12. Computerized Testing In Licensure

C. David Vale
Insurance Testing Corporation

Follow this and additional works at: http://digitalcommons.unl.edu/buroslicensure

Part of the Adult and Continuing Education and Teaching Commons, Educational Assessment, Evaluation, and Research Commons, and the Other Education Commons

http://digitalcommons.unl.edu/buroslicensure/18

This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Licensure Testing: Purposes, Procedures, and Practices by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Computerized testing has come out of the laboratory and into the field. By rough estimates, over a million licensure and certification examinations are currently given by computer each year, and the number is rising. Computerized testing is not appropriate for every application, however. Computerized tests always result in significantly greater direct costs than paper-and-pencil tests. To justify their use, a computerized test must result in a net dollar saving. This means that something in the process of computerization must offer a cost reduction that more than offsets the direct cost of computerization. The purpose of this chapter is to identify the areas in which computerization can result in dollar savings and to help the reader determine if, and in what form, computerized testing is appropriate for a specific application.

It may be possible to make the case that a computerized test is useful because it can implement new question types or questioning strategies and thus measure something that cannot be measured by other means. Such an application has yet to be demonstrated in licensing. This chapter will thus ignore this possibility, dealing exclusively with the use of computerization of traditional test questions as a means of saving costs.

SCHEDULING EFFICIENCY—AN OBVIOUS ADVANTAGE

The success of computerized testing in licensure today is due in large part to the scheduling improvements it has offered. Consider a typical paper-and-pencil license testing program: Tests are given every 2 weeks and must be scheduled 2 weeks in advance. Say a candidate decides on October 1 to take a licensure test. The scheduling deadline for the October 14 test has just passed and the first test available is October 28. The candidate takes and fails the test, learns of the failure on November 10, and must reschedule for November 25. A typical computerized testing program is different: Tests are given daily and candidates
need to register only one day in advance. Thus, the candidate could fail the first
test on October 2, study hard that night, and take the retest on October 3. Assuming
the candidate passed the second time in either scenario, the result of computerization
would be a time saving of almost 2 months. If passing a test stands between
a candidate and a career, a 2-month time saving can be significant.

Why does a computerized testing program offer such scheduling improve­
m ents? The direct costs in a testing program can be divided into five categories: (1) registering a candidate to take a test, (2) providing a place for the candidate to take the test, (3) providing a medium on which to present the test, (4) providing someone to proctor the examination, and (5) scoring and reporting the results. An optimal administration design must balance all five of these categories. If the criterion for design is minimal cost, the least expensive combination of elements must be found.

Paper-and-pencil administration offers significant freedom to choose a low­cost design. The minimal expense in administration is achieved by requiring the candidate to mail an application and a check (avoiding telephone and credit-card charges), administering the test in idle space that is normally used for other purposes (e.g., Saturday in a high-school cafeteria), presenting the questions on an inexpensive medium (e.g., paper), using part-time personnel earning supplemental (lower wage) income to administer the test, and limiting expensive equipment to a single site (e.g., scoring and reporting results from a central office). The optimal economic design results in the often seen massed administration of paper-and­pencil tests and 2- to 4-week advance registration requirements.

A computerized testing program has less freedom in design. The media for test presentation are not readily portable; this suggests implementation in a permanent site. The media, as well as the space to store them, are relatively expensive; this suggests that relatively few be used. When the costs of equipment and space are balanced against the cost of proctoring, small, frequent sessions usually result. In its optimal configuration, computerized administration is significantly more expensive than paper-and-pencil administration. Historically, this naturally gave rise to the offering to candidates of improved services such as rapid scheduling and score reporting.

Computerized administration is not essential to achieve the scheduling advantages typically obtained through computerized testing. However, when the design appropriate for computerization (and yielding the scheduling advantages) is applied to paper-and-pencil testing (e.g., small, frequent sessions; rapid scheduling; onsite score reporting), its costs are nearly as great as full computerization. The direct cost of a computer system adequate for implementing multiple-choice licensure tests is only about $300 per testing station per year, which translates to about one dollar per test in a center that gives one test per station per day. Thus, if daily testing is implemented, the additional costs of computerization are small.

Scheduling improvements, from a scientific perspective, are not very interesting. Psychometric journals rarely publish articles documenting the time saved through efficient handling of candidates. As a point of comparison with psychometric savings discussed below, however, remember that the time savings achieved through scheduling improvements are on the order of 1 to 2 months.
Note, however, that these time savings translate into dollar savings only when the time has value. Time typically has great value when a candidate must pass a test to get a license to practice a profession. When the translation is achieved by comparing the earning power of an unemployed individual with that of a licensed individual, the figures are large enough to defy belief. Anecdotal experience suggests that these savings are meaningful to licensure candidates. Time has less value if the candidate can practice the profession on a provisional license while attempting to pass the test. Similarly, time has less value to certification candidates than to license candidates because the connection between having the certification and earning money is less direct. If the decision to computerize a test is based on the improvements possible in scheduling efficiency, it is wise to first verify that the time saved is truly valuable.

**SOME PRACTICAL ISSUES IN COMPUTERIZATION**

Although the time savings through changes in the approach to scheduling may appear to strongly recommend the computerized approach, there are some practical issues that should at least be considered before embarking on the path of computerization.

**Computer Anxiety and the Unique Nature of the Medium**

Two concerns have been expressed since computerized tests were first proposed: First, are the results of a computerized test comparable to those of a paper-and-pencil test? Second, will the computer create undue anxiety in the examinees that will affect their performances on the tests?

The answer to the first question is relevant only if a test is administered in both computerized and paper-and-pencil modes. In that case, fairness is an issue. However, if a test is only administered in computerized mode, the fairness issue does not exist. The paper-and-pencil mode is in no sense a standard to which the computerized mode will be compared.

Nevertheless, studies comparing the two modes have found differences to be rare. Kiely, Zara, & Weiss (1986) found no differences between modes for unspeeded Armed Services Vocational Aptitude Battery (ASVAB) subtests, when the entire item fit on a single computer screen. Even items containing graphics showed no difference. The differences they found were for reading-comprehension items that required the candidate to scroll the screen to see the passage. White, Clements, and Fowler (1985) found comparable scores on the Minnesota Multiphasic Personality Inventory (MMPI) administered in both modes, although they noted that the availability of a “cannot say” response on the computer resulted in significantly more omits. Lukin, Dowd, Plake, and Kraft (1985) found no differences between scores on measures of anxiety, depression, or psychological reactance across modes. Moreno, Wetzel, McBride, & Weiss (1984) found arithmetic, vocabulary, and reading comprehension tests of the ASVAB similar across modes. Greaud & Green (1986) did, however, find a substantial difference between modes for a speeded test. Thus, to summarize, if the computer simply presents text (or high-quality graphics), the candidate is not rushed (i.e., the test is
not speeded), and no complicated manipulations (e.g., scrolling a long screen) are required, the results from the two modes are psychometrically equivalent.

Regarding the question of computer anxiety, although it undoubtedly exists in isolated cases, it is not prevalent. Burke, Normand, and Raju (1987) found no difference in anxiety for the two modes. They also found that examinees preferred taking the test on the computer. White, Clements, and Fowler (1985) found that examinees preferred the computerized mode. Lukin et al. (1985) found that 85% of examinees preferred the computerized mode. Wise, Barnes, Harvey, and Plake (1989) found that neither the degree of anxiety toward computers nor the amount of experience with computers had any effect on test scores. In summary, examinees tend to prefer the computerized mode of administration and do not appear to suffer anxiety toward it.

Availability and Economics of Computerized Testing Centers

Recall that the significant advantages obtained through computerized testing result from the rapid, convenient scheduling and the immediate availability of results. It is easy, with commercially available software, to set up a computer to administer a test, even an adaptive one. It is quite feasible to set up a local area network and collect results from multiple testing stations in a database. But it is a major endeavor requiring significant testing volume to set up a cost-effective wide-area testing network complete with the management and support personnel necessary to operate it.

How does such a network operate? Consider as an example ITC’s (Insurance Testing Corporation) network of insurance testing centers. All exam registration (money collection) and scheduling is done centrally in St. Paul, Minnesota. Candidates can register for their exams by mail or by phone (paying with a credit card). Candidates who register by phone can schedule their exams in the same call; those who register by mail must call to schedule. All scheduling is done interactively; candidates do not express preferences for dates and times with their mail registrations. Candidates can take their tests at any of the 58 centers in the network at any time a seat in the chosen center is available.

The testing centers consist of testing computers connected to redundant network servers through a local area network. The server computers contain the tests. All test material is encrypted using the National Bureau of Standards’ Data Encryption Standard (DES). The servers are also stored in a thick steel safe that is bolted to the floor of the testing center.

Each night, when the registration phone center in St. Paul closes, testing schedules are assembled for each of the testing sites. These schedules are sent electronically to each of the sites using fast modems and standard phone lines. (Except for periodic modem communication, such as this, the sites operate autonomously.) Typically, the test item banks are stored at the sites and only test assembly information is sent with the schedule. If a candidate chooses to take a test at an out-of-state location (e.g., a Pennsylvania test at an Oregon center), the complete test will be sent; only those tests administered frequently are stored at a site.

The next morning at each center, 30 to 45 minutes before the first scheduled test of the day, a test proctor logs into the testing center’s computer system by
entering a password. As part of the log-in process, electronic mail sent from St. Paul is displayed for the proctor to read. The system is then ready to administer tests. At that point, the testing system initiates a call to St. Paul to communicate that it is up and running. (If sites do not report in 30 minutes prior to the first scheduled test, alarms go off in St. Paul.)

As candidates arrive, their identifications are checked, they are seated at testing stations to take their tests. The testing stations are standard personal computers with slightly modified keyboards; the relevant keys are color coded and a few of the key descriptions have been changed. Although the proctors generally explain everything a candidate will need to know to take a test, each candidate receives an on-line tutorial that provides the detail essential to taking the test.

When a candidate finishes a test, his or her results are presented on the computer screen. A paper copy of the score report is printed at that time and is usually ready by the time the candidate emerges from the testing room. In some states, these score reports are considered official. In most, however, the communication of results to the states is electronic.

When a site closes for the day, test results for all candidates who tested are electronically communicated to St. Paul. There they are stored in a database and assembled for reporting to the states. This reporting generally takes place the next morning, less than 24 hours after the test was taken.

Figure 1 shows the direct cost of operation of 45 testing centers, for which cost data were available as of this writing, as a function of center size. This figure was based on data through the first 9 months of 1993. The abscissa represents the number of testing hours per year. The ordinate is the cost per hour of testing. (Actual dollar values are not included as they are considered confidential information.) As may be intuitively obvious, the cost per hour drops as the testing volume
at a site increases. This is because certain fixed costs of establishing a center need to be paid, whether tests are given or not (e.g., rent). Although some aspects of the fixed costs can be tailored to the anticipated volume of the site (e.g., the amount of office space), others cannot. In ITC’s centers, fixed costs that do not vary according to the volume of the center include costs of the redundant network servers, a steel safe in which to put the network servers, and a telephone line. Also, the time to open a center (45 minutes before the first candidate arrives) is the same regardless of whether 2 or 60 tests are given that day.

The costs shown in Figure 1 are for centers that have been optimized for cost to the greatest degree ITC’s center concept would allow. Even so, costs rise dramatically as the annual testing hours fall below about 1,000. Political, rather than economic, concerns require ITC to have a few such centers. For insurance tests, ITC has found that an average of between three and four centers per state is needed. A national testing program giving 2-hour tests would have to administer almost 90,000 exams per year to get to the 1,000-hour point, where the cost curve flattens out. This is an optimistic figure, however, because it is unlikely that any program will be able to evenly distribute its examinations across centers.

As of this writing, there are two testing networks available to administer tests that are national in scope. One is operated by Sylvan-Kee Systems. The other is operated by Drake Training and Technologies. The ITC network is also available in specific regions, but does not approach national scope. This means that the choice of testing networks for the implementation of computerized tests is somewhat limited. Although the costs of using such a network vary by application and vendor, the number of vendors and available testing stations has not grown large enough yet that national computerized testing services are a commodity.

The availability of testing networks is a key issue in the implementation of a computerized test. Although the economics of time suggest that candidates will support rather hefty fees for the convenience of computerization, it remains to be seen in practice how high a fee candidates will endure without complaint. Fees as high as $30 per hour are occasionally mentioned for national service of small programs; but because the actual fees are negotiated and usually private, exact numbers are difficult to pin down. In the case of insurance and real-estate candidates, a mandatory per-test increment of $30 ($10 to $15 per hour) for computerization does not seem to cause problems. Whether candidates would readily accept a per-test surcharge of $75 to $100 is an empirical question.

Legal Defensibility of Computerized Tests

Perhaps the most comprehensive review of the potential legal challenges to a computerized test is contained in a compendium entitled “Collected Works on the Legal Aspects of Computerized Adaptive Testing” (NCSBN, 1991), a collection of works commissioned by the National Council of State Boards of Nursing in anticipation of its effort to implement computerized adaptive forms of the examinations it publishes for the licensure of Registered and Licensed Practical nurses. After pointedly noting that there was no case law directly on point (because no one had yet been sued over a computerized test), the contained works consider the
possible mechanisms of legal attack on computerized or adaptive tests. This discussion is largely drawn from that document and a paper by Mehrens and Popham (1992); readers interested in further details are directed to those sources.

In considering the possibility of legal challenge, it is worth noting that the successful suit will not be based simply on a candidate's distaste for computers or tests, but must have some basis in law. There are relatively few laws on which a challenge can be based. The first possibilities are the 5th and 14th amendments to the United States Constitution. The Constitution prohibits the federal and state governments from denying life, liberty, or property without due process of law and requires these governmental units to provide all citizens with equal protection under the law. A license is considered property.

As discussed by O'Brien (1991), constitutional cases are difficult to make. First, the due process principles require only that the requirements for allowing an individual to practice a profession bear a rational relationship to his or her fitness to do so; historically this has only required that the examination ask questions related to the domain of knowledge required by the profession. Second, claims alleging violation of the equal protection requirements must prove intent; if a process appears neutral, it need bear only a "fair and substantial relationship" to the competence required by the license. Thus, a challenge to a computerized test on constitutional grounds is likely to be successful only if it can be shown that it was intentionally used to discriminate unfairly or to deny a license.

Beyond Constitutional grounds are statutory ones. Title VII of the Civil Rights Act of 1964 significantly extends the equal-protection concept for minorities and other protected classes. Title VII allows a case to be made if discrimination occurs, even if it is not intentional. Furthermore, its application is not limited to governmental units. Finally, the Rehabilitation Act of 1973 and the Americans with Disabilities Act (ADA) of 1990 prohibit discrimination against people with disabilities and require reasonable accommodation of such individuals.

In general, the research literature has not shown that computers discriminate against minorities. The challenges to computerization appear far more likely to be based on ADA. Accommodations for physical disabilities have long been made by most organizations offering licensure tests in any mode. The ADA brings mental disabilities more to the forefront, however. As O'Brien (1991) points out, the ADA may require the accommodation of computer-phobes, a subgroup of test-phobes. Practical experience suggests that learning disabilities are a frequent source of requests for alternate testing modes. Legally, if a licensed professional supports a candidate's request for an alternate testing mode, there appear to be two defenses for denying it. First, the accommodation must be "reasonable." This implies that the accommodation should not compromise the integrity of the test and that it should not be outrageously expensive; of course, what compromises the test or constitutes outrageous expense may be the subject of litigation. Second, the individual should be otherwise "qualified." Although case law with respect to the Rehabilitation Act of 1973 seems clear that this means an individual must meet all of the requirements for a license in spite of a handicap, not except for it (O'Brien, 1991), case law has not developed with respect to ADA. Current belief is that an
individual should not be barred from taking a test simply because he or she will be unable to meet other requirements for licensure (Warren, 1992).

Does this present special problems for a computerized test beyond those that exist in a paper-and-pencil test? Potentially, it does. Although candidates are equally free to request alternate forms of any test based on their disabilities (e.g., oral, rather than paper and pencil), requesting a paper-and-pencil form rather than a computerized one is a relatively frequent request. If the test is pre-formed, this is only a logistic inconvenience. If the computerized test is tailored based on examinee responses, it may not be feasible to administer a comparable test via paper and pencil.

GREATER EFFICIENCY THROUGH MODERN PSYCHOMETRIC METHODS

Computerization allows tests to be made psychometrically more efficient by tailoring them to the candidates who take them. There are two ways to tailor a test. First, the difficulty of the test items may be adjusted to the ability of the candidate. A test is more efficient if it does not waste time giving items that are clearly too difficult or too easy for the candidate. Second, the length of the test may be tailored to the candidate. There is no point in continuing a test when the measurement is sufficiently accurate to achieve the purpose for which the test was intended. Tailoring the difficulty of a test has typically been called computerized adaptive testing (CAT; Wainer, 1990; Weiss, 1983). Tailoring the length of a test has been referred to by a variety of names including sequential testing (Linn, Rock, & Cleary, 1972; Reckase, 1983; Weitzman, 1982), adaptive mastery testing (AMT; Kingsbury & Weiss, 1983; Weiss & Kingsbury, 1984), and computerized mastery testing (CMT; Lewis & Sheehan, 1990). To properly explore the potential utility of these techniques, however, an appropriate statistical framework is necessary. Item Response Theory (IRT; Hambleton & Swaminathan, 1985; Lord, 1980) offers such a framework.

At this point the reader should be aware of two things: (a) the remainder of this section makes heavily mathematical arguments regarding the utility of adaptive and sequential testing for licensure and certification programs, and (b) the conclusions of these arguments are of interest primarily to those programs that administer several thousand examinations each year. Readers representing smaller programs who would not even consider using adaptive or sequential testing methods can skip the rest of this section without a loss of useful information.

Item Response Theory

Item Response Theory refers to a family of mathematical models that express the probability of an item response as a function of numerical item characteristics and the underlying ability of the examinee. IRT is of use to computerized testing because it both allows the computation of comparable scores when different items are administered to candidates and suggests which items will be most appropriate for assessing the ability of a given candidate.

IRT models differ in the number of abilities they encompass, the number of item parameters they include, the form of the function that relates the item response
to the underlying ability, and the type of item responses they accept. The most general form of IRT model to be widely accepted in practical ability or achievement testing applications is the three-parameter logistic model. It requires a dichotomous (e.g., right/wrong) item response and describes the probability of a correct response as a logistic ogive (an s-shaped function) in three item parameters and one ability parameter. Mathematically, the model is specified in Equation 1.

\[
P(u_g = 1|\theta) = c_g + (1 - c_g)\Psi(z_g) \quad [1a]
\]

or

\[
P(u_g = 1|\theta) = \Psi(z_g) + (1 - \Psi(z_g))c_g \quad [1b]
\]

where

\[
\Psi(z) = 1/(1 + \exp(-z))
\]

and

\[
z_g = 1.7a_g(\theta - b_g).
\]

In Equation 1, \(u_g\) is the scored response to item \(g\): 0 for incorrect, 1 for correct. The ability parameter is represented by the Greek letter theta (\(\theta\)). The item parameters are \(a_g\), \(b_g\), and \(c_g\). The constant 1.7 is a historical artifact that causes the logistic model to closely resemble its cousin, the normal model. It remains as a convenience to those psychometricians who think of \(a\) parameter magnitudes in that scale.

Equations 1a and 1b are mathematically equivalent. Equation 1a is the form typically seen, because it is computationally simpler. Equation 1b is useful for illustration, however, because it is more amenable to a conceptual treatment. To wit, consider that \(\Psi\) represents the probability that the examinee knows the correct answer to the item. This model, in concept, implies that there is a bell-shaped probability (density) distribution relating the relative likelihood that examinees at points along the theta dimension will know the correct answer. This distribution is centered on the difficulty (\(b\) parameter) of the item and its dispersion is related to the \(a\) parameter (the standard deviation of the distribution is \(.588a\)). The probability that an examinee will know the correct answer is equal to that proportion of the distribution that is below the examinee’s ability level (\(\theta\)). Equation 1b then gives the probability that an examinee with ability equal to a value of \(\theta\) will answer the item correctly. This probability is computed as the sum of the probability that the examinee knows the correct answer (\(\Psi\)) plus the joint probability that the examinee does not know the answer (1-\(\Psi\)) and successfully guesses (\(c_g\)).

Figure 2 gives a graphical depiction of several three-parameter test items. The horizontal axis indicates the underlying ability, typically expressed on a standard scale ranging, practically, from about -3 to +3. The \(a\) parameter indicates how well the item discriminates among levels of ability and relates to the slope of the curve. High \(a\) parameters result in steep slopes near the middle of the curve and shallow slopes at the tails. The \(b\) parameter refers to the difficulty of the item and is equal to the point on the horizontal axis that corresponds to the vertical midpoint of the curve (i.e., \([1+c]/2\)). Difficult items have curves that plot toward the right side of
the horizontal axis. The $c$ parameter is the pseudo-guessing parameter, conceptually equivalent to the probability a candidate of very low ability would have of answering the item correctly. Although it is reasonable to expect this to be the reciprocal of the number of alternatives, in practice there is some variability around this value depending on other characteristics of the item.

The $a$, $b$, and $c$ parameters that gave rise to Item 1 were $(.4, .0, .25)$; this represents an item of modest discriminating power, middle difficulty, and probably four alternatives. Item 2 is a more discriminating version with the same difficulty $(.8, .0, .25)$. Finally, Item 3 is like Item 2, but more difficult $(.8, .5, .25)$.

Two reduced versions of the three-parameter model are also popular. If the $c$ parameters are all assumed to be zero, the two-parameter model results. This model is appropriate if it is not possible to answer the items correctly by guessing. If, in addition to holding the $c$ parameters at zero, all $a$ parameters are held to a constant value, the one-parameter logistic or Rasch model results.

In concept, the Rasch model does not seem appropriate for use with multiple-choice licensure items; correct guessing is obviously possible and items probably differ substantially with regard to how well they discriminate (correlate with) ability. There is an ongoing debate among psychometricians, however, regarding which model is practically appropriate (Traub, 1973; Hambleton & Swaminathan, 1985). Although the Rasch model makes some conceptually unappealing assumptions regarding two of the parameters, available statistical techniques do not allow these parameters to be estimated accurately when the three-parameter model is used. It has long been known that the individual parameters are difficult to

![Figure 2. Three Item Characteristic Curves](image-url)
estimate, in part because errors in the estimation of one parameter can be compensated by errors in another and several sets of item parameters can yield models that fit the data about equally well (Thissen & Wainer, 1982). Proponents of the Rasch model would say this suggests using a simpler model. Advocates of the three-parameter model would counter that declaring the parameters by fiat at values known to be incorrect (e.g., zero for the c parameter) is probably more harmful than poorly estimating the parameters using the best techniques available. For analyses presented in this chapter, the three-parameter model has been used exclusively. The analyses are intended to set bounds on the maximum improvement that can be expected through psychometric means; thus, the model that (if its assumptions are met and its parameters are accurately estimated) will give the best results was used.

Regardless of the model, a major appeal of IRT is the method of scoring it allows. The curves shown in Figure 2 are referred to as item characteristic curves (ICCs), item response functions (IRFs), or response likelihood functions. They express the probability of a correct response as a function of ability (or whatever psychological dimension theta may represent). Inversely, they express the likelihood of a level of ability given a correct response. Each item has complementary response functions for correct and incorrect responses. Figure 3 shows the IRF for both correct and incorrect responses to the same item. The increasing function is for the correct response, indicating that the probability of a correct response goes up as ability increases. The corresponding IRF to the incorrect response indicates decreasing probability of an incorrect response as ability rises.

The individual IRF does not allow much of an estimate of ability, based on the item response. If the response is correct, any higher level of ability is more likely. But the utility of IRT is in how it combines IRFs from responses to multiple items. If the assumptions of IRT hold, the likelihood of a pattern of item responses (e.g.,

![Figure 3. Complementary Item Response Functions](image)
those obtained by a given examinee) can be obtained by simply multiplying the individual response functions together. The assumption necessary to allow this is local independence, a character resulting from unidimensionality. In essence, what this means is that if all of the items in a test measure a single trait (in a factor analytic sense), the responses to items given to someone whose ability level is constant (typical, during the course of a test, for most examinees) will be statistically independent. It is a basic tenet of probability that the joint probability of independent events is the product of their individual probabilities.

Figure 4 shows the IRFs for responses to the three items used for Figure 2, two answered correctly and one (the difficult one) answered incorrectly. It also shows the resulting likelihood function. A good estimate of the candidate’s ability is that level of ability corresponding to the maximum of the likelihood function. This is called the maximum-likelihood ability estimate. In this example, the maximum-likelihood estimate of theta is .23. Note that an estimate can be obtained from any set of test items and expressed on this same ability scale; scores thus computed will be comparable, even if they are obtained from different sets of items.

The likelihood function can, without compromising its character as a likelihood function, be scaled to any size that is convenient. One common scaling is to make the area under the curve equal to one. This done, the likelihood can be considered a Bayesian posterior probability density function, indicating the distribution of abilities that would result if all possible candidates with the same set of responses to the same items were plotted. (If the scaling is accomplished without changing the shape of the distribution, an uninformative or uniform Bayesian prior has been applied.) The standard deviation of that posterior distribution is akin to the standard error of measurement (SEM). (It differs in that the classical SEM refers to a distribution of observed scores around a true score and this is a Bayesian

![Figure 4. Likelihood as the Product of Item Response Functions](image)

Figure 4. Likelihood as the Product of Item Response Functions
distribution of true scores around an observed score. They are equivalent, however, if an uninformative prior is applied.) A laudable measurement objective is to minimize the variance of this distribution. This can be accomplished by administering more items, better items, or items more appropriately matched to the examinee.

A useful index provided by IRT is the item information function. Mathematically the information function is the ratio of the squared slope of the IRF to the conditional variance of the item response at a level of theta. The formula for information in the three-parameter logistic model is given by Equation 2 (after Birnbaum, 1968, Eqs. 20.2.3 and 20.4.16).

\[ I(\theta, g) = \frac{\left[ \frac{P'_g(\theta)}{P_g(\theta)} \right]^2}{P_g(\theta) [1 - P_g(\theta)]} \]  

and

\[ I(\theta, g) = \frac{2.89 (1 - c_g) a_g^2 \psi^2 [z_g]}{\psi [z_g] + c_g \psi^2 [-z_g]} \]  

where

\[ P'_g(\theta) = 1.7 (1 - c_g) a_g \psi [z_g] \]  

and

\[ \psi [z] = \exp(-z) / (1 + \exp(-z))^2 \]

Equations 2a and 2b are equivalent. Equation 2a presents a conceptual formulation of information; 2b presents a computational one. The numerator of Equation 2a is the squared derivative of the item response function. As the IRF becomes steeper, the information increases. The denominator is the conditional variance of the dichotomously scored item. Note that the variance of such an item at a point on the theta scale (i.e., the conditional variance) is solely determined by the probability of a correct response at that point.

Practically, information indicates how effectively a given item will reduce the variance of the posterior distribution (and thus the SEM) as a function of the item characteristics and the point on the theta dimension. Figure 5 shows graphs of the information functions for two items. The flatter of the curves (Item 1) is for a middle-difficulty item (.4, .0, .25) with a modest \(a\) parameter. The more peaked of the curves is for a more difficult item (Item 3) with a higher \(a\) parameter (.8, .5, .25). Several things are important to note from the figure. First, items with high \(a\) parameters generally have higher information peaks, indicating that they can do a better job of shrinking the SEM. Second, note that the point along the theta dimension at which the curve peaks varies with the difficulty of the item. Third, note that the higher the information peak, the more rapid the drop-off; items with high \(a\) parameters provide their advantage over a relatively small range of ability.
It may be obvious at this point that the efficiency of a test can be improved by the judicious choice of items. Information could be maximized (and SEM minimized) by selecting those items that provide the highest level of information at the candidate’s level of ability. The fact that the test must be administered to determine what this level is has given rise to the adaptive test, a test that attempts to administer items most appropriate to its estimate of the examinee’s ability at any point in the test. A simple adaptive strategy begins by assuming an initial estimate of ability near the population mean and choosing items and updating ability sequentially throughout the course of the test. At each stage, the next item is chosen based on the current estimate of ability. After each item is administered, the estimate of ability is updated.

Recall that IRT scoring results in a posterior distribution. The mean or mode of this distribution can be taken as an estimate of ability. Its standard deviation can be taken as an estimate of the SEM. In a pure measurement application, the interest is in obtaining a posterior distribution with as small a variance as possible. In classification (e.g., licensure) testing, there is a passing point to be considered. Then the interest is in classifying the candidate on the proper side of the passing score with as little chance of error as possible. Figure 6 illustrates the situation with a cut score. The curve represents the posterior probability density of a 120-item test composed of items with \( a = .5, c = .25 \), and difficulties peaked at the candidate’s ability of \( \theta = .3 \). The probability of misclassification is the proportion of the posterior distribution that falls on the wrong side of the passing point, set here at \( .0 \) and indicated by the arrow. Both the mean and the variance of the posterior distribution are important in determining the probability of...
misclassification. If an acceptable probability of misclassification can be specified, the test can be terminated when the portion of the distribution that overlaps the cut score reaches this level. This is essentially the AMT procedure (Weiss & Kingsbury, 1984).

Applicability of Psychometric Improvements to Licensure

Few in the psychometric community would argue against the utility of adaptive testing or tailored termination (sequential testing), in the proper applications. But the application is critical to determining the utility. For example, the average discriminating power of the item pool (average $a$ parameter) is critical to establishing how much advantage an adaptive test will have over a conventional one. Similarly, an adaptive test excels at providing high information over a wide range of ability, which is more appropriate for a measurement than a classification application. Furthermore, the position of the passing point in the distribution of ability is significant to determining the utility of tailored termination. Rather than attempting to summarize published research results descriptive of specific situations, this section provides a mathematical model that allows the utility of the methods to a specific environment to be ascertained, subject to a few simplifying assumptions.

Consider the concept of an ideal test. The ideal test makes assumptions known in reality to be unduly optimistic. In the results shown below, four such assumptions were made: (a) The items fit the IRT model perfectly; (b) the item parameters are estimated without error; (c) the item pool is very large, in fact infinite in size; and (d) in the case of an adaptive test, the test is adapted perfectly, with no allowance made for the fact that an examinee's level of ability must be known a priori to do this.
Note that these are significant assumptions, but they are directional. No real test could perform any better than a test evaluated under these assumptions. Obviously, there is no advantage to be gained by using items that do not fit the model, by using parameters other than the true ones, by using a smaller item pool, or by adapting a test other than perfectly. Thus, the ideal test provides a bound of how well a test can perform. The bound is useful because, if the ideal test does not provide sufficient benefit to suggest the more complicated adaptive procedure, neither will the real test. Note also that these assumptions favor an adaptive test more than a conventional one; a conventional test cannot take advantage of perfect adaptation. Thus, these assumptions also place a bound on the relative advantage of the adaptive test.

As a meaningful application of the concept, consider the following reasonable application environment: For many licensure and certification examinations, the range and distribution of item difficulty can be tailored as desired. Assume that items are available at any level of difficulty desired by the testing algorithm. Experience with insurance licensure item banks and anecdotal data informally collected from other researchers suggest that a reasonable \( a \) parameter value would be \( .5 \). Similarly, experience suggests that although there is some variability among items, the average \( c \) parameter for four-alternative items is about \( .25 \). Thus, assume \( a \) parameters fixed at \( .5 \) and \( c \) parameters fixed at \( .25 \). Finally, for the first evaluation, assume the passing point is set at \( \theta = .0 \), a value that would (assuming a standard normal distribution of ability) result in a 50% passing rate.

A few characteristics of IRT will assist in the analyses of the ideal test (and allow exact analytic solutions rather than simulated ones). The characteristics, detailed by Birnbaum (1968), are that:

1. The item information functions (Equation 2) can be added together to obtain the test information.
2. Maximum-likelihood ability estimates tend to be normally distributed around a mean equal to the true value of the parameter they estimate (\( \theta \)).
3. The variance of the distribution of maximum-likelihood estimates is given by the reciprocal of the test information function evaluated at the value of the parameter (\( \theta \)).

These characteristics imply, for a mastery decision, that the probability of misclassification for any particular level of ability can be obtained from that portion of the distribution of ability estimates that fall on the wrong side of the passing point. Thus,

\[
P(Misclass|\theta) = \Phi(-|\theta - \theta_c|/SEM)
\]

where

\[
\Phi[x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)dt
\]

and

\[
SEM = \sqrt{\frac{1}{\sum_{g} I(\theta,g)}}
\]

[3]
An ideal conventional test would be constructed of items that provided the most information at the passing point. An ideal adaptive test would be constructed by selecting items that provided maximum information at the ability level (theta) of each candidate tested. The peak of the information function occurs at $b_g = \theta$ for items in which guessing is not possible. For items where guessing is possible, the ideal difficulty (Birnbaum, 1968, Eq. 20.4.22) is

$$b_g = \theta - \frac{1}{1.7a_g} \ln \left( \frac{1 + \sqrt{1 + 8c_g}}{2} \right)$$  \[4\]

Note that, as Equation 4 implies, the ideal difficulty of an item when guessing is possible is somewhat easier than when guessing is not possible.

Thus, a comparison of the classification accuracy of conventional and adaptive tests is quite straightforward using a bank of items that differ only in difficulty. An ideal conventional test is composed of items with difficulty fixed to provide maximum information at the passing point. An ideal adaptive test, composed of items peaked at each candidate’s true ability level, provides a level of information at all ability levels that is equivalent to the maximum level provided only at the passing point by a conventional test.

The comparison of fixed versus variable test length is a bit more complicated, however. Consider the situation in which a fixed-length test is terminated early when the ability estimates and standard errors leave an acceptably small probability of misclassification. This will result in shorter tests for those individuals with ability levels distant from the passing point. The overall probability of misclassification will rise, however, if tests are only shortened. The result that a shorter test leads to higher misclassification probability does not yield a meaningful comparison of fixed versus variable test length. To properly compare fixed-length and variable-length tests, with respect to misclassification probability, either the misclassification probability or the average test length must be held constant. To achieve a truly fair comparison, the items saved by early test termination for candidates with ability levels distant from the passing point must be reallocated and given to candidates closer to the passing point. How should test lengths be optimally distributed?

As a tool for redistribution, consider the derivative of the misclassification probability with respect to test length. This derivative, a function of the test characteristics and the point on the underlying ability (theta) continuum, indicates how much reduction in misclassification probability can be achieved for each item delivered. The derivative, assuming here for simplicity that items differ only in difficulty, is given by Equation 5. (Note that without this assumption of item equivalence, the evaluation of relative test length is not meaningful.)

$$\frac{dP(Misclassification)}{dL(\theta)} = \phi \left| \theta - \theta_c \sqrt{L(\theta)I(\theta,g)} \right| \frac{1}{2\sqrt{L(\theta)}}$$  \[5\]

where $I(\theta,g)$ is the information provided by any of the equivalent items at ability $\theta$ and $L(\theta)$ is the test length in items. (Note that for a fixed level of theta, the
information value for the items will be constant; a conventional test will have all items peaked to provide maximum information at the passing point and an adaptive test will have all items peaked to provide maximum information at the candidate’s ability level.) This derivative indicates where to get the “most bang for the buck” in terms of items administered. In concept, optimal allocation can be achieved by taking test length from where it will do the least good (low derivative) and putting where it will do the most good (high derivative). Note that, for a specific level of theta, the derivative decreases as test length increases. Therefore, a point will retain the highest derivative only until test length reaches the point where the derivative is higher at another point along theta. Although the concept of moving items around until an optimal allocation is achieved is appealing in concept, practically it is difficult and computationally time-consuming. The ultimate objective of such reallocation, however, is to achieve a distribution of test lengths that causes the derivative to be constant.

For a specified constant, Equation 5 can be solved (numerically) for the optimal test length \( L(\theta) \) at any value of theta. The overall test length for theta distributed standard normal is thus:

\[
L = \int_{-\infty}^{+\infty} L(\theta)\phi(\theta)d\theta \tag{6}
\]

For a specified average length \( L \), Equation 6 can be solved (again numerically) for optimal conditional lengths (those that result in a constant derivative and average length \( L \)). The overall probability of misclassification can then be computed, based on the conditional lengths, as

\[
P(\text{Misclass}) = \int_{-\infty}^{+\infty} P(\text{Misclass}|\theta)\phi(\theta)d\theta \tag{7}
\]

Figure 7 shows the misclassification probabilities as a function of test length for all types of ideal test. Both adaptive tests provide minor improvements over their non-adaptive counterparts. Larger differences obtain between fixed and variable-length versions.

Figure 8 shows the transpose of Figure 7, the test lengths required to obtain a given overall probability of misclassification. The distances between the curves indicate items saved by the various testing strategies. Note that a fixed-length adaptive test shows a relatively constant saving of about four items. Figure 9 shows the proportionate reduction in test length of three testing strategies compared to the fixed-length conventional strategy. The variable-length tests show the larger savings, especially when a low misclassification probability is desired.

Thus, in theory, significantly greater savings are possible through tailored termination than through tailored item difficulty. It is informative, however, to look at the optimal distribution of test lengths. Figure 10 shows optimal adaptive test lengths to achieve an average test length of 120. Two somewhat troublesome issues are apparent from Figure 10. First, optimal test lengths near the passing
point ($\theta=0$) exceed 300 items. Although this number of items may be manageable on the part of the examinee, it is sufficiently different from the average or the reasonable low point (at $\theta=1$, a bound outside which roughly one third of the candidates will fall) to cause scheduling difficulty. Perhaps of greater concern, however, is the drop in test length very near the passing point. The optimal length function suggests a sort of triage: Terminate when you are confident a candidate

![Figure 7. Misclassification Probability as a Function of Average Test Length](image1)

![Figure 8. Length Required for Misclassification Probability](image2)
has passed or failed, give a long test if you are not sure, and quickly write off candidates that are too close to call. Flipping a coin to decide the fate of marginal candidates, although mathematically optimal, may be politically unwise. (Note that this problem would not occur in a real test, however, because a number of items would have to be administered to determine that the candidate was too close to call.)

Figure 9. Length Reduction Compared to Fixed-Length Conventional

Figure 10. Optimal Lengths for 120-item Average Length Adaptive Test
The optimization strategy can be altered to fill in the void around the passing point. If the optimization algorithm is so altered, the test lengths required become as shown in Figure 11. (Note that only the variable-length tests are affected by this modification.) The savings resulting from variable termination are uniformly reduced by about 10 items.

The resulting proportionate reductions in test length (with the void filled), compared to a fixed-length conventional test, are shown in Figure 12. As a practical point of comparison, consider a 120-item fixed-length conventional test. This would yield a misclassification probability of .086. At this level of error, a fixed-length adaptive test will reduce test length by about 3%, a variable-length conventional test will reduce it by about 22%, and a variable-length adaptive test will reduce it by about 30%.

Consider practically what this means. If a fixed-length adaptive test is used rather than a 120-item conventional test, it need only be 116 items long. Assuming that the conventional test is a 2-hour test, the candidate will be able to go home 4 minutes early. The real savings are with the variable-length adaptive test. A candidate should come planning to spend 5 hours testing. Typically, the candidate will go home about 3 1/2 hours early. Sometimes the candidate will go home after just a few minutes. Are there any savings? To save, on average, about half an hour, a candidate has had to block out 5 hours rather than 2. Although a testing center of moderate size (10 or more stations) will be able to take advantage of the average for scheduling, it is likely that a variable-length test will still require a longer time block to ensure that everyone can test; this will translate into higher exam fees. Is this a saving? Perhaps a less significant one than the 1 to 2 months saved by simple computerization.

![Figure 11. Length Required When the Void is Filled](image-url)
It can be argued, with some justification, that the above analysis is too harsh on the tailored tests. Specifically, the mastery problem is most difficult when the passing point is set right in the middle of the ability distribution. With the cut set at \( \theta = .0 \), as above, 50% of the candidates would pass. Consider a somewhat simpler classification problem with the cut set at \( \theta = .5 \). In this case, about 31% of the candidates would pass. Figures 13 and 14 correspond to Figures 7 and 8 above (those with the void not filled). Note that all test forms achieve comparable error rates with fewer items, but that proportionate reductions in test length (compared to a fixed-length conventional test) are remarkably similar in relative and absolute magnitude. Even with the cutting score shifted substantially from the center of the ability distribution, the fixed-length adaptive test offers only modest improvement over its conventional counterpart.

Although the ideal test concept has been applied to only two variations of one testing application here, the application seems a reasonable depiction of the typical licensure testing environment. In this environment, there seems to be relatively little advantage available from adaptive testing. Furthermore, to take advantage of the item savings available through tailored termination seems to result in unpredictable variation in testing times to a degree that is unacceptable from an operational perspective. Note that the more simplistic approach of terminating an otherwise fixed-length test when a candidate has clearly passed it will result in less variability. Its disadvantage, however, is only that a few candidates will get to go home unexpectedly early and the net psychometric result will be an increased error rate.

![Figure 12. Length Reduction When the Void is Filled](image-url)
Figure 13. Item Savings With Passing Point at 0.5

Figure 14. Length Reduction with Passing Point at 0.5
Although the above analyses may suggest that tailored item difficulty or test length may not yield great practical advantage in licensure or certification applications where a pass/fail decision must be made, there will be applications where they do not cost much and are useful. For those cases, there are a few additional details of test design worth discussing.

Content Stratification of a Tailored Test

Although most licensure tests make a pass-fail decision based on a single score, many of these tests also report subtest score results. Furthermore, many of these tests stratify their content to a great degree, sometimes associating the content of each item to a point in a job analysis. IRT and the adaptive testing methods discussed above assume a unidimensional test. Stratification implies multidimensionality. What are the implications of such stratification on practical test design?

IRT and tests based on IRT assume that all items in the test measure the same dimension. According to the IRT model, the only selection that should occur is to maximize the precision of measurement—that is, select items with high $a$ parameters, low $c$ parameters, and $b$ parameters near the theta level of the candidate. When a test consists of subtests that clearly measure different characteristics (e.g., arithmetic, vocabulary, and reading comprehension), Thomas and Green (1989) have shown that it is better to measure each characteristic separately and then average the scores on them rather than to treat them all as a single unidimensional test.

Licensure tests generally consist of subtests that measure characteristics that are less distinct. A test of life insurance knowledge, for example, may be divided into subtests on policy forms, policy options, and policy riders. For purposes of conventional test construction, each section may be further subdivided. Yet, factor analysis and all other analyses may fail to confirm any psychometric distinction between even the subtests, much less their subdivisions.

The issues regarding how to analyze these data are quite complex. First, the psychometric perspective would argue for analyzing the test as a whole; psychometrically it hangs together and better item calibration can be obtained by treating it this way. Politically, it would make more sense to calibrate the item bank by subtests; if the subtests are all calibrated along a (single) common dimension, any differences among subtest scores provided in a diagnostic score report are indicative only of measurement error and not actual competence differences. Operationally, it would be best to treat each category of stratification as a single dimension because the simple (nonstratified) adaptive testing strategies could be used within each.

Kingsbury and Zara (1989) have suggested one way of stratifying an adaptive test. Their model is appropriate when the items are calibrated along a single dimension and behave, psychometrically, as if they measure the same thing; the need for stratification is political rather than psychometric. Specifically, what they suggest is that the item pool be stratified according to content and that the percentage of items to be drawn from each stratum be specified. The adaptive procedure then, at each stage, selects the psychometrically best item from the
Another technique for accomplishing stratification has been suggested by Swanson and Stocking (1993). Their technique, more mathematical and less algorithmic, applies a compensatory optimization approach. Items are described by their characteristics (e.g., being an arithmetic item) and a target test profile, in terms of the characteristics, is specified. The characteristics may be differentially weighted. The item selection process then sums the weighted deviations of actual characteristics from the target ones and selects the item that minimizes the summed deviations. Unlike the Kingsbury and Zara approach, the Swanson and Stocking approach does not guarantee stratification precisely as prescribed.

Consider one final stratification strategy, suggested here an unresearched but imminently practical solution to the stratification problem. Consider first the algorithm for the fixed-length case. Begin by grouping all items into content strata and assigning an item quota to each stratum. Begin item selection with an unrestricted search for the best item. Then, as each stratum quota is reached, mark all items in that stratum as unavailable. As the final item is selected, there will be only one stratum that has not reached quota. If the stratum quotas are integers, the exact number of specified items will be drawn. This technique has computational simplicity and exact stratification as advantages over the Swanson and Stocking method. Over the Kingsbury and Zara method it offers the advantage of extending the choices for the psychometric best item while still assuring proper content stratification.

As a modification of this method for variable length, start by assigning quotas based on the shortest test that will be administered. Select items only from strata having at least half an item remaining in their quotas. As the test grows beyond the minimum length, adjust the quotas. Note that stratification can never overfill a stratum by more than half an item so if the test reaches a length where all quotas are integers, the stratification will be exact.

Finally, as a reminder, note that stratification is an issue only if items within a scoring dimension (i.e., an IRT dimension) are considered dissimilar. If political considerations are consistent with psychometric ones, no within-dimension stratification will be necessary; both will agree that the items all measure the same thing and differ only in their psychometric characteristics.

Randomization

Computerized testing, to achieve the scheduling advantages discussed earlier, must be offered on a relatively continuous (e.g., daily) basis. This means that individuals who do not pass the test on the first attempt may be exposed to the test several more times before passing. It is important that each test they take be sufficiently different from the previous ones that their passage is indicative of mastering the domain and not just a specific test. Furthermore, test coaching for a specific test often takes the form of training for the test rather than for the substance of the test. A test for which the exact item content cannot be predicted is effective in reducing the utility of such coaching.
Adaptive tests will, to a degree, be unique on each presentation. In a pure adaptive test, a candidate will receive the same items on a second administration only by answering the items in the same way each time. If the first attempt to pass was not successful, the second attempt using this strategy is not likely to be either. The issue of coaching is still relevant, however, and certain specific patterns of responses to a pure adaptive test will lead to a passing score every time.

Randomization may be introduced into any of the test types discussed above. The precise mechanism depends on the type of test. Consider first an adaptive strategy. A pure adaptive test selects what it considers the most appropriate item for administration at each stage. Randomization may be added by selecting the two, three, or more most appropriate items at each stage and then randomly choosing among them. (See Kingsbury and Zara, 1989, for a description of Randomesque item selection, as proposed for use in a nurse licensure examination.) Scoring, via IRT, is done the same way, regardless of whether randomization is introduced.

Consider next a conventional test, one in which a fixed set of items is administered regardless of the response the examinee makes. In concept the simplest solution is to construct a collection of sets of parallel items. If such a collection were available, in which all items in a set were psychometrically equivalent, parallel random forms could be constructed by randomly selecting one item from each of the sets in the collection to form a test. In practice, it is virtually impossible to assemble truly parallel item sets. Tests drawn as described would have varying psychometric characteristics. If the tests are to be scored using IRT, this is not a problem. If, however, the scoring is to be done by traditional proportion correct, additional psychometric balancing is required.

One approach to psychometric balancing that has been applied in work at the Insurance Testing Corporation (ITC) involves paired random sampling and balancing according to a similarity criterion. Specifically, items are drawn in pairs, resulting in twice as many items as are required for the test. (If the bank is stratified and three items are required from a stratum, six items are drawn from that stratum.) Then the item from each pair that, in concert with the items thus far selected, maximizes a similarity function is selected. The process may iterate to convergence for better balance.

The similarity function for this procedure should reflect the overall parallelism of the tests. One simple function is the difference between the average difficulty of the test and that of the item bank as a whole. More comprehensive functions compare the similarity of the test characteristic curve or the information function of the test to that of the item bank as a whole. Details and performance of the methods are beyond the scope of this chapter, however.

Finally, consider a test with variable termination. It is possible to order the items selected for the conventional test such that, at any stage in the test, similarity with the target is maximal. Unless the items were all equivalent, however, IRT scoring would be required to determine when to terminate the test. If IRT is used for scoring, no psychometric balancing would be necessary.

Passing Points for Tailored Tests

In concept, a passing point for any of the types of computerized tests is no different than a passing point for a conventional test. If a conventional test is
12. COMPUTERIZED TESTING IN LICENSURE

administered on a computer and scored, like a paper-and-pencil test, via the number of items answered correctly, the issues in setting and using a passing point are identical to those faced when using a paper-and-pencil test. Tailored tests raise several issues that cloud the concept, however. First, how can a passing point be set on the theta continuum? Second, how can a passing point be set on a reference test in the more familiar number-correct scale? Finally, can and should a passing point from a paper-and-pencil test be transferred to a computerized test?

There are several ways to set a passing point relative to the theta continuum. If a reference test exists, it will have a test characteristic curve. The test characteristic curve for a fixed test is simply the sum of the item characteristic curves, or

\[
T(\theta) = \sum_g P(u_g = 1|\theta)
\]  

[8]

If the passing point has already been set in terms of the number correct, the passing point in theta can be determined by solving Equation 8 for theta. Graphically this can be accomplished by identifying the theta value that corresponds to the number correct at which the passing point is set.

There are two somewhat more elegant ways to set the passing point using raw judgmental data, both derived from techniques of maximum-likelihood scoring. First, if the Angoff (1971) procedure is altered and judges are asked to evaluate whether the minimally competent candidate would most likely answer the question correctly or incorrectly, the standard likelihood equation shown as Equation 9 (after Hambleton & Swaminathan, 1985, p. 84) could be used to estimate theta from the dichotomous judgments.

\[
\sum_g a_g \frac{(u_g - P_g(\theta))(P_g(\theta) - c_g)}{P_g(\theta)(1 - c_g)} = 0.
\]  

[9]

If data are available from a passing-point study done using the classic Angoff method, the passing point can be obtained by solving Equation 10 for theta. Note that Equations 9 and 10 are identical, except \( R_{gj} \) (judge j’s probabilistic rating) has been substituted in Equation 10 for the scored item response. This rating, as is typically true of the Angoff technique, is for the proportion that will answer correctly, not the proportion who would know the correct answer. The passing point would be set at the average theta of the judges, whether set using Equation 9 or 10.

\[
\sum_g a_g \frac{(R_{gj} - P_g(\theta))(P_g(\theta) - c_g)}{P_g(\theta)(1 - c_g)} = 0.
\]  

[10]

The method of Equation 10 is similar to that proposed by Kane (1987), except for one significant difference. Kane had suggested averaging the item judgments across judges and then applying a formula comparable to Equation 10. Kane’s approach assumes that all judges agree on a common theta and differences are due to errors in judging the proportion. It seems more reasonable to believe that the judges would have different opinions regarding the ability level of the minimally
competent candidate and differences in ratings would result both from errors in judgment and differences of opinion. The average opinion (different from the opinion level underlying the average rating) seems a more reasonable value to select for a passing point. In addition to greater consistency with the conceptual model of IRT, it also allows the use of different sets of items across judges.

A passing point set on the theta scale can easily be transferred back to the number-correct scale of any reference test for which item parameters, relative to the theta scale, have been estimated. The passing point on the reference test is simply the test characteristic curve (Equation 8) evaluated at the passing point on theta. Regarding a choice of reference test, this can be a conventional test previously used, a theoretical reference test based on hypothetical item parameters, or a test composed of all the items in the item bank; IRT affords great flexibility in how the score may be expressed.

SUMMARY

Computerized testing offers significant advantages over paper-and-pencil testing. Although the psychometric advantages for licensure and certification testing appear to be small, the scheduling advantages that typically occur with computerization are great, especially when the time saved has value. Furthermore, issues of anxiety, comparability, and defensibility of the computerized mode do not appear to be significant. The difficulty in computerizing a test appears in finding a service network that can deliver the tests in a timely manner and in geographically appropriate locations. The small number of testing networks currently available renders the feasibility of implementing any small program questionable at this time. This difficulty will pass as more networks become available and more computerized tests make the operation of these networks more efficient and affordable.

REFERENCES


Thomas, T. J., & Green, B. F. (1989) Item presentation controls for computerized adaptive testing: Content balancing vs. mini-CAT (Report 89-1). Baltimore: Johns Hopkins University, Psychometric Laboratory.


