SOFTWARE REVIEW

XCALIBRE™ Marginal Maximum-Likelihood Estimation Program, Windows™ Version 1.10

Mark J. Gierl and Terry Ackerman
University of Illinois at Urbana-Champaign

In this, the first software review for *Applied Psychological Measurement*, XCALIBRE™, a marginal maximum-likelihood estimation (MMLE) program for computing item and trait (θ) parameters for the two-parameter logistic (2PLM) and the three-parameter logistic (3PLM) item response theory (IRT) models, is evaluated. The criteria used to evaluate this program were established, in part, by the Editor for Software Reviews, Richard Luecht, as well as by us. We identified criteria that should be satisfied to produce a computer program that is useful to both researchers and practitioners using IRT. The criteria are discussed throughout this paper and are also presented, more succinctly, in the Appendix. The review is divided into five sections: program description, program interface and performance, output, documentation, and our evaluative conclusion.

Program Description

The use of IRT in research and applied settings requires computer-based procedures for estimating item and θ parameters. XCALIBRE for Windows™, Version 1.10, is a computer program available from Assessment Systems Corporation that estimates item and θ parameters for the 2PLM and 3PLM. It also provides classical test theory item and test analyses. The program can accommodate tests with as many as 750 items per examinee with no practical limit on the total number of examinees (other than disk space).

XCALIBRE estimates the parameters from dichotomously scored test data using the expectation-maximization (EM) algorithm to implement MMLE. The use of priors for the item and θ distributions facilitates the MMLE process. XCALIBRE assumes that the underlying unidimensional trait measured by the test is normally distributed along the θ continuum [i.e., $\theta \sim N(0,1)$]. For items, normal distributions are assumed for the *a*, *b*, and *c* parameters (i.e., the item discrimination, item difficulty, and pseudoguessing parameters, respectively); however, the user can modify the shape of the prior densities by specifying the mean and standard deviation of the prior distributions (default values are also supplied).

XCALIBRE proceeds through the EM cycles until the maximum number of iterations specified by the user has been reached or until the sum of the absolute changes in the item parameters is .05 or less from one cycle to the next. (This default tolerance and the maximum number of iterations can be changed by the user.)

Once the item parameters are estimated, XCALIBRE computes the standardized residuals for each item so that the data-model fit can be evaluated and, if requested by the user, produces IRT θ estimates for each examinee using maximum-likelihood, Bayes modal, or expected a posteriori (EAP) estimation methods. XCALIBRE also has a linking feature that allows the user to equate new item parameters to a pre-existing scale using anchor items.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 3, September 1996, pp. 303–307

0146-6216/96/030303-05$1.50

Program Interface and Performance

XCALIBRE is flexible and easy to use. The program is easy to install in a Windows environment; the graphical user interface is logically formatted; and the work area is clearly labeled and structured. Two folders are available. The first folder, called the *file folder*, controls the input and output files. The user specifies the location of the input data file and the analysis output file. The user has the option of generating an IRT estimated θ score file and an external item statistics file so that the estimated IRT parameters can be easily exported to another program. The user also has the option of importing a file containing item names or identifiers so the output is easier to interpret.

The second folder, called the *options folder*, controls the analysis options. The user can select a 2PLM or 3PLM and the prior distributions for the items. The user can also specify the number of EM cycles and apply the floating priors option (i.e., allow the mean for the item parameter prior distributions to be updated after each EM cycle). Program defaults are provided for each of these alternatives. The options in each folder provide the user with flexibility and control when conducting an IRT analysis.

To evaluate the efficiency of preparing a data file and running the program, three datasets containing student response vectors from a test administered by a large-scale achievement testing program were analyzed using the 3PLM and the XCALIBRE program defaults. (To report on the accuracy of the XCALIBRE parameter estimation procedure would require an empirical study that goes beyond the scope of this largely descriptive review. However, empirical studies comparing XCALIBRE to other IRT estimation programs have been conducted; see, for example, a study by Yoes, 1996.)

We timed each calibration run using a small, medium, and large dataset. All runs were performed with a Pentium™ 66 MHZ computer with 16 megabytes of RAM. The small file contained dichotomously scored data from 500 examinees for a 30-item subtest taken from a larger 60-item test. The medium file contained dichotomously scored data from 1,000 examinees for a 60-item test. The large file contained dichotomously scored data from 2,000 examinees also for a 60-item test.

The data files were easily formatted for use with XCALIBRE. Each file contained four control lines that described different features of the dataset followed by the scored (0,1) item response data. In the first control line, the user must specify the number of items, the code for omitted responses, the code for items not reached, and the number of ID characters. The second and third lines contain the keyed response for each item and the number of alternatives for each of the items, respectively. The final control line allows the user to specify which items are to be included in the analysis.

The program was relatively fast and efficient for all three files. The small file required approximately 1 minute and 19 seconds to estimate item and θ parameters. The medium file required approximately 3 minutes and 45 seconds for the calibration run; the larger file required approximately 7 minutes and 12 seconds. The convergence criterion was reached after four iterations for the small dataset and after three iterations for both the medium and large datasets. During each analysis, XCALIBRE provided a percentage-complete bar to monitor the run time. It also reported the total elapsed time when the run was finally completed. These run-time reports were surprisingly useful because they monitored the analysis and provided feedback about the estimation process. This feature would be invaluable if large datasets were processed or if a slower computer was running the program because the user could accurately anticipate the time needed to execute the program.

Unfortunately, the version of XCALIBRE reviewed by us did not contain on-line help. This important omission forces the user to rely on the XCALIBRE manual when running the program, and detracts from this otherwise flexible and easy-to-use graphical interface.

Output

XCALIBRE output is satisfactory but not encompassing. The program provides the user with a compre-

hensive summary of the program input—the input and output files are reported along with the analysis configuration—as well as item-level and test-level output. The item-level output includes the estimated item parameters for the 2PLM or the 3PLM and associated standard errors; the sample size; the item proportion correct; and two item-test correlations [$r(u_i, X)$, the product-moment correlation between the dichotomous item score, u_i, and the number-correct total score, X, and $r(u_i,\theta)$, the product-moment correlation between u_i and the continuous trait, θ]. A standardized residual fit statistic is also reported for each item. The item information function, an important feature used in test development, is not included as part of the item-level summary. It would be helpful if item information values were provided, even if they were limited to a select number of quadrature points.

XCALIBRE also provides a comprehensive classical item analysis in which the p value (proportion of examinees) and the two item-test correlations—$r(u_i, X)$ and $r(u_i, \theta)$—are reported for each of the multiple-choice alternatives (i.e., the keyed option and the distractors). The program also flags items with out-of-range parameters, items with distractors containing unusually high point-biserial correlations, and items with a standardized residual statistic exceeding a prespecified value that indicates possible item misfit. In short, XCALIBRE provides a comprehensive item analysis containing many informative item-level variables. It also alerts the user to potential problems with the data.

The XCALIBRE output has some noteworthy limitations. First, the graphics are weak. The test response function and the test information functions are the only two graphs provided, and both use asterisks (*) to plot discrete data points (as opposed to creating a smooth line); this approach makes the functions difficult to interpret. Second, and more importantly, item-level graphics are not produced, even as an option. IRT is an item-based theory rather than a test-based theory, and graphics depicting item features such as the item response function, standard error, information function, and data-model fit should be produced for the user. The graphs currently produced by XCALIBRE provide the user with little practical information about these salient item features.

Documentation

The documentation that accompanies XCALIBRE is satisfactory. The manual provides a step-by-step outline for formatting the data, running the program, interpreting the annotated output for one example, and trouble-shooting problems. The XCALIBRE user's manual is concise (a mere 40 pages) but still comprehensive and instructional. The authors provide an excellent overview of the MMLE process. Unfortunately, the overview of IRT trait MMLE/EM estimation is less complete. The estimation algorithm is briefly described and there is little reference to existing literature.

A weakness in the XCALIBRE documentation is an inadequate number of examples. The program diskette includes only one dataset. The user's manual provides a description of sample output generated using the program defaults with a 3PLM for 400 examinees responding to a 30-item multiple-choice test. The example is straightforward and overly simplistic when compared to the more complicated situations that researchers and practitioners frequently encounter. Thus, more examples are needed demonstrating the capabilities and options available in XCALIBRE. The program diskette(s) should include more datasets and the manual should provide more sophisticated analyses—analyses that the program can conduct and that test users can expect in practice. For example, how does the output from a 2PLM and 3PLM compare and contrast? How can datasets be linked using common items? What salient features characterize mixed models (i.e., runs in which parameters for both the 2PLM and 3PLM are estimated) as briefly mentioned in chapter 3 of the XCALIBRE manual? The XCALIBRE user's manual describes the program but fails to demonstrate how the program can be applied in a variety of testing situations. Assessment Systems Corporation also offers technical product support for XCALIBRE by phone, fax, or email, and an Internet Web site.

Evaluative Conclusion

XCALIBRE for Windows is a contender but not a leader in the IRT software market. XCALIBRE satisfies the basic requirements of an IRT program: It estimates item and θ parameters for a test. XCALIBRE has some excellent features, including a program interface that is logically-structured and well-described, which makes the program easy to use; it provides a comprehensive item analysis; and it comes with instructional documentation. However, XCALIBRE also lacks some important features that should be included with any IRT package. For example, it has no on-line help, thereby forcing the user to rely on the program manual; it produces test-level rather than item-level graphics; and it has only one sample run. In short, XCALIBRE serves as a basic package that will probably meet the needs of most researchers and practitioners who use IRT.

Appendix

The evaluation of XCALIBRE for Windows, Version 1.10, was guided by the following criteria. These criteria were developed through our experiences with IRT, and with computer software in both research and applied settings. Descriptions of the evaluative criteria follow, including example questions that helped guide our thinking.

1. Item and Trait Level Parameter Estimation

The use of IRT requires procedures for estimating the parameters of the items and of the examinees' trait levels on a test.

- Does the program provide item parameter estimates?
- Does the program provide trait level parameter estimates?
- Can different methods be used to estimate the item and trait parameters?

2. Flexibility and Ease of Use

The computer program must be flexible, providing users with options for different testing situations, yet easy to operate, given that most users will have limited knowledge about IRT.

- Are the data easily formatted and prepared for the analysis?
- Is the graphical user interface well structured (e.g., are the icons clearly marked and are the function buttons carefully described)?
- Does the program contain an editor for the data, input, and output files?
- Does the program contain on-line help?
- Is the program fast and quickly executed?
- Can the program be used with large samples and with large tests containing many items?

3. Comprehensive Output

The output must be comprehensive—providing the user with information at both the item and test level.

- Does the output include a classical item analysis?
- Does the output include an IRT analysis in which summaries of the model fit, the standard errors, and the information functions are available at the item and test level?
- Does the program graph the item response functions and the test response function?
- Does the program graph the item information function and the test information function?
- Is the output well-formatted and easy to read?
- Can the input and output files be easily exported to other applications?

4. Manual and Technical Support

The operation and execution of the program must be clearly presented in the manual and the user should receive free technical product support.

- Does the program come with a comprehensive manual in which data formatting, program formatting,

user options, estimation procedures, and error codes are carefully described?

- Does the program come with datasets and does the manual contain detailed sample runs in which typical examples encountered in test development are examined and explained?
- Are program features documented and explained in the manual (e.g., linking functions, scaling options, comparisons across parameter estimation options)?

References

Yoes, M. E. (1996). *An updated comparison of microcomputer-based item parameter estimation procedures used with the 3-parameter IRT model* (ASC Technical Rep. 95-1R). St. Paul MN: Assessment Systems Corporation.

Editor's Note

At the time that this review was conducted, only a 16-bit version of XCALIBRE was available. A 32-bit version is now available that will significantly improve the speed of the calibrations. The price of the 32-bit version is the same as the 16-bit version. In addition, since this review was written an updated version with on-line help has been released.

Summary

Program Name: XCALIBRE™: Marginal Maximum-Likelihood Estimation Program, Windows Version 1.10

Vendor: Assessment Systems Corporation
2233 University Avenue, Suite 200
St. Paul MN 55114, U.S.A.

Voice: (612) 647-9220
Fax: (612) 647-0412
Email: sales@assess.com (Sales)
info@assess.com (Product Information)
mailbox@assess.com (General)
Worldwide Web: http://www.assess.com (Free Product Demos)

Price: $399.00, plus shipping.

System Requirements: 80386 or higher computer running Microsoft Windows 3.1, Windows NT, or Windows 95.
2 Mb free RAM and 1 Mb free disk space.

Capabilities: Item response theory two- and three-parameter logistic model marginal maximum-likelihood estimation of dichotomous data, using the EM algorithm.
Classical item analysis.
Up to 750 items per examinee.
Support for omits and not reached items.
Sample size limited only by disk space.