DOCUMENT RESUME

ED 408 330 TM 026 583

AUTHOR Wise, Steven L.
TITLE Overview of Practical Issues in a CAT Program.
PUB DATE Mar 97
NOTE 8p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Computer Attitudes; Equal Education; Equated Scores; *Item Banks; Item Response Theory; Review (Reexamination); Scoring; Student Attitudes; Student Motivation; Test Anxiety; *Test Construction; Test Items; *Testing Problems; Timed Tests
IDENTIFIERS Large Scale Assessment

ABSTRACT

Computerized adaptive testing (CAT) has become increasingly common in large-scale testing programs. This paper considers relevant practical issues that are likely to be faced by the developers and managers of a CAT program. The first cluster of issues is that of item pool development and maintenance. It includes such considerations as item pool specifications, the choice of item response theory model, and other concerns in constructing and choosing test items. The second cluster of items involves administering and scoring the CAT. Proficiency estimation method, test items, item review, and equating CAT scores to paper-and-pencil tests are areas that must be considered. The third cluster involves protecting the integrity of the CAT item pool, considering security and coaching concerns. A fourth cluster includes issues involving examinees. These issues (whether or not to allow item review, how to set time limits, how to address examinee anxiety, test taker motivation, and test equity) are areas that must be explored for fair and useful tests. (SLD)

Overview of Practical Issues in a CAT Program

Steven L. Wise, University of Nebraska-Lincoln

Paper presented at the 1997 annual meeting of the National Council on Measurement in Education, Chicago, IL

Overview of Practical Issues in a CAT Program

Computerized adaptive testing (CAT) has become increasingly common in large-scale testing programs. The primary advantage of a CAT to test developers and administrators is its promise of efficient testing; in theory, examinee testing times can be dramatically reduced while maintaining the quality of measurement provided by traditional paper-and-pencil tests. This advantage is particularly attractive to testing programs that have traditionally administered long tests. In such testing contexts, the potential problems of examinee fatigue and consequent diminished effort can be alleviated by use of a CAT.

Virtually all operational CATs use measurement methods based on item response theory (IRT) in the selection of test items and the estimation of examinee proficiency. The invariance principle of IRT allows one to administer different sets of items drawn from a unidimensional item pool to different examinees, yet estimate their relative levels of proficiency on a common scale of measurement. The CAT's efficiency is realized through the targeting of item difficulty to examinee proficiency. Such items, according to the principles of IRT, provide maximal information in proficiency estimation.

The CAT procedure is basically a two-step process. At step one, an item is chosen whose difficulty is matched to the examinee's current (or initial) proficiency estimate. At the next step, the examinee's response to the administered item is scored and the examinee's proficiency estimate is updated. These two steps are then repeated until some stopping criterion is met, which is usually a predetermined number of items or a desired level of measurement precision. By this process, the CAT algorithm converges on a final proficiency estimate for the examinee.

Although, in theory, CAT is a relative simple idea, the reality of the planning, implementation, and maintenance of a CAT program is substantially more complex. The purpose of this symposium is to provide a discussion of the challenging practical issues that must be addressed in planning, implementing, and maintaining a CAT program. Each symposium participant has extensive experience in designing and managing CAT programs. On the basis of discussions among the participants, a set of practical issues has been developed; this set has been subdivided into four major clusters. Each of the symposium presenters will focus on a particular cluster, providing (a) a presentation of the particular practical issues that managers of CAT programs are likely to face, (b) a discussion of the theoretical and empirical research relevant to each issue, and (c) provide recommendations for measurement practice regarding each issue.

The following is a listing and brief description of the relevant practical issues that are likely to be faced by the developers and managers of a CAT program. It should be noted, moreover, that these issues are substantially interrelated and decisions made regarding one issue are likely to influence or constrain the decisions made regarding other issues in the list.

Cluster 1: Item Pool Development and Maintenance

- Pool Specifications. This issue involves planning an item pool that (a) matches the content areas in the test specifications, (b) has a sufficient number of items per content area, and (c) has an adequate distribution of item difficulty within each area.
- Choice of IRT model. The choice of IRT model to has important implications regarding (a) how much data are needed for adequate item calibration and (b) CAT item selection strategies.

- Collecting the item calibration data. On one hand, to calibrate item data based on paper-and-pencil administrations of operational tests requires the assumption that the paper-and-pencil and computerized versions of each item will have the same IRT parameters. On the other hand, to develop computerized test forms and administer them to examinees in a non-operational (i.e., no-stakes) test administration invites the problem of low examinee motivation affecting item parameter estimates.
- Pool Dimensionality. This poses somewhat of a contradiction. Virtually all CAT programs are based on unidimensional IRT models. Yet the specification of different content areas in the item pool implies that the item data will be multidimensional. How does the test developer address the various content areas while maintaining adequate unidimensionality?
- Adding items to the pool. As the CAT program matures, there will be likely be a need to add new items to the pool. It is challenging to design strategies for gathering the data needed for calibrating these new items.
- Deleting items from the pool. There will also likely be a need to retire items from the pool. What criteria should be used to make this decision?
- Recalibrating item parameters. It is likely that the IRT parameters of at least some of the items will change over time. How can data be collected to re-calibrate the parameters of the items in the pool?

Cluster 2: Administering and Scoring the CAT

- Proficiency estimation method. Which method will be used to estimate examinee proficiency? Common choices are maximum likelihood, Bayesian, or modal Bayesian. If a Bayesian method is used, what prior distribution should be specified?

- Initial test item(s). What should be the difficulty level of the initial CAT item(s)? How does one avoid exposure issues with the initial item(s)? How large should the difficulty step-size be for the first few items administered?
- Content Balancing. How should the items be administered to maintain content balancing congruent with the test specifications?
- Item selection. Which methods of item selection should be used in identifying items to administer from the pool?
- Stopping criterion. Should a fixed number of items be administered to each examinee, or should each examinee receive enough items to reach a prespecified level of measurement precision (i.e., reliability)?
- Item constraints. An item that has been administered to an examinee may provide cues to the correct answer of other items in the pool. Should the pool be constrained to not administer any of these items?
- Item review. Should examinees be allowed to review, and possibly change, their answers to previously administered items?
- Time limits. How does one establish a time limit that is fair to all examinees? If item review is allowed should the time limits be sufficient for all examinees to have an opportunity to review?
- Equating CAT scores to paper-and-pencil tests. In many testing programs, both paper-and-pencil and CAT versions will be used. How does one equate the scores from these tests? Under which circumstances should a paper-and-pencil test be used in lieu of a CAT?

Cluster 3: Protecting the Integrity of the CAT Item Pool

- Pool security. The higher the consequences associated with a CAT, the more likely that persons or organizations will try to acquire information regarding the particular items in the pool.

- Exposure control. A key aspect of pool security concerns the relative frequency with which items are administered from the pool. The more frequently that an item is administered, the more likely that it can become "known" to an examinee in advance.
- Test disclosure. Testing programs are sometimes forced to publicly disclose information about the item pool. How can this be accomplished while maintaining pool security?
- Coaching. An inevitable outcome of a high-stakes CAT program is the emergence of coaching schools directed toward preparing examinees to take the CAT. While some coaching schools provide legitimate test preparation, others seek to develop an extensive knowledge of the CAT pool, or teach examinees strategies to "beat" the CAT.

Cluster 4: Examinee Issues in CAT

- Item review. This issue is by far the area of greatest concern expressed by examinees. It represents another dilemma, however. Providing item review detracts from the efficiency of the CAT, both in terms of testing time and of item targeting. On the other hand, there are decades of research indicating that allowing examinees an opportunity to review, and possibly change their answers, is likely to legitimately increase test performance.
- Time Limits. Establishing a reasonable time limit for a CAT is challenging because (a) examinees may receive tests of different lengths and (b) examinees will receive tests of different average difficulty—which may require differential amounts of time to complete.
- Examinee anxiety. Increased anxiety during a test has been shown to lower test performance. What characteristics of a CAT are potentially anxiety increasing?

- Motivation. Without consequences associated with test performance, many examinees will not try to do their best on a test. This issue has implications for establishing an item pool, which should be developed under consequential conditions.
- Equity. Examinee subgroups may react differently to a CAT administration, which may confound test performance and threaten score validity. What aspects of a CAT are most likely to pose difficulties?

TM026583

AERA CONFERENCE 1997

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: OVERVIEW OF PRACTICAL ISSUES IN A CAT PROGRAM

Author(s): STEVEN L. WISE

Corporate Source: UNIVERSITY OF NEBRASKA - LINCOLN

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all **Level 1** documents

☒

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all **Level 2** documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

☐

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Steven L. Wise	Printed Name/Position/Title: STEVEN L. WISE
Organization/Address: DEPT. OF EDUCATIONAL PSYCHOLOGY UNIVERSITY OF NEBRASKA LINCOLN, NE 68588-0345	Telephone: 402/472-2736 FAX: E-Mail Address: SWISE@UNLINFO.UNL.EDU Date: 3/25/97

(over)