
**Bradley Hanson Award
Information on Past Recipients**

Past Recipients

- 2010 Joseph Patrick Meyer, III; University of Virginia
2009 Ying Cheng; University of Notre Dame
2008 *No recipient.*
2007 Jianbin Fu; Educational Testing Service
2006 Won-Chan Lee; University of Iowa / CASMA
2005 Seonghoon Kim; Keimyung University & Gary Skaggs; Virginia Tech University
-

Project Descriptions

2009 Joseph Patrick Meyer, III; University of Virginia

J-Metrik – A Comprehensive and User-friendly Psychometric Software

Psychometric software applications use obsolete interfaces and complicated syntax files for operation. Consequently, the risk of error is high in operational test analysis and the quality of measurement instruction is limited by computer science topics and computing issues. *JMetrik* addresses these problems and advances the state of psychometric computing. It is a user-friendly and comprehensive psychometric software application that is accessible to K-12 teachers, graduate students, and psychometricians alike. In the spirit of Brad Hanson's work, *JMetrik* is free and open source software. It is available for download at www.ItemAnalysis.com.

JMetrik is a pure Java application that involves a data management system, point-and-click operation, and a consistent user interface for Windows, Linux, and Mac operating systems. It is efficient and user friendly. The data management system is driven by an integrated database capable of storing millions of records per table. Users may define a workspace that contains many projects and each project may contain many data tables. To facilitate organization, projects and data tables are displayed in a tree like structure in the interface. A user simply clicks a table name to analyze it - one data set, many methods of analysis. *JMetrik* currently offers a variety of statistical methods that may be executed by a simple point-and-click interface. The program may also be run by syntax. Indeed, the menu options and dialog boxes simply generate the syntax files that are processed by the program.

Regardless of how they are executed, a variety of methods are currently available including item analysis, differential item functioning analysis, and nonparametric item response theory as well as common descriptive statistics and graphics. The item analysis includes multiple methods of estimating reliability: Cronbach's alpha, Guttman's lambda-2, Feldt-Brennan, and Feldt-Gilmer (see Feldt & Brennan, 1989). Huynh's raw agreement and decision consistency indices are also included (Huynh, 1976). Each reliability estimate is accompanied by the appropriate confidence interval (Fan & Thompson, 2001). Differential item functioning methods include the Cochran-Mantel-Haenszel procedure (Holland & Thayer, 1988; Zwick, Thayer, & Mazzeo, 1997), the

standardized mean difference (Dorans, Schmitt, & Bleistein, 1992), ETS DIF classifications (Zwick & Ercikan, 1989), and nonparametric item characteristic curves (Bolt & Gierl, 2006). Several methods will be added over the next two years according to the software development plans.

Item response theory (IRT) methods have been developed and are currently being tested. Rasch, partial credit, and rating scale models are available (Wright & Masters, 1982). Mean/mean, mean/sigma, Haebara, and Stocking-Lord equating methods have also been developed (see Meyer, in press) and will soon be integrated into *JMetrik*. The IRT and equating methods will be seamlessly integrated into *JMetrik*. Parameter estimates will be saved in new database tables and these tables will serve as input for the equating procedures. The integration of IRT and equating methods into *JMetrik* and the additional methods detailed below are part of the work proposed for this award.

The following methods and features will be integrated into *JMetrik* by the end of December 2011. Specific dates are listed at the following website: http://www.itemanalysis.com/jmetrik_development_plans.php.

Psychometric features:

- o Item response theory estimation and simulation: Rasch, partial credit, and rating scale models (feature already developed and in testing)
- o Item response theory equating from the eqboot library (Meyer, in press): Mean/mean, mean/sigma, Haebara, Stocking-Lord (feature already developed and being integrated)
- o Traditional equating: Mean, linear, equipercentile
- o Generalizability theory: Generalizability study and decision study analysis for basic designs
- o Cognitive diagnosis models: DINA model by maximum likelihood estimation (see de la Torre, 2009)
- o Reliability: stratified alpha estimation

Software features:

- o Data sub setting
- o Save analysis result to database table
- o Capability of calling *JMetrik* from external programs for conducting simulations (batch processing feature already developed)

2009 Ying Cheng; University of Notre Dame

Variable-length Computerized Adaptive Testing System for Cognitive Diagnosis

Cognitive diagnosis has received a great deal of attention recently, especially since the No Child Left Behind Act (2001) mandated that diagnostic feedback should be provided to teachers, students and parents. Traditionally, assessments only provide a summative score which represents the sum of individual scores from all the items on the test. This single, summative score, however, may fail to capture which specific skills a student has mastered, and which ones he or she has not. Cognitive diagnosis, on the other hand, aims to provide detailed diagnostic feedback to test users, and consequently help link educational assessment with instructional intervention (Hartz, 2002; McGlohen, 2004; Hartz, Roussos, & Stout, 2002).

One challenge that cognitive diagnostic assessment faces is efficiency. Therefore, we would like to propose implementing cognitive diagnosis in computerized adaptive testing (CAT), a testing mode that has been known to be more efficient than traditional paper-pencil testing. It tailors the test to each individual's proficiency level, so that test takers do not have to answer items that are too easy or too difficult for them. Research has shown that a CAT that is half as long as a conventional paper-pencil test can maintain the same level of measurement precision (Wainer et al., 2000). Tatsuoka (2002), Tatsuoka and Ferguson (2003) and Xu, Chang, & Douglas (2003) are among the first that examined the problem of combining computerized adaptive testing with cognitive diagnostic models (CD-CAT). They proposed two item selection heuristics, one based on Kullback-Leibler information (i.e., the KL algorithm), and the other based on Shannon entropy (i.e., the SHE algorithm). On the basis of the KL algorithm, Cheng (2008) developed the posterior-weighted KL information or PWKL algorithm, and the hybrid KL algorithm (HKL) which considers not only the posterior but also the distance between latent classes. The simulation study in Cheng (2008) showed that the PWKL and HKL algorithms outperform the KL and SHE algorithms uniformly. Cheng (2008) also proved that the SHE algorithm can be considered a special case of the KL algorithm.

The aforementioned studies share one common feature, namely that they all investigated cognitive diagnostic computerized adaptive testing (CD-CAT) with fixed test length, meaning that the test terminates when a fixed number of items are received by each examinee. We believe that a variable-length CD-CAT will be more flexible and more helpful if it can be of variable-length, meaning that the test length can be tailored more tightly to each test taker's performance. The ultimate goal of the research proposed here is to develop a versatile platform to facilitate research on CD-CAT, which allows both fixed-length and variable-length tests to be administered. To achieve that goal, I plan to conduct a simulation study which complements the discussion of fixed-length CD-CAT in the current literature. The proposed study involves studying (a) competing cognitive diagnosis models, (b) different test lengths, (c) different convergence criteria, and (d) different item selection algorithms.

2007 Jianbin Fu; Educational Testing Service

The Development of a General Purpose Computer Program for the Estimation and Equating of Item Response Theory Models

The project involves the development of a general purpose collection of computer programs for the estimation and equating of item response theory (IRT) models using the freely available statistical software package R. By incorporating several different IRT models and equating procedures within the same software package, it is a very flexible product that can be easily adapted to accommodate the variety of different models currently used in IRT, as well as the various types of equating designs that lend themselves toward different linking procedures. The specific purposes of the project are threefold: (a) to provide a unified implementation in R of a class of techniques to carry out the estimation and equating of a wide range of popular IRT models, such that a user with minimum knowledge of R language is able to use it; (b) to make open source code available so that users can reuse the functions and easily develop variations and extensions to meet specific needs, and (c) to promote R as a valuable computing tool in the educational measurement research community. The ultimate goal is that all significant statistical techniques in education measurement have their implementations in R so that they become easy to find and use within a single common environment. Dr. Fu has written various R programs to carry out statistical computations for research projects on educational measurement that have involved IRT (see, e.g., Li, Bolt, & Fu, 2005a, 2005b; Li & Cohen, 2003). These programs have been used for the following purposes: (a) test characteristic curve linking, followed by observed score equating and true score equating under the graded response model, the partial credit model, the generalized partial credit model, the nominal response model and the two-parameter normal ogive (2PNO) testlet model (Bradlow, Wainer, & Wang, 1999), and (b) the maximum marginal likelihood estimation using an EM algorithm (MML-EM) for the multiple-group 2PNO testlet model with parametric priors (Li, Bolt, & Fu, 2005a).

Importantly, these applications have involved IRT models of considerable complexity. Indeed, many other IRT models and their associated linking/equating procedures can be formulated as special cases or easily extended from the models/procedures that have already programmed. Consequently, Dr. Fu plans to extend this previous work in several directions by (a) expanding the program to handle other linking methods such as mean-mean and mean-sigma, and other IRT models such as the 1-3PL(PNO) models, the three-parameter testlet model (Wainer, Bradlow, & Du, 2001) and the polytomous testlet model (Wang, Bradlow, & Wainer, 2002); (b) expanding the program to accommodate the MML-EM estimation of the multiple-group versions of IRT models, including the two-parameter polytomous multidimensional IRT models (Moustaki, 2000; Muraki & Carlson, 1995) and the three-parameter dichotomous multidimensional IRT model (Beguin & Glas, 2001; Reckase, 1997) as well as sub-models, which compose the majority of popular IRT models; (c) refining and optimizing the codes to be object-oriented for easy maintenance and other extensions, and (d) creating an R package to hold all functions including the writing of help files, and uploading the package to the Comprehensive R Archive Network (CRAN, <http://www.r-project.org/>) as a contributed add-on package to the freeware R for users to download, as well as disseminating the program by other means such as publishing a Computer Software Exchange note on *Applied Psychological Measurement*, and/or uploading the program to a psychometric software exchange website.

Integrating Procedures and Software for Estimating Classification Consistency

Research on classification consistency has been on-going for about 30 years. No current document exists that systematically discusses the many approaches that have been developed. Furthermore, publicly available computer programs do not exist for many of the procedures, and there are no integrated and well-documented computer program packages that, in total, permit users to apply all of the available procedures. Hopefully, this project would remedy these problems.

Specifically, the purpose of my project is:

- (a) to summarize procedures developed so far for estimating classification consistency;
- (b) to create integrated computer software for implementing those procedures; and
- (c) to communicate these issues to various measurement communities.

As an outcome of this project, a monograph will be written, which summarizes various IRT and non-IRT procedures for estimating classification consistency, and provides practical guidelines for users. Also, several computer programs will be created with written manuals. A brief summary of the estimation procedures and computer software that have been developed is provided below. For the sake of convenience, the various estimation procedures are categorized here into two broad categories, IRT and non-IRT approaches. Non-IRT procedures for tests consisting of dichotomously-score items include Huynh (1976), Subkoviak (1976), and Hanson and Brennan (1990). When the test is composed of items more complex than dichotomous ones, the following non-IRT procedures can be used: Livingston and Lewis (1995); Breyer and Lewis (1994); Woodruff and Sawyer (1989); Brennan and Wan (2004); and Lee (2005a). IRT procedures include Huynh (1990); Schulz, Kolen, and Nicewander (1999); and Lee, Hanson, and Brennan (2002) for dichotomously-scored items, and Wang, Kolen, and Harris (2000) for polytomously-scored items. For tests that consist of mixtures of dichotomous and polytomous items, the procedure discussed by Wang et al. (2000) can be extended directly using mixed IRT models.

Some of the procedures mentioned above have been implemented in the following computer programs. Class Consistency (Hanson, 1995) computes classification consistency as discussed in Hanson and Brennan (1990). BB-CLASS (Brennan, 2004) is more general than Class Consistency and can also compute results for the Livingston and Lewis (1995) procedure. MULT-CLASS (Lee, 2005b) computes results based on the multinomial and compound multinomial models as discussed in Lee (2005a). Finally, IRT-CLASS (Lee & Kolen, 2006) is intended to be used for tests that are scaled using dichotomous, polytomous, or mixtures of different IRT models. These computer programs are available from CASMA website: www.education.uiowa.edu/casma.

Manuscripts

- Lee, W., Brennan, R. L. & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement*, 33, 374-390.
- Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory* (CASMA Research Report No. 27). Iowa City, IA: University of Iowa.
- Lee, W. (2005). *Classification consistency and accuracy under the compound multinomial model* (CASMA Research Report No. 13). Iowa City, IA: University of Iowa.

Computer Software

- Lee, W. (2008). *MULT-CLASS: A computer program for multinomial and compound-multinomial classification consistency and accuracy (Version 3.0)*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. (Available from www.education.uiowa.edu/casma).
- Lee, W., & Kolen, M. J. (2008). *IRT-CLASS: A computer program for item response theory classification consistency and accuracy (Version 2.0)*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. (Available from www.education.uiowa.edu/casma).

Revitalizing and Investigating the IRT Command Language Program

The goal of measurement theory is concerned with the justification of various measurement procedures and with the meaningfulness of their results. In item response theory (IRT), the goal is closely related to the statistical techniques for estimating item and proficiency parameters (i.e., calibrating) and developing calibration software to implement the techniques. The techniques are usually quite complicated and hardly applicable to practical problems without the help of calibration software. It has been a matter of regret in the field of educational measurement that technical support and improvement of the open-source computer program *IRT Command Language* (ICL) stopped because of the sudden death of Dr. Bradley A. Hanson, the developer of the program.

I believe that ICL is one of the best IRT calibration programs based on reliable frameworks in terms of psychometric, statistical and computing techniques. ICL is the only widely available IRT calibration computer program that is open source, meaning that all of the source code is available to the public. However, the sudden loss of the developer left it to the educational measurement community to make it more user-friendly and efficient, let alone to maintain it. Regrettably, practical applications of ICL are rarely found in the literature, and there are gaps in the documentation of the theoretical details for ICL. This situation motivated the present project.

The main purpose of the project is to revive the usage of the computer program ICL by comparing its performance to the performance of other commercial IRT calibration software and revealing and documenting the rationale for algorithms used in ICL. To the best of my knowledge, areas that need detailed description of the algorithms include multiple-group estimation, bootstrapping item parameter estimates, and pretest item calibration. Of course, functions of ICL for implementing these areas need to be fully tested with a variety of real and simulated data. The following briefly summarizes the objectives of the project:

1. Dissect, screen, and arrange source code of ICL, so that users can easily use the code for their own purposes.
2. Conduct a series of simulation studies to test the performance and behavior of ICL in comparison to other commercial calibration computer programs. Situations for simulation include those that need separate calibration, concurrent calibration, fixed-parameter calibration, and scale transformation for linking.
3. Document some algorithms built into ICL in technical papers that describe them in detail.
4. Based on the first three points above, make suggestions for further improvement of the performance of ICL.
5. Make suggestions about how to create a graphical user interface (GUI) version of ICL, so that users can use it more easily and more effectively.

Manuscripts

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43, 355-381.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, 19, 357-381.

2005 Gary Skaggs, Virginia Polytechnic Institute and State University

A Psychometric Software Exchange Site

This project involves developing a website for exchanging psychometric software. Brad felt strongly about both having a software exchange website and making software available and free, and I am happy to have the opportunity to continue that vision. My own interest in software issues came out of research I have been working on recently as I noticed how dependent our field is on the availability of software or on the programming skill of the researcher in creating software. This project is an outgrowth of my working with Brad on psychometric software issues. At the 2001 Annual Meeting, Brad presented the report from an NCME ad hoc committee he chaired on issues related to psychometric software. He asked me to be a discussant at that session. After that session, I surveyed several measurement journals on software use and published the findings in an article in *Educational Measurement: Issues and Practice* (Spring, 2004). One of the recommendations in that article was the creation of a psychometric software exchange website. This had also been recommended in the report of Brad's committee, and Brad had developed a prototype exchange on his website. Unfortunately, since his untimely death, this exchange is no longer available. I still firmly believe that such an exchange would be valuable to the measurement community and facilitate our research.

Specifically, I plan to use the funds from the award to: (1) obtain input from NCME members on the structure and function of the exchange; (2) meet with an NCME representative group (e.g. Board of Directors or appropriate committee) to finalize the design of the exchange, including submission procedures and criteria; and (3) begin to implement it. To obtain input from NCME members, I propose to create a brief survey and administer it via the NCME listserv. Based on the results of the survey, I will propose an initial design for the exchange and present it to the appropriate NCME group, probably at the 2006 Annual Meeting. I have some initial thoughts on the website. First, although the site would provide links to commercially available software, a major focus of the site would be to house freeware, open-source applications, and subroutines and functions written by individual researchers. Second, software can be accessed and then downloaded through keywords (e.g. simulation, equating, etc.). Third, there should be some mechanism for software users to provide comments or recommendations to the research community. My hope is that this website will make our research efforts more efficient and productive.

Manuscripts

Skaggs, G. (2004). Software use in psychometric research. *Educational Measurement: Issues and Practice*, 23(1), 28-33.