Open-source IRT:
A comparison of BILOG-MG and ICL features and item parameter recovery

Alan D. Mead
Scott B. Morris
David L. Blitz
Illinois Institute of Technology

Author Note

Correspondence concerning this article should be addressed to Alan D. Mead, Institute of Psychology, Illinois Institute of Technology, 3101 South Dearborn/2nd floor, Chicago IL 60616. E-mail: mead@iit.edu

Abstract

BILOG is the *defacto* standard for dichotomous IRT model estimation. However, BILOG is a commercial product and limited in other ways. Hanson provides an open-source alternative, ICL, and this paper compares ICL to BILOG in terms of features and obtained item parameter estimates. In general, BILOG has more features, especially with respect to assessing model-data fit. One notable feature of ICL is built-in support for bootstrap estimates. ICL and BILOG produced very, very similar estimates of item parameters.

Open-source IRT:
A comparison of BILOG-MG and ICL features and item parameter recovery

In their seminal article introducing BILOG and comparing it to LOGIST, Mislevy and Stocking (1989) noted the central role of estimation software for researchers interested in using item response theory (IRT). In the intervening 18 years, BILOG-MG has become the defacto standard for estimating the parameters of dichotomous IRT models (Rupp, 2003).

This paper compares BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 2003; hereafter designated BILOG) and ICL (Hanson, 2002a; 2002b), a relatively new program for estimating the parameters of IRT models.

The proposed talk has three goals: (a) to introduce Hanson's (2002a) ICL software; (b) to describe common and unique features of ICL and BILOG so that users may choose the best program for a given application; and (c) to demonstrate that the two programs are equivalent in accomplishing their primary task of estimating item parameters.

Introducing ICL. Recently, Hanson (2002a) released stand-alone software for estimation of IRT model parameters, called IRT Command Language (ICL). ICL is free , open-source software (e.g., Raymond, 1999) similar to the popular *R* statistical software (de Leeuw & Mair, 2007) and is licensed in a way that allows it to be modified and extended. In fact, ICL is actually IRT estimation functions (ETIRM; Hanson, 2000) embedded into a fully-featured programming language called Tcl ("tickle"; Welch, Jones & Hobbs, 2003) and thus allowing relatively complex operations. ICL is available for Windows, Macintish and Linux.

Comparing ICL and BILOG-MG

One reason for BILOG's universal acceptance was its introduction of marginal maximum likelihood (MML) estimation in a Bayesian framework (Bock & Aitkin, 1981; Mislevy, 1986). The MML estimation algorithms introduced in BILOG were a significant statistical and practical advance over the programs available then, especially the popular LOGIST program (Mislevy & Stocking, 1986; Lord, 1980).

Other reasons for BILOG's wide usage are very practical. BILOG has many features that are designed for applied work. For example, BILOG will read and score raw data in many different formats and BILOG allows flexibility regarding the estimation from data of complex sampling schemes. BILOG has also enjoyed professional support, including a well-written manual and technical support from the publisher. The program has been maintained and its features have expanded. Today, BILOG comes with a Windows-based "shell" program that allows users to build command syntax from menus and pull-down lists. BILOG's '-MG' suffix indicates the version which handles estimation in a multigroup situation (Bock & Zimowski, 1996) and is now the only generally available version.

A natural question is the degree to which ICL is similar to BILOG. This "apples and oranges" comparison defies simple lists of similar and separate features. Many of the features directly available in BILOG are available indirectly in ICL (but must be programmed). The ICL manual (Hanson, 2002b) is particularly helpful, providing a series of example ICL command files which accomplish various tasks. We will confine our comparison of BILOG-MG and ICL to the features documented in their respective manuals (including examples) rather than assuming any special Tcl knowledge on the part of the ICL user.

Estimation. Both BILOG and ICL provide maximum marginal likelihood (MML; Bock & Aitkin, 1981) estimation via the EM algorithm (McLachlan & Krishnan, 1997; Dempster, Laird, & Rubin, 1977) and both implement a Bayesian framework (Mislevy, 1986). Many of the details of the estimation in ICL are provided in Woodruff and Hanson (1997) and Hanson (1998). One significant difference is that ICL does not implement Fischer scoring in the same way as BILOG; BILOG users will notice this in two ways. First, there are no "Newton cycles" following the "EM cycles" during estimation. And second, ICL does not compute the item variance-covariance matrix. As an alternative for users requiring the item variances or covariances, ICL implements a bootstrapping feature which can be used to generate data that can be analyzed with SAS or SPSS to estimate these values.

Estimation options. Both programs allow the user to influence parameter estimation, such as setting priors on structural parameters and specifying the maximum numbers of estimation cycles. ICL provides a larger selection of options for expert users, although many of these options may not be useful for the average user. ICL also provides more flexibility regarding the prior distributions for item parameters (beta, normal, and log-normal are available).

Model fit information. Both programs provide information about convergence but otherwise BILOG provides considerably more information about model fit. BILOG includes provides chi-square indices of the fit of individual items (or residual information for short tests), summaries of the estimates, and "fit plots" which overlay the empirical proportion correct for various "bins" of theta-hat values upon the ICC. This is a significant practical advantage for BILOG.

Documentation. Both BILOG and ICL have well-written manuals. The manual describing BILOG (du Toit, 2003) also chapters providing an overview of the estimation procedures, other IRT programs from SSI, and a chapter of historical material. The chapter on BILOG contains some introductory information, documentation of the commands and options, examples, and file formats. The ICL manual (Hanson, 2003b) is similar to the BILOG chapter of the SSI manual, providing introductory information, documentation of the commands, and examples. The documentation is divided between the basic commands required for a default single-group dichotomous model estimation and more advanced commands. Both programs come with PDF copies of the documentation, allowing easy searching for information. The ICL manual would profit from the addition of an index.

Data processing. Both programs provide data processing options. For example, BILOG will score raw multiple-choice options. ICL does not provide this capability directly. Also, missing

responses are always ignored by ICL, while BILOG provides three models for missing data: wrong, partially right, and ignored.

Item parameter recovery study

Although one previous study of on-line calibration (Ban, Hanson, Wang, & Harris, 2001) found similar results using ICL and BILOG, no previous research has been reported which directly compares the abilities of the two programs to recover item parameters. To directly address this issue of the comparability and accuracy of the item parameter estimates produced by the two programs, we conducted an item-parameter recovery study using simulated data.

We hypothesized that BILOG might have implementation decisions (either statistical or logical) that resulted in better estimation for very small or large samples. Therefore we generated samples of three different sizes: N=100; N=1,000, and N=50,000. For each condition, items responses were generated for a 50 item test based on operational item parameters from a high-stakes certification exam. The true, generating parameters of the IRT model for each item are given in Table 1. The procedure was replicated five times. [We considered a larger number of replications; however, the results were remarkably stable.]

Method

Item responses were generated using a Fortran 90 program. The true ability of the examinees was generated from a standard normal distribution with M = 0 and SD = 1 using the IMSL (1984) pseudo-random number generator DRNNOR. Next, the probability of a correct response to each item was computed for each examinee as a function of the individual's ability according to a 3PL IRT model. Then, the individual's response to each item was computed by generating a uniform random number between 0 and 1 using the IMSL DRNUN routine, and assigning a score of 1 if this number was less than the individual's probability for that item, and a score of 0 otherwise.

Results

The main outcome variable was the root mean square error (RMSE) between ICCs. We chose to compare differences in ICCs because for some items, different values of item parameters can yield very similar ICCs. We computed the RMSE using 41 points from -3.0 to 3.0.

Three comparisons were made:

- ICC's computed from ICL estimates compared to ICC's computed from the true, generating parameters
- ICC's computed from BILOG estimates compared to ICC's computed from the true, generating parameters
- ICC's computed from ICL estimates compared to ICC's computed from BILOG estimates

This resulted in hundreds of comparisons. We summarized by computing the mean and standard deviation of the RMSE statistic across the 50 items. Presented in Table 2, the comparisons of ICL estimates to truth (IT) and of BILOG estimates to truth (BT) are very similar in size and larger than the comparison of ICL estimates to BILOG estimates (IB). This seems to indicate that both programs are equally good at recovering item parameters and that they actually achieve comparable results (as opposed to disparate results that have a comparable accuracy). This is hardly surprising given the similarity of the core estimation algorithms. Table 2 does not indicate any interaction between sample size and recovery accuracy.

Discussion

At their heart, ICL and BILOG both estimate the parameters of dichotomous IRT models and they perform this function using similar algorithms and in a uniformly accurate manner. Slight differences in the second or third decimal of the RMSE statistic are unlikely to make any practical difference and can probably be ignored.

While both programs offer many similar features, BILOG is clearly the more mature product with a number of practical advantages such as: additional features, greater ease-of-use, and professional support. In any serious analysis of real data, BILOG has a decided advantage in terms of the rich information about model fit provided in the Phase 2 output. Indeed, most examination programs that rely upon BILOG today would find ICL's default output to be inconveniently sparse.

ICL offers some users greater power. For example, ICL can be used to generate simulated item responses for research. And ICL offers built-in boot-strapping functionality for empirically estimating the sampling distribution of item parameter estimates. In some research using simulated data, model fit may not be an issue and ICL may be more convenient than BILOG. For those who grasp Tcl, ICL offers a unique opportunities to extend ICL to include additional functionality. In the form of ETIRM, ICL offers a unique opportunity for programmers to build professional-grade IRT parameter estimation into item banking software, simulation studies, and other applications. Finally, because ICL incorporates both dichotomous and polytomous models, ICL may be a simpler solution to the estimation of mixed-format examinations.

Limitations. We are currently evaluating the theta-hat estimates produced by the two programs; our talk will report those results. We also plan comparisons for multigroup situations.

In addition, this study compared data generated from the 3PL with a normally-distributed theta density—that is, perfectly coincident with the assumptions of the IRT model and the estimation software. We are currently exploring the effect, if any, of using non-normal theta densities.

References

Ban, J. Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of online pretest item–calibration/scaling methods in CAT. *Journal of Educational Measurement, 38*, 11–212.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., & Zimowski, M. F. (1996). Multiple group IRT. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory*, pp. 433-448. New York: Springer-Verlag.

de Leeuw, J. & Mair, P. (2007). An introduction to the special volume on ``Psychometrics in R". *Journal of Statistical Software*, 20(1), 1-5.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1–38

du Toit, Mathilda (Ed.). (2003). *IRT from SSI*. Chicago: SSI Scientific Software International.

Hanson, B. A. (2002a). *IRT Command Language (ICL)*. Computer software. [Available at http://www.b-a-h.com/software/irt/icl/index.html]

Hanson, B. A. (2002b). *IRT Command Language*. Computer software manual. [Available at http://www.b-a-h.com/software/irt/icl/icl_manual.pdf]

Hanson, B. A. (2000). *Estimation Toolkit for Item Response Models (ETIRM)*. Computer software. [Available at http://www.b-a-h.com/software/cpp/etirm.html]

Hanson, B. A. (1998). *IRT Parameter Estimation using the EM Algorithm*. [Available at http://www.b-a-h.com/papers/note9801.html]

International Mathematical and Statistical Library (1984). *User's Manual: IMSL Library, Problem-solving software system for mathematical and statistical FORTRAN programming* (vol. 3, ed. 9.2) Houston, TX: Author.

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

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: John Wiley & Sons.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*(1), 57-75.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 117-195.

Raymond, E. S. (1999). *The Cathedral & the Bazaar*. Sebastopol, CA: O'Reilly [Available at http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/]

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3*(4), 365-384.

Welch, B. B., Jones, K., & Hobbs, J. (2003). *Practical programming in Tcl and Tk* (4th Edition). Upper Saddle River, NJ: Prentice Hall. [Available at http://beedub.com/book/]

Woodruff, D. J., & Hanson, B. A. (1997). Estimation of item response models using the EM algorithm for finite mixtures. Paper presented at the Annual Meeting of the Psychometric Society (Gatlinburg, Tennessee, June). [Available at http://www.b-a-h.com/papers/paper9701.html]

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3: Item analysis and test scoring with binary logistic models. Chicago, IL: Scientific Software. [Computer software.]

Table 1. True item parameters.

Item	a	b	c	Item	a	b	c
1	0.40237	-1.64510	0.11956	26	0.49745	-0.77670	0.04026
2	1.50722	1.10230	0.18490	27	0.58603	-0.77812	0.22593
3	0.80505	0.17244	0.26850	28	0.69610	1.24704	0.50000
4	0.15805	-2.05920	0.26024	29	0.24525	-0.66669	0.29986
5	0.45750	1.52293	0.26201	30	0.87542	0.35110	0.03471
6	0.66742	1.49058	0.38434	31	0.73514	-0.56731	0.16151
7	0.34609	-0.03989	0.26903	32	0.43242	-1.36105	0.22960
8	0.20172	-1.29105	0.27716	33	0.68566	1.22426	0.36349
9	0.52335	-1.76923	0.27812	34	0.57749	-2.19104	0.15466
10	0.37082	-1.36278	0.21690	35	0.52161	-1.97909	0.11212
11	0.50803	0.66152	0.22796	36	0.70127	0.64502	0.11812
12	0.58292	0.82928	0.14032	37	0.57184	-1.29226	0.20895
13	1.17285	1.08667	0.12216	38	0.35913	-1.18656	0.28738
14	0.62464	0.15067	0.31497	39	0.22536	0.66184	0.17440
15	0.48563	1.06429	0.14727	40	0.50863	-1.31509	0.20115
16	1.03591	0.35649	0.28852	41	1.14187	0.82448	0.14375
17	0.41226	-0.50816	0.26086	42	0.53855	0.74164	0.23957
18	0.75952	-0.33662	0.47837	43	0.58855	0.12126	0.15586
19	0.98741	0.12531	0.32846	44	0.75182	0.43824	0.15527
20	0.61734	-0.40191	0.23452	45	1.70122	1.30561	0.33408
21	0.38579	-1.52071	0.26428	46	0.71508	1.09795	0.23546
22	0.40970	-0.34555	0.14183	47	0.41586	0.76006	0.09524
23	0.46755	0.70829	0.08476	48	0.48227	0.22206	0.08606
24	0.61385	0.52302	0.24969	49	0.91812	1.24238	0.19714
25	0.14550	0.60789	0.21611	50	0.87570	1.47565	0.12994

Table 2. Means and standard deviations of the RMSE between ICCs across 50 items (five replications shown).

	M_{IT}	S_{IT}	M_{BT}	S_{BT}	M_{IB}	S_{IB}
N=100	0.037	0.034	0.036	0.037	0.021	0.036
	0.031	0.035	0.030	0.036	0.018	0.028
	0.038	0.043	0.038	0.040	0.018	0.043
	0.040	0.041	0.034	0.038	0.024	0.034
	0.041	0.050	0.042	0.047	0.018	0.028
N=1,000	0.014	0.016	0.016	0.019	0.008	0.013
	0.018	0.016	0.022	0.023	0.009	0.021
	0.019	0.022	0.020	0.028	0.010	0.019
	0.019	0.023	0.017	0.022	0.009	0.013
	0.020	0.022	0.019	0.020	0.010	0.017
N=50,000	0.003	0.003	0.003	0.003	0.001	0.001
	0.003	0.003	0.003	0.003	0.001	0.001
	0.003	0.003	0.003	0.003	0.001	0.001
	0.003	0.005	0.003	0.004	0.001	0.001
	0.003	0.003	0.003	0.002	0.001	0.001

Note: M=Mean; S=Standard deviation; IT=ICL estimates compared to true, generating parameters; BT=BILOG compared to true, generating parameters; IB=ICL estimates compared to BILOG estimates