, but also to weight the information provided by each curve by the

Correspondence may be addressed to R. J. De Ayala, Measurement, Statistics, and Evaluation, Benjamin Building, University of Maryland, College Park, MD 20742.

number of examinees used to calculate P_i . It should be noted that estimation programs that do provide an option to output the predicted and empirical regressions do not provide an indication of the number of examinees used to obtain P_i .

The program. The OCCPGM program creates the OCC from the proportion of examinees in a percentile interval who correctly answer the item and the number of percentile intervals the user desires (range of intervals is 2 ... 21); percentiles are based on number correct score and not θ . For example, if the user wants to create the OCC for an item based on deciles, then the proportion of individuals correctly answering the item is calculated for each decile. By using a percentile metric, the two concerns mentioned above are addressed. An item's ICC is obtained by using the model specified by the user (the 1PL, 2PL, and 3PL models are supported) and the item's parameter estimates to calculate the probability of a correct response conditional on θ .

Computer language. The program is written in Pascal/VS (2.1) for the IBM operating under VM/CMS.

Input. Two input files are required. File 1 is either a dichotomous data set or raw responses with a key of correct answers, whereas the second file contains the estimated item parameters; the estimated item parameters may be obtained from any estimation program. Program parameters allow specification of specific (or all) items whose OCC and ICC are desired, as well as selection of the IRT model (1PL, 2PL, and 3PL) to use for the computation of ICCs, the number of percentile intervals, and whether or not to use the logistic scaling constant.

Output. The output file contains the plots of OCC and ICC for all items for which figures were requested. A second optional file (the user specifies its creation) contains the values plotted in the first file; this file may be created as a tab-delimited file. This second output file allows the user to use this file as input to a more sophisticated plotting or graphing program so that figures may be created on a plotter or laser printer.

Limitations. The maximum number of items is 200, with a maximum of 4,000 examinees; these values should handle most practical testing problems. Due to the dynamic allocation of certain variables that are a function of the number of items, operating system restrictions on

the user's virtual machine (e.g., memory restrictions, etc.) may not allow the user to use the program's maximum number of items and examinees. The number of percentile intervals is constant for all items. Tests must consist of at least 3 items with valid responses from 0 to 6.

Documentation. Instructions for operation of the program (i.e., input program parameters, input and output files, etc.) as well as sample output is provided. A sample data set corresponding to the sample output is included, to allow the user to verify the program's correct operation.

Availability. A copy of the program may be obtained from the author for a \$5 handling fee and a tape. CMS-formatted disk files will be dumped to the user-supplied tape, and the user may access the program with the tape load command.

REFERENCES

- BOCK, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- HAMBLETON, R., & SWAMINATHAN, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- KINGSTON, N. M., & DORAN, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement*, 9, 281-288.
- LUDLOW, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, 10, 217-229.
- MISLEVY, R. J., & BOCK, R. D. (1982). *BILOG, maximum likelihood item analysis and test scoring: Logistic model*. Mooresville, IN: Scientific Software, Inc.
- SMITH, R. M. (1988, April). *A comparison of the power of Rasch total and between item fit statistics to detect measurement disturbances*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- SMITH, R. M., & WRIGHT, B. D. (1987, April). *A new method of testing person fit in the Rasch model*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- WINGERSKY, M. S., BARTON, M. A., & LORD, F. M. (1982). *LOGIST users' guide*. Princeton, NJ: Educational Testing Service.
- WRIGHT, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85-101). Princeton, NJ: Educational Testing Service.
- YEN, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.

(Manuscript received November 20, 1989;
revision accepted for publication January 15, 1990.)