23 The Use of Computers in Assessment

BRIAN E. CLAUSER
National Board of Medical Examiners, Philadelphia

LAMBERT W.T. SCHUWIRTH
University of Maastricht

SUMMARY

Computers have had a pervasive impact on contemporary life, an impact that has been felt strongly in educational assessment. Interestingly, although the potential advantages of computerized testing have been studied and discussed for more than two decades, it is only within the last few years that computers have begun to have a large-scale practical impact on assessment. This chapter will examine the role of computers in medical education assessment. The chapter is broadly divided into three sections. The first examines the use computers as a platform for the delivery of standard testing formats (e.g., multiple-choice items). In this application, the computer eliminates costs associated with printing test materials, allows for enhancements such as audio and video, and may allow for increased flexibility in scheduling and administration. The computer also facilitates the implementation of various types of adaptive testing. Computer-adaptive testing (CAT) is a means of improving the efficiency of testing by targeting the test to the individual examinee. Most computer-adaptive testing procedures are based on item response theory (IRT). A brief (and relatively non-mathematical) description of IRT will be presented followed by a discussion of the logic and potential of CAT. This section will end with a description of some recent innovations related to computerized testing, including alternative item selection and test construction methodologies and procedures for automated test assembly.

The second section of this chapter will focus on computer simulations for use in assessment. As with other aspects of computer-based testing, discussion, planning and research date back decades but results from large-scale implementation are only now becoming available. This section begins with an historical perspective, describing the precursor of the computer simulation, the paper-and-pencil based patient management problem. Two examples of currently available simulations are described. A significant question with the use of computerized patient management simulations is, "How should the performance be scored"? A conceptual framework

International Handbook of Research in Medical Education, 757–792.
G.R. Norman, C.P.M. Van der Vleuten, D.I. Newble (eds.)

for considering scoring approaches is described and recent published results are summarized. Validity and reliability issues are discussed for assessments based on computerized simulations.

The third section of this chapter briefly describes what may be the future of computers in assessment. Research in the areas of natural language processing and virtual reality provide a basis for sensible speculation about the potential for the use of technology in future assessment. Recommendations are made to temper enthusiasm so that technology may be a means to improved assessment rather than an end in itself.

INTRODUCTION

This chapter examines the role of computers in medical education assessment from three perspectives. The first section considers the use of computers as a platform for the delivery of standard testing formats. In this application, the computer may eliminate costs associated with printing test materials, allows for enhancement of the stimulus materials, and provides increased flexibility in scheduling and administration. The computer also facilitates the implementation of various types of computer-adaptive testing. The second section focuses on computer simulations for use in assessment, including a detailed description of two simulation formats that are currently in use. A significant question with the use of computerized patient management simulations is, "How should the performance be scored?". A conceptual framework for considering scoring approaches is presented. The third section describes what may be the future of computers in assessment. Research in the areas of natural language processing and virtual reality are presented as a basis for speculation about the use of technology in future assessment. Recommendations are made to temper enthusiasm so that technology may be seen as a means to improved assessment rather than an end in itself.

COMPUTER DELIVERY OF STANDARD TEST FORMATS

Logistical advantages and disadvantages of computer administration

The use of computers for test administration offers both theoretical and practical advantages. Practically speaking, paper administration of an examination requires printing test books. This simple requirement has numerous implications. In medical evaluation, it is often appropriate to have high-quality color photographs included in the stimulus materials. However, color printing remains a significant expense; so much so that testing programs may be forced to limit or eliminate the use of this sort of stimulus. With computer delivery, presenting colored stimulus materials requires only that the photograph be digitized. Eliminating the printing costs may

practically remove restrictions on the use of color pictures which are only one type of stimulus material that the computer facilitates. Although audio and video have been used in certification tests, presentation has been problematic. As with color, the computer makes presentation of this type of material relatively convenient, assuming that the required hardware and software are available.

In addition to the costs and practical limitations in terms of the nature of stimulus materials, printed test booklets may create a substantial logistical challenge. For large-scale high-stakes testing, a significant problem is created in the distribution of the booklets. To ensure security, booklets must be printed, shipped, and stored at the administration site, under secure conditions. The loss of a single copy of the test (for even a brief period) calls into question the security of the examination. In the common circumstance that test materials are to be reused, the problem does not end with the test administration. Security must continue after administration while test materials are collected and shipped back to the test administrator. Again, if even a single copy is missing, the security of the test material may be compromised.

By contrast, when the test is administered on computer, encryption procedures can be used to ensure that the test material is only available during the period that the authorized examinee is viewing the item on the screen. Systems involving separate passwords given to the test administrator and the examinee can be implemented to ensure that the material is only decrypted under authorized conditions. Scoring can be completed on-line or an electronic file can be returned to be scored. In either case, there is no need for further transportation of the secure test materials and the encrypted files can simply be deleted when testing is complete.

Computerization also has the potential to offer flexibility in test administration. For large-scale assessments, tests are typically administered on a relatively small number of dates annually. Administrative efficiency and printing costs dictate this restriction. By contrast, when examinations are administered via computer-testing, centers can be established and test forms can be administered at the examinee's convenience. This flexibility may also appear highly advantageous in small-scale (intramural) testing contexts. Testing on demand makes it convenient to establish educational programs in which students work at their own rate, demonstrating mastery at each level before moving to the subsequent unit.

This section has described some of the "promise" of computer-based testing from a practical and logistical perspective. Before moving on to examine the additional theoretical advantages of computerization, a few comments on the practical limitations are warranted. Clearly computerization eliminates printing costs. However, it does introduce other administration expenses. The evaluator must either establish the required computer center(s) or pay for administration at commercial centers. In general the cost of computer administration will exceed the cost of paper administration for large-scale testing (when economy of scale makes printing more efficient). In smaller-scale administrations, where the cost of high-quality printing may be much higher per examinee, this relationship may shift.

Although certain aspects of administration may become more convenient because of computerization, issues arise which may not have an analogue in paper administration. For example, when color pictures are part of the test material, it is possible to print the pictures so that each test booklet will contain an accurate and uniform representation. By contrast, the image on a computer screen may vary from machine to machine, depending on specifics of the hardware and software, and also on monitor adjustments. Similar issues may arise when audio or video are included in the stimulus materials. Even when the test is entirely text based, the specifics of the computer equipment may lead to concerns about the standardization of administration (e.g., the size and quality of the monitor may influence the readability of the material).

Another issue that arises in computerization of assessment is the possible impact of examinee computer experience. When examinees differ in terms of how often they use computers, issues of comfort and familiarity may become a concern. In some circumstances there may be examinees who experience anxiety when using a computer. Although there is empirical evidence suggesting that computer experience may have a negligible impact on outcomes, the issue requires consideration when planning a computerized assessment.

As described previously, computerization may dramatically reduce the vulnerability of a testing program to certain types of security breaches. If test booklets are not printed, there is no need for concern that a booklet may be taken, copied, and distributed before or after the administration. If the test is administered as a single simultaneous sitting for all examinees, this may create the optimal security conditions. However, if computerization means moving from occasional administrations to ongoing availability, it will most likely be necessary to reuse items from day to day. This creates the possibility that examinees may memorize or otherwise carry away specifics about the items. Organized efforts to collect information may lead to subsequent examinees having information about large numbers of items. Varying test forms administered and constructing substantial numbers of forms from a pool containing a large number of items may effectively minimize the advantage that any individual is likely to gain from information gathered by other examinees. The drawback is that item writing is typically expensive and writing the items required to produce numerous forms with minimal overlap will be costly.

Finally, the convenience of scheduling may prove to be a limited advantage. No doubt some examinees will find a single scheduled test date inconvenient. Illness or other personal conflicts my be problematic. But the other side of the convenience of examinee scheduling may be that when all (or most) examinees wish to test during a brief period, the available computer facilities may be insufficient to accommodate them. There are likely to be circumstances in which the guarantee of test availability during some critical period is more attractive than the convenience on self-scheduling without that guarantee.

Theoretical advantages of computerized testing

For more than two decades, researchers have examined the potential advantages associated with computer-adaptive testing (CAT). The logic behind CAT is that each item response provides information about the examinee; however, the amount of information provided varies. The information provided by an item is maximized when the difficulty of the item is well matched to the proficiency of the examinee (i.e., when the probability of the examinee responding correctly is about 0.5). The mathematical definition of information and the logic behind this relationship are complex but it is intuitively apparent that if a highly proficient examinee is presented with a very easy item and answers correctly, the evaluator has learned little. By contrast, if the highly proficient examinee responds correctly (or incorrectly) to a very difficult item, this response provides some basis for raising or lowering the estimation of that examinee's proficiency. CAT allows the evaluator to select a set of test items that is optimally targeted to the examinee's proficiency. In effect, this approach selects a set of items that minimize measurement error in the vicinity of the individual examinee's estimated proficiency.

Item response theory

Classical test theory provides a variety of extremely useful tools for evaluating the performance of tests and test items. One limitation of this framework is that the resulting conclusions about test and item performance must be interpreted in the context of the specific sample of examinees whose responses were evaluated. Similarly, inferences about examinees must be interpreted in the context of the sample of items which they completed. For example, to describe the difficulty of a test item (scored as 0 or 1) in the classical test theory framework, an index would be calculated representing the proportion of examinees scoring 1 on the item. This index, referred to as the p-value, is informative, but it may vary dramatically if the item is re-administered to a more (or less) competent sample of examinees. Similarly, a description of examinee proficiency is likely to take the form of a percent correct or number correct score. This value will also vary if the examinee is retested with a sample of more (or less) difficult items.

In contrast to classical test theory, IRT provides a means of estimating item difficulty which is theoretically invariant with regard to the sample of examinees that responded to the item. Similarly, it provides an estimate of examinee proficiency which is invariant to the sample of items used to measure the examinee. Practically, it is this invariance which makes proficiency estimates based on computer-adaptive testing possible. This means that a potentially unique set of items can be administered to each examinee to optimize the information available about that examinee's proficiency.

In the typical conceptualization of CAT, an examinee is presented with a small number of items. Based on the responses to those items an estimate is made of the examinee's proficiency. The next item is then selected to provide maximum

http://www.springer.com/978-1-4020-0466-7

International Handbook of Research in Medical Education

Norman, G.R.; van der Vleuten, C.P.M.; Newble, D.I. (Eds.)

2002, XIII, 1106 p. In 2 volumes, not available separately., Hardcover
ISBN: 978-1-4020-0466-7