On the consistency of individual classification using short scales

Emons, W.H.M.; Sijtsma, K.; Meijer, R.R.

Published in:
Psychological Methods

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Psychological Methods
2007, Vol. 12, No. 1, 105–120

1082-989X/07/$12.00 DOI: 10.1037/1082-989X.12.1.105

On the Consistency of Individual Classification Using Short Scales

Wilco H. M. Emons and Klaas Sijtsma
Tilburg University

Rob R. Meijer
University of Twente

Short tests containing at most 15 items are used in clinical and health psychology, medicine, and psychiatry for making decisions about patients. Because short tests have large measurement error, the authors ask whether they are reliable enough for classifying patients into a treatment and a nontreatment group. For a given certainty level, proportions of correct classifications were computed for varying test length, cut-scores, item scoring, and choices of item parameters. Short tests were found to classify at most 50% of a group consistently. Results were much better for tests containing 20 or 40 items. Small differences were found between dichotomous and polytomous (5 ordered scores) items. It is recommended that short tests for high-stakes decision making be used in combination with other information so as to increase reliability and classification consistency.

Keywords: classification consistency, decision-making on short scales, individual decision making, reliability of short scales

Long cognitive tests and personality inventories can be stressful to children and adults suffering from, for example, concentration and attention problems, chronic physical fatigue, or brain damage due to hereditary defects or traumatic events (Donders, 2001; Goring, Baldwin, Marriot, Pratt, & Roberts, 2004; Kosinski et al., 2003; Reise & Henson, 2003; Stuss, Meiran, Guzman, Lafleche, & Willmer, 1996). Thus, there is a need for short tests and inventories that alleviate the burden of testing in various domains, such as clinical child psychology, mental health care, and medicine. Also, short questionnaires may increase response rates to mailed questionnaires (Edwards, Roberts, Sandercock, & Frost, 2004) in, for example, opinion and marketing research.

An example of a short inventory is the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), which consists of 11 questions and requires only 5–10 min to administer. This inventory is aimed at evaluating the mental state of psychiatric patients and consists of vocal responses in the domains of orientation, memory, and attention. As the authors emphasize, the quantitative assessment of cognitive performance via lengthy tests is a problem for elderly patients suffering from, for example, dementia syndromes because they are able to cooperate only for short periods. Other examples include an 8-item questionnaire that measures pathological dissociative experiences (Waller, Putnam, & Carlson, 1996), a 5-item version of the Test Anxiety Inventory (J. Taylor & Deane, 2002), and a 7-item questionnaire on alcohol drinking behaviors (Koppes, Twisk, Snel, van Mechelen, & Kemper, 2004); see Cooke, Michie, Hart, and Hare (1999) and Denollet (2005) for other examples.

Tests, including short ones, are often used in practice for classifying individuals, for example, into groups of those who will receive treatment and those who will not receive treatment. Treatment might refer to psychological or medical therapy but might also refer to counseling, a job, or a course. Classification problems can also involve three or more proficiency levels identified as nonoverlapping intervals on a continuous scale that are determined by standard setting procedures (e.g., Ercikan & Julian, 2002).

This study deals with the influence of random measurement error when observed test scores are used to classify individuals. In particular, the smaller the number of items in the test, the greater we expect the influence of measurement error to be on test scores and the decisions based on these test scores. The level of uncertainty caused by measurement error varies across individuals taking the test: Individuals closer to the cut-score are classified with less certainty than are respondents farther away. This suggests that an interval should be around the cut-score in which uncertainty may be unacceptably large for individual decision making in some classification problems. We hypothesize that for short scales this interval covers a large part of the scale, even if highly discriminating items that provide maximum information with respect to measurement in the vicinity of the cut-score have

Wilco H. M. Emons and Klaas Sijtsma, Department of Methodology and Statistics FSW, Tilburg University, Tilburg, the Netherlands; Rob R. Meijer, Department of Measurement and Data Analysis, University of Twente, Enschede, the Netherlands.

Correspondence concerning this article should be addressed to Wilco H. M. Emons, Department of Methodology and Statistics FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, the Netherlands. E-mail: w.h.m.emons@uvt.nl

been used. Support of this hypothesis by research results may provide grounds for careful and perhaps reserved use of short tests when decisions have far-reaching consequences.

Goals of This Study

Before we discuss the goals of this study, we introduce two important proportions. The first proportion, denoted π, is called the *certainty level.* Proportion π is chosen by the researcher to reflect the importance of correct decisions for the classification problem at hand: It defines the lower bound of the proportion of hypothetical independent repetitions of the test (Lord & Novick, 1968; to be defined shortly) in which an individual is classified correctly. For example, if $\pi = .9$ the researcher requires at least 90% of the hypothetical independent repetitions of the test to lead to the correct classification, and a lower value of π obviously expresses that a lower certainty level is deemed acceptable. The second proportion is called the *classification consistency* (*CC*). The value of *CC* varies for different values of π. For example, for $\pi = .9$ the *CC* equals the proportion of individuals from a given diagnostic group for whom the classification decision is correct in at least 90% of hypothetical independent repetitions of the test. Suppose that for $\pi = .9$ we find that $CC = .64$; this means that 64% of the individuals in the group are classified correctly in at least 90% of the hypothetical test administrations. It also means that for 36% of the individuals, the test score contains too much random measurement error to classify them correctly with a lower bound given by π. These people are located closer to the cut-score than are the other 64% (e.g., Hambleton & Slater, 1997; Subkoviak, 1976).

The first goal of this study is to establish, for given certainly level π, the influence of test length and other test and item characteristics on the *CC* in a particular diagnostic category. The second goal of the study is to determine the bounds of the interval around the cut-score in which the individuals for whom the test score contains too much measurement error, in our example 36% of the group, are located. This interval is called the *unreliability interval.* Like the *CC,* the unreliability interval is studied in relation to test length given realistic test and item characteristics. It will become clear that the bounds of the unreliability interval are needed for computing the *CC*; thus, the bounds and the *CC* are related, and predictions for one have implications for the other. Because the classification problem is a problem of random measurement error, we predict that the *CC* is smaller, and the unreliability interval is longer as test length decreases, holding constant all other properties of the test, the population, and the cut-score.

This article is organized as follows. First, we give a general definition of classification consistency. Second, we discuss some psychometric prerequisites that are needed in this study. Third, we discuss how the unreliability intervals and the *CC* are found, given a fixed certainty level π. Fourth, we discuss the design of a computational study in which, for a given distribution of test scores and a given certainty level, the test length, the cut-score, and the psychometric properties of the test and its constituent items are varied. Each of the design factors is expected to influence the unreliability intervals and the *CC.* Fifth, we present the results of a computational study. Finally, we discuss the results and provide directions for future research.

CC and Related Topics

CC

Lord and Novick (1968, p. 30) define for each indivdual taking a particular test a distribution of observable test scores with a mean that is equal to the true score. An individual's test score resulting from one administration of the test can be conceived of as a random draw from his or her distribution of test scores conditional upon his or her true score. This distribution is known as the *propensity distribution.* Now, suppose that, hypothetically, the same test is administered infinitely many times to the same individual and that these repetitions are independent (Lord & Novick, 1968, pp. 29–30). Also suppose that we know the individual's true score and, on the basis of the comparison of the true score and the cut-score of the classification problem, the individual's correct classification. Then we can determine the percentage of observable test scores from the propensity distribution that would classify the individual correctly. This percentage can be computed exactly in a computational study with the known properties of the test, the individual's true score, and a known cut-score. Given a desirable certainty level π, within a particular diagnostic category we select the individuals for whom the proportion of observable test scores from their propensity distributions that classify them correctly exceeds π; this selection determines the *CC* for that category. Because the spread within the propensity distributions is caused only by random measurement error, classification is more often correct for people whose true scores are far away from the cut-score (Hambleton & Slater, 1997).

For the sake of simplicity, we only consider classification into two disjoint, mutually exhaustive categories that are separated by a cut-score. On the basis of his or her true score, each individual belongs to one of these categories, and it is known which category this is. A test with known psychometric properties is administered infinitely many times to each individual. The only source of variation in an individual's test scores is random measurement error. We set the certainty level equal to, for example, $\pi = .9$, and we compute the corresponding *CC* for the group of people who belong to this category. The bounds of the unreliability intervals are also determined.

Additional Remarks

A high certainty level such as .9 represents a situation in which a decision is considered highly important. For example, the treatment might be expensive or it might involve a risk of some mental or physical damage for those who do not need it. Thus, one has to be certain to a high degree that individuals are assigned correctly to (non) treatment—that is, π must be high—one wants the number of these individuals to cover a large part of the group—that is, the *CC* also must be high. Other classification problems might involve other certainty levels and more than two disjoint and mutually exhaustive categories (e.g., Ercikan & Julian, 2002). A greater number of categories would involve developments similar to those outlined in this study for the simple case of two categories.

CC was originally defined (Ercikan & Julian, 2002; also, see Bechger, Maris, Verstralen, & Béguin, 2003; Huynh, 1976; Livingston & Lewis, 1995) as the percentage of people assigned to the same diagnostic category by two hypothetical independent repetitions of the same test. Notice that two draws from the propensity distribution provide less accurate information about *CC* than infinitely many draws, thus evaluating the whole propensity distribution.

CC is different from classification accuracy (e.g., Ercikan & Julian, 2002; also, Hambleton & Slater, 1997; Livingston & Lewis, 1995; Swaminathan, Hambleton, & Algina, 1974; Traub & Rowley, 1980). *Classification accuracy* is the degree to which, for a certain cut-score, a *single* test administration leads to the same classifications when either the true ability score or the estimated ability score is used. Ercikan and Julian (2002) express classification accuracy as the proportion of agreement across categories. Unlike *CC*, classification accuracy evaluates classification effects of a single test administration, and each individual is assumed to be classified equally reliably.

Psychometric Prerequisites

Let the test contain J items, and let items be indexed by j and k, with $j, k = 1, \ldots, J$. Let random variable X_j denote the score on item j and x_j denote the realization of this score; for example, $x_j = 0, 1$ for incorrect or correct solutions of items from cognitive tests, or $x_j = 0, \ldots, m$ for ordered levels of agreement on rating scales in personality inventories or other questionnaires. Let respondents be indexed by v and the sample size be denoted N so that $v = 1, \ldots, N$.

Given a fixed certainty level π, the unreliability interval and the *CC* were determined in a computational study that used *item response theory* (IRT) models. IRT models are ideal probabilistic test models for manipulating the test situation in a computational study (Embretson & Reise, 2000; Van der Linden & Hambleton, 1997). IRT models also enable the evaluation of the contribution of each individual item to the measurement precision of the test by means of Fisher's information function (e.g., Baker & Kim, 2004; Van der Linden, 2005; also see Reise & Henson, 2000).

IRT models define the relationship between the probability of obtaining a particular score on an item and the latent trait that is assumed to drive responses to the items in the test. We define the probability of obtaining a score x_j as a function of latent trait θ as $P(X_j = x_j|\theta)$. For binary item scores, this is the *item response function* (IRF), also denoted as $P_j(\theta \equiv P(X_j = 1|\theta)$, and for polytomous item scores this is the *category response function* (CRF), also denoted as $P_{jx_j}(\theta) \equiv P(X_j = x_j|\theta)$, for $x_j = 0, \ldots, m$.

Unreliability intervals and the *CC* were studied by using tests consisting entirely of binary scored items and tests consisting entirely of polytomously scored items. For binary items, we used the Rasch (1960) model or the *one-parameter logistic model* (1PLM). Let b_j be the parameter that locates the IRF on the θ scale such that $P_j(\theta) = .5$; hence, b_j is the location or difficulty parameter. The IRF of the 1PLM is defined as

$$P_j(\theta) = \frac{\exp(\theta - b_j)}{1 + \exp(\theta - b_j)}. \quad (1)$$

For ordered polytomous item scores, we used the *graded response model* (GRM; Samejima, 1997). For each item score, $x_j = 1, \ldots, m$, a response function is defined. This response function has location or threshold parameters b_{jx_j} ($x_j = 1, \ldots, m$) and a slope parameter a_j, which depends on j only, such that

$$P(X_j \geq x_j|\theta) = \frac{\exp[a_j(\theta - b_{jx_j})]}{1 + \exp[a_j(\theta - b_{jx_j})]}. \quad (2)$$

Note that $P(X_j \geq 0|\theta) = 1$ by definition. This response function, which is also known as the *item step response function* (ISRF), is related to the CRF by means of

$$P_{jx_j}(\theta) = P(X_j \geq x_j|\theta) - P(X_j \geq x_j + 1|\theta). \quad (3)$$

Notice that if $a_j = a$ for all J items, the ISRFs reduce to functions that are similar to those in the 1PLM (Equation 1), and if $a = 1$ they are equal. Fixing a in both models is a convenient way to make dichotomous-item tests and polytomous-item tests comparable when different choices of a represent different levels of discrimination.

The contribution of the J item scores X_j to the maximum-likelihood estimation of latent trait θ (the result of which is the maximum-likelihood estimate $\hat{\theta}$) is given by Fisher's information function. Let $I(\theta)$ denote the information function for the whole test and let $I_j(\theta)$ denote the information function for item j. Then, the contribution of the J items to the maximum-likelihood estimation of θ is the sum of the item contributions (Baker & Kim, 2004, chap. 3),

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta), \tag{4}$$

and the standard error of the asymptotic normal $\hat{\theta}|\theta$ is given by

$$SE(\hat{\theta}|\theta) = I(\theta)^{-1/2}. \tag{5}$$

The information function and the standard error can be used to assemble tests such that they measure the most reliably at the cut-score, denoted θ_c, that is used to separate the treatment and the nontreatment groups. For the 1PLM, the smallest standard error at θ_c is obtained for items with $b_j = \theta_c$ (Figure 1A; see also Baker & Kim, 2004, p. 73). For the GRM, $I_j(\theta)$ can have several peaks. For classification, it often suffices to choose items for which θ_c lies somewhere in between the m location parameters, provided $I_j(\theta)$ has a near constant and relatively high value in that region (Figure 1B; see also Baker & Kim, 2004, pp. 220–223).

Finally, the "classical" test score or total score on J items is defined as random variable X_+, such that

$$X_+ = \sum_{j=1}^{J} X_j. \tag{6}$$

Because both the $\hat{\theta}$ scale and the X_+ scale are used in practice for decision making, we point out the monotone relationship between both scales. Let T_v be the expected (i.e., true score) value of X_{+v}, as defined in classical test theory (Lord & Novick, 1968, p. 30). For binary items with monotone nondecreasing IRFs, θ_v and T_v are monotone related as

$$T_v = \sum_{j=1}^{J} P_j(\theta_v) \tag{7}$$

(Lord, 1980, p. 46) and for polytomously scored items with monotone nondecreasing ISRFs as

$$T_v = \sum_{j=1}^{J}\sum_{x=1}^{m} xP_{jx_j}(\theta_v) = \sum_{j=1}^{J}\sum_{x=1}^{m} P(X_j \geq x_j|\theta_v) \tag{8}$$

(e.g., Sijtsma & Hemker, 2000). Because of these monotone relationships, we may switch from one scale to the other. This proves to be convenient in this study.

Classification Into Two Categories

We study the following situation. We choose a cut-score θ_c and assume that people with $\theta < \theta_c$ do not need treatment and that people with $\theta \geq \theta_c$ do need treatment. Because θ and the true score T are monotonically related, classification on the basis of T and a cut-score T_c that corresponds to θ_c is identical to classification on the basis of θ and θ_c. In practice, one has $\hat{\theta}$ or X_+ but not θ or T, respectively. We use a distribution for θ, a cut-score θ_c, and the 1PLM and the GRM to simulate a testing and classification problem, and we use X_+ and T_c for the actual classification. This enables us to study the exact influence of random measure-

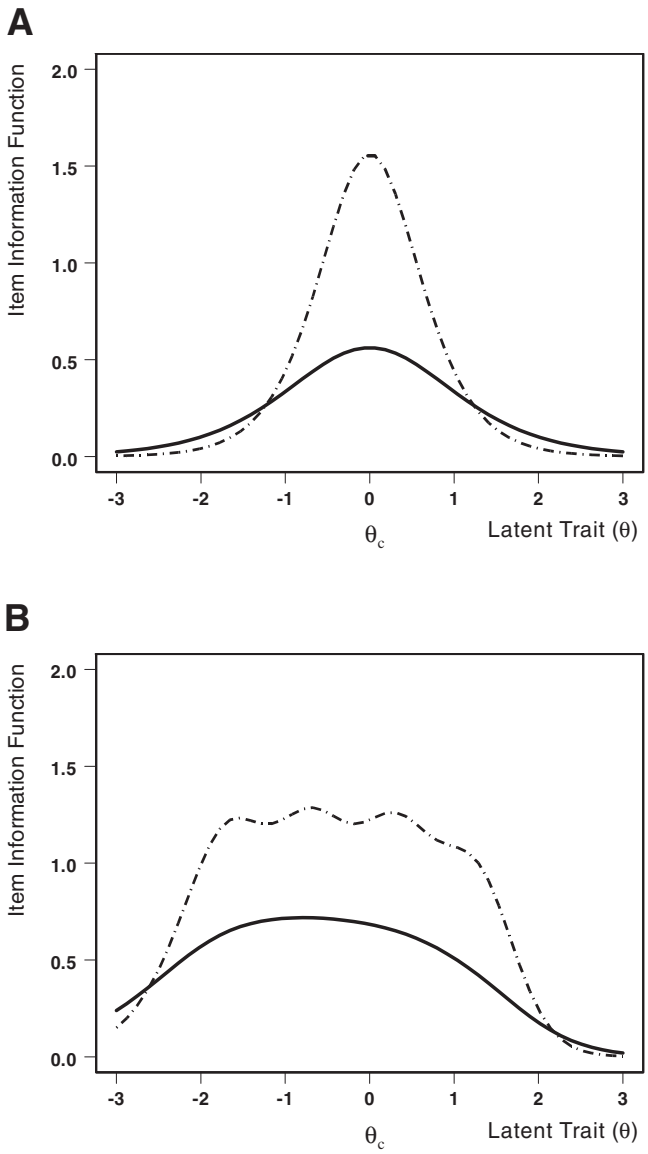

Figure 1. Information curves for (A) dichotomous item j, with $b_j = 0$ for the one-parameter logistic model, at $\theta_c = 0$, for low discrimination power ($a_j = 1.5$; solid curve) and high discrimination power ($a_j = 2.5$; dashed-dotted curve); and for (B) polytomous ($m + 1 = 5$) item k, with $b_{k1} = -1.5$, $b_{k2} = -0.5$, $b_{k3} = 0.5$ and $b_{k4} = 1.5$ (i.e., $\bar{b}_k = 0$) for the graded response model, again at $\theta_c = 0$, for low discrimination power ($a_k = 1.5$; solid curve) and high discrimination power ($a_k = 2.5$; dashed curve).

ment error in X_+ on the unreliability interval and the *CC* given a preset certainty level π.

The closer θ is to θ_c, the more the conditional distributions of $X_+|\theta$ and $X_+|\theta_c$ overlap, and the more classification on the basis of the fallible X_+ score resembles flipping an unbiased coin. Thus, only for θs that are far enough from θ_c in either direction will classification on the basis of X_+ exceed certainty level π. On the basis of this line of reasoning, we identify a lower bound, $\theta_l < \theta_c$, below which the probability of being classified correctly as not needing treatment on the basis of X_+ exceeds a preset value π; and, similarly, an upper bound, $\theta_u > \theta_c$, above which the probability of being classified correctly as needing treatment on the basis of X_+ exceeds that same value π. Interval (θ_l, θ_u) is the unreliability interval. The higher the value of π, the further the bounds are driven away from θ_c in either direction, and the longer the unreliability interval becomes.

The bounds θ_l and θ_u are formalized as follows. Given the choice of π, and given the cut-score θ_c, the psychometric properties of the test and the items, and the distribution of θ in the group under consideration, we determine lower bound θ_l ($\theta_l < \theta_c$), such that

$$P(X_+ < T_c|\theta < \theta_l) \geq \pi; \tag{9}$$

and, similarly, upper bound θ_u ($\theta_u \geq \theta_c$), such that

$$P(X_+ \geq T_c|\theta \geq \theta_u) \geq \pi. \tag{10}$$

Figure 2 graphically shows how the bounds θ_l and θ_u are determined for a hypothetical test of $J = 10$ binary items (technical details are given later and in the Appendix). Figure 2 shows the test response function, defined as $E(X_+|\theta)$ (Lord, 1980, p. 49). We use either Equation 7 or Equation 8 to determine the value of T_c that corresponds to θ_c. For decreasing values of θ ($\theta < \theta_c$), we determine for each θ the distribution of $X_+|\theta$. As θ decreases further, the distribution $X_+|\theta$ shifts further down along the X_+-axis (see Figure 2), whereas its spread becomes smaller as it approaches the bounds of X_+; that is, for smaller θ the distribution of $X_+|\theta$ has both smaller mean and variance. For decreasing θ, we continue determining distributions $X_+|\theta$ until a proportion π of the X_+ values fall below T_c. The value of θ at which this happens is the lower bound θ_l. Only for individuals whose θ values are smaller than θ_l do we know that in at least a proportion π of the repetitions are they assigned to nontreatment. The procedure for finding upper bound θ_u is similar.

Given the availability of bounds θ_l and θ_u, *CC* is operationalized as follows. For notational convenience, we use set notation $\bar{D}$ if $\theta \in \{\theta < \theta_c\}$ and D if $\theta \in \{\theta \geq \theta_c\}$. Consistent (C) classification can either refer to category $\bar{D}$, denoted as $C\bar{D}$, or to category D, denoted as CD. For a given π and corresponding unreliability interval (θ_l, θ_u), we determine proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$; both represent

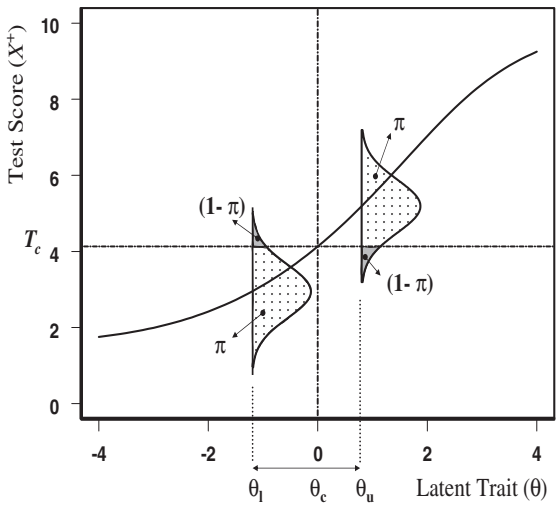

Figure 2. Distributions of test score X_+ conditional on θ, determined such that level of classification consistency π identifies θ_l (left-hand distribution) and θ_u (right-hand distribution) given cut-score θ_c. The S-shaped curve is the test response function.

levels of *CC* but for different diagnostic categories. Given a distribution for θ, these proportions are equal to

$$P_\pi(C\bar{D}) = \frac{P(\theta < \theta_l)}{P(\theta < \theta_c)}, \quad \text{and } P_\pi(CD) = \frac{P(\theta \geq \theta_u)}{P(\theta \geq \theta_c)}. \tag{11}$$

The values of θ_l and θ_u for which Equation 9 and Equation 10 hold were obtained by using an iterative algorithm based on interval bisection; details can be found in the Appendix. Each iteration requires the distribution of $X_+|\theta$, which was obtained as follows. For dichotomous items with varying location parameters b_j, the distribution of $X_+|\theta$, denoted $\phi(X_+|\theta)$, is the generalized binomial (Kendall & Stuart, 1969, p. 127; Lord, 1980, p. 45). The generalized binomial cannot be expressed in closed form and, therefore, a recursion formula (Lord & Wingersky, 1984; see also Kolen & Brennan, 1995, pp. 182–183) was used to generate this distribution. For polytomous item scores with varying threshold parameters b_{jx_j}, the distribution $\phi(X_+|\theta)$ is a generalized multinomial (e.g., Kolen & Brennan, 1995, pp. 219). The generalized multinomial distribution cannot be expressed in closed form either, and a recursive algorithm was used to generate this distribution (Kolen & Brennan, 1995, pp. 219–221; Thissen, Pommerich, Billeaud, & Williams, 1995). More specifically, the recursion formula first evaluates $\phi(X_+|\theta)$ for the first two items, which contains the probabilities of X_+ given θ for $X_+ = 0, 1, \ldots, 2m$. In each of the $J - 2$ consecutive steps s ($s = 1, \ldots, J - 2$), the distribution of $X_+|\theta$ is expanded to the distribution $\phi(X_+|\theta)$ for $s + 2$ items. For dichotomous items, this recursion formula specializes to the recursion formula of Lord and Wingersky (1984). More details can be found in the Appendix.

Research Questions

The goal of this study can now be formulated in terms of research questions that can be investigated in a computational study. For different certainty levels π and a standard normal distribution of latent trait θ, we determine the influence of (a) test length (J), (b) cut-score (θ_c), (c) item score (dichotomous or polytomous), (d) item discrimination (parameter a_j), and (e) item difficulty (parameter b_j), on the CC proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ and the bounds of the unreliability interval (θ_l, θ_u).

Method

Analysis Steps

The computations that lead to the bounds θ_l and θ_u and the proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ follow the next sequence of steps:

1. We choose a cut-score θ_c that defines a particular area under the right-hand tail of the standard normal distribution for θ, denoted $f(\theta)$.
2. Given θ_c, we obtain the corresponding cut-score, T_c, by using either Equation 7 or Equation 8.
3. We choose the certainty level, π. This choice determines the length of the (θ_l, θ_u) unreliability interval.
4. We determine interval (θ_l, θ_u) by using the algorithm explained in the Appendix.
5. We compute the proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ by using areas under the standard normal given values of θ_c, θ_l, and θ_u.

An interval (T_l, T_u) corresponding to (θ_l, θ_u) may be obtained by using Equation 7 or Equation 8, but such an interval will prove to be problematic, as we explain later. Thus, we only report a few noteworthy results for true score intervals.

The analysis steps were repeated for several combinations of (a) test length, (b) cut-score, (c) item score (dichotomous or polytomous; implied by the chosen IRT model), (d) item discrimination power, and (e) location and spread of item difficulties. The design characteristics and their expected influence on the proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ and on the unreliability interval (θ_l, θ_u), are discussed next.

Independent Variables

First we enumerate the expected influence of each of the independent variables on the proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ and on the unreliability interval (θ_l, θ_u). Second, we describe the specific choices made for each of the independent variables. We had the following expectations about effects:

1. *Test length:* Longer tests are expected to yield greater proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ and shorter intervals (θ_l, θ_u) because the influence of random measurement error variance relative to true score variance is smaller.
2. *Cut-score:* Let group D be a minority of the population and let its members have the highest θs. It is expected that a more extreme cut-score—equivalent to a smaller group D size—yields a greater proportion $P_\pi(C\bar{D})$ and a smaller proportion $P_\pi(CD)$, a result that is well-known from personnel selection problems (Wiggins, 1973; H.C. Taylor & Russell, 1939). Unreliability intervals are expected to be shorter as the cut-score is more extreme.
3. *Item scores:* It is expected that J polytomous items will yield greater proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ than J dichotomous items because the variance of the corresponding X_+ scores is greater for polytomous items and this is expected to reduce the influence of random measurement error variance relative to true score variance. As a result, the unreliability intervals are expected to be shorter for polytomous-item tests than for dichotomous-item tests.
4. *Discrimination values:* Test information increases as item discrimination increases. Thus, it is expected that proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ will increase and that unreliability intervals will be shorter as item discrimination increases.
5. *Location of items and spread of item difficulties:* For the 1PLM, the closer an item's location parameter is to θ_c, the greater this item's contribution is to the test information function (Equation 4) and, equivalently, to the reduction of the standard error of the maximum-likelihood estimate $\hat{\theta}$ (Equation 5). The shape of the test information function is determined by the locations of the J items. The next three predictions about the influence of the item difficulties on the proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ and the unreliability interval (θ_l, θ_u) can be made safely.

   If $b_j = \theta_c$ $(j = 1, \ldots, J)$, test information is maximal at θ_c. As a result, the interval (θ_l, θ_u) has minimal length and the proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ are maximal.

   If individual b_js are different and their mean (de-

noted $\bar{b}$) equals θ_c (i.e., $\bar{b} = \theta_c$), the proportions are smaller and the intervals longer the more the *b*s differ.

If $b_j = \theta_0$ ($b_j = 1, \ldots, J$), the greater the absolute distance between θ_0 and θ_c the smaller the proportions and the longer the intervals.

In other cases, the interplay of the mean and the spread of the item difficulties produces a test information function for which the influence on the proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ and the unreliability intervals is difficult to predict. For the GRM, predictions similar to those for the 1PLM are more difficult because each item has *m* location parameters, and the relationship between item location and maximum information is not as straightforward as in the 1PLM.

Specific choices of values of independent variables:

1. *Test length:* Test length was $J = 6$, 8, 10, 12, 20, and 40. We consider the first four values typical of short tests, $J = 20$ typical of medium-length tests, and $J = 40$ typical of long tests.

2. *Cut-score:* Given a standard normal density $f(\theta)$, different sizes of Group *D* correspond with 50%, 25%, 10%, and 5% of the right-hand tail of $f(\theta)$. The corresponding cut-scores are $\theta_c = 0$, 0.675, 1.285, and 1.645, respectively. The cut-score is meaningful given that we know to what percentage of the right-hand tail of $f(\theta)$ it refers. Thus, in discussing results it is sometimes more convenient to talk about this percentage (denoted as PERC) instead of the cut-score.

3. *Item scores:* Binary item scores were modeled using the 1PLM (Equation 1) and polytomous item scores were modeled with the GRM (Equation 2). Each of the *J* polytomous items had five ordered-answer categories ($m = 4$), meaning that four ISRFs are defined as each having a difficulty parameter (b_{jx_j}, $x_j = 1, \ldots, 4$). Tests consisted of *J* dichotomous items or *J* polytomous items.

4. *Discrimination power:* We used simulations to determine realistic values for the discrimination parameters, such that for the shortest tests (i.e., $J = 6$) the *a*s would produce values of Cronbach's (1951) alpha approximately between .60 and .80. These are values typically reported for short tests (e.g., Goring et al., 2004; Knight, Goodman, Pulerwitz, & DuRant, 2000; Murphy & Davidshofer, 1998, p. 142). On the basis of these simulations, both the 1PLM and the GRM items were found to have relatively low discrimination power when $a_j = 1.5$ (alpha was approximately .60) and relatively high discrimination power when $a_j = 2.5$ (alpha was approximately .80). For all *J* items within the same test, the *a*s were chosen to be equal.

5. *Location of items and spread of item difficulties:* For the 1PLM, the mean item difficulty, $\bar{b}$, was either equal (Figures 3A and 3C) or unequal (Figure 3B) to the cut-

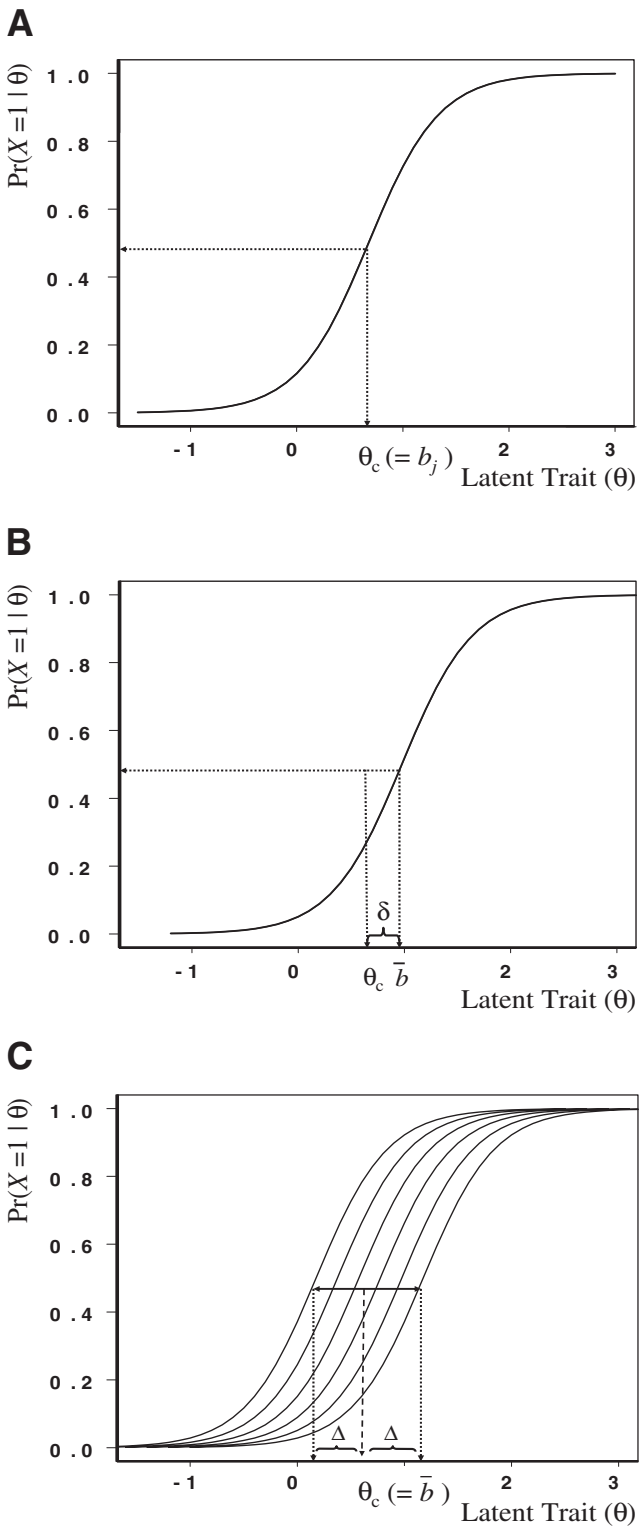

Figure 3. Item response functions for six one-parameter logistic model items with $a_j = 2.5$, $j = 1 \ldots, 6$: (A) all six items located at θ_c ($b_j = \theta_c = 0.675$, $j = 1, \ldots, 6$); (B) all six items located at $\theta_c + \delta$, with $\theta_c = 0.675$ and $\delta = 0.3$ ($b_j = \theta_c + \delta = 0.975$, $j = 1, \ldots, 6$); (C) all six items evenly spread around $\theta_c = 0.675$, within range $(\theta_c - \Delta; \theta_c + \Delta)$, and with mean $\bar{b} = \theta_c$ and $\Delta = 0.5$.

score, θ_c. This was formalized as $\bar{b} = \theta_c + \delta$, with $\delta = -.30, -.15, 0, .15, .30$. Notice that δ gives the distance of the mean $\bar{b}$ to the cut-score θ_c; thus, it quantifies how much the items are "off target" on average. For example, $\delta = 0$ means that $\bar{b} = \theta_c$; so the items are centered at the cut-score. Also, the J item difficulties within one test were either equal (zero spread; see Figures 3A and 3B) or unequal (positive spread; see Figure 3C). Item difficulties varied in equidistant steps from $\theta_c - \Delta$ to $\theta_c + \Delta$, for Δ fixed at either 0, .50, or 1. For $\Delta = 0$, zero spread was obtained.

To keep the study within manageable proportions, only main effects of the item locations (δ) and the spread of the item difficulties (Δ) were studied. In particular, for each combination of the other design factors—test length, cut-score, item score (IRT model), and item discrimination power—results were obtained for the following: $\delta = 0$ and $\Delta = 0, .50, 1$; and for $\delta = -.30, -.15, .15, .30$ and $\Delta = 0$.

Given that predictions about the influence of item locations on proportions and intervals are not straightforward for polytomous items, we make use of the knowledge that item j is more informative about the maximum likelihood estimate $\hat{\theta}$ as cut-score θ_c is more in the middle of the m location parameters b_{jx_j}, $x_j = 1, \ldots, m$, with $m = 4$. Thus, for item j ($j = 1, \ldots, J$) the four difficulty parameters b_{jx_j}. Across the J polytomous items, we defined $\bar{b} = \theta_c + \delta$, with $\delta = -.30, -.15, 0, .15, .30$, similar to the definition for the 1PLM. Similar to 1PLM items, for GRM items the mean item step difficulties, $\bar{b}_j$, were equidistant from $\theta_c - \Delta$ to $\theta_c + \Delta$, or Δ fixed at 0, .50, or 1. The choices of δ and Δ were similar to those for the 1PLM.

We chose certainty level $\pi = .9$, which expresses that highly consistent decisions are considered important. We also discuss some results for $\pi = .8, .7$ and .6, keeping other design characteristics fixed. The design of the study is summarized in Table 1.

Dependent Variables

The dependent variables were the proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ (Equation 11) and the bounds of the unreliability interval, θ_l and θ_u.

Results

The results are manyfold and show much detail, but we concentrate on the main results with respect to the influence of the design factors on the proportions $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$. First, results are discussed for $\pi = .9$ and all items located at the cut-score. Main effects for test length, cut-score, item score, and item discrimination are discussed, followed by some interesting detailed results. Then, some results are discussed for smaller values of π and for $\pi = .9$ and items that show variation in item locations. Finally, we discuss some results for the (θ_l, θ_u) unreliability intervals, translate them to T_l, T_u intervals, and discuss the problems encountered.

Results for $\pi = .9$: All Items Located at Cut-Score θ_c

For $\pi = .9$ and all items located at the cut-score θ_c, Tables 2 and 3 give the (θ_l, θ_u) intervals and the CC proportions, $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$, for varying test length, cut-score (expressed as percentage PERC of the area under the standard normal θ distribution for the treatment group D), and item discrimination. Table 2 gives results for dichotomous items generated by means of the 1PLM, and Table 3 gives corresponding results for polytomous items generated by means of the GRM.

Test length. Longer tests were predicted to produce greater CC proportions, $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$. This result was found consistently as shown in each panel in Tables 2 and 3.

Table 1
Factors and Factor Levels of the Computational Study

Factor description	Symbol	Levels/values
Fully crossed factors		
Cut-score (percentage of individuals in diagnostic category)	θ_c, PERC	0 (50%), 0.675 (25%), 1.285 (10%), 1.645 (5%)
Test length	J	6, 8, 10, 12, 20, 40
IRT model (dichotomous vs. polytomous items)		1PLM, GRM
Item discrimination power	a_j	1.5, 2.5
Fully crossed factors combinations: Item parameters		
Distance between mean difficulty and θ_c with no spread of item difficulties	δ	$-.30, -.15, 0, .15, .30$
Spread of difficulties with mean difficulty equal to θ_c	Δ	0.00, 0.50, 1.00

Note. The latent trait θ has a standard normal distribution. PERC = percentage of individuals in diagnostic category; IRT = item response theory; 1PLM = one-parameter logistic model; GRM = graded response model.

Table 2
Intervals (θ_l, θ_u) and Proportions of Consistent Classification $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$ for Dichotomous-Item Tests (1PLM), Different Test Lengths, Discrimination Levels, and PERCs

	Low discrimination				High discrimination			
J	θ_l	θ_u	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$	θ_l	θ_u	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$
PERC = 50 ($\theta_c = 0$)								
6	−0.74	0.74	.46	.46	−0.45	0.45	.66	.66
8	−0.64	0.64	.53	.53	−0.38	0.38	.70	.70
10	−0.56	0.56	.58	.58	−0.34	0.34	.74	.74
12	−0.51	0.51	.61	.61	−0.31	0.31	.76	.76
20	−0.39	0.39	.70	.70	−0.23	0.23	.82	.82
40	−0.27	0.27	.79	.79	−0.17	0.17	.87	.87
PERC = 25 ($\theta_c = 0.675$)								
6	−0.07	1.42	.63	.31	0.23	1.12	.79	.52
8	0.04	1.31	.69	.38	0.29	1.06	.82	.58
10	0.11	1.24	.73	.43	0.34	1.01	.84	.62
12	0.17	1.18	.75	.47	0.37	0.98	.86	.65
20	0.29	1.06	.81	.58	0.44	0.91	.89	.73
40	0.40	0.95	.88	.69	0.51	0.84	.93	.81
PERC = 10 ($\theta_c = 1.285$)								
6	0.54	2.03	.78	.21	0.84	1.73	.89	.42
8	0.65	1.92	.82	.28	0.90	1.66	.91	.48
10	0.72	1.84	.85	.33	0.94	1.62	.92	.53
12	0.77	1.79	.87	.37	0.98	1.59	.93	.56
20	0.89	1.67	.90	.47	1.05	1.51	.95	.65
40	1.01	1.55	.94	.60	1.12	1.44	.97	.74
PERC = 5 ($\theta_c = 1.645$)								
6	0.90	2.39	.86	.17	1.20	2.09	.92	.33
8	1.01	2.28	.89	.23	1.26	2.03	.94	.40
10	1.08	2.21	.91	.27	1.31	1.98	.95	.44
12	1.14	2.16	.92	.31	1.34	1.95	.95	.48
20	1.26	2.03	.94	.42	1.41	1.88	.97	.58
40	1.37	1.92	.96	.55	1.48	1.81	.98	.71

Note. All item locations at θ_c. 1PLM = one-parameter logistic model; PERC = percentage of individuals in diagnostic category.

Cut-score. A more extreme cut-score—equivalent to a smaller PERC—was predicted to yield a greater proportion $P_{.9}(C\bar{D})$ and a smaller proportion $P_{.9}(CD)$. This result was found consistently for different test lengths and item discriminations; the reader may follow the four panels for different PERCs from top to bottom in Tables 2 (dichotomous items) and 3 (polytomous items).

Item scores. Polytomous items were predicted to yield greater proportions, $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$, than dichotomous items. Indeed this was found; one may compare corresponding entries in Tables 2 and 3. However, differences were small, often no more than a few hundredths. Thus, although the effect is in the predicted direction, it is not as pronounced as expected.

Item discrimination. Greater item discrimination was predicted to produce greater proportions, $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$, than was lower item discrimination. Tables 2 and 3 show that this prediction was supported by the results: In each table, one may compare the proportions in the left half with the corresponding proportions in the right half.

Some detailed results. We concentrate on classification in category *D*. For PERC = 50 (i.e., $\theta_c = 0$), for $J = 6$ dichotomous items and low item discrimination, $P_{.9}(CD) = .46$; this means that 46% of the persons who had $\theta \geq \theta_c$ were assigned to *D* by at least 90% of the test repetitions (Table 2). For smaller PERC values, $P_{.9}(CD)$ decreased considerably: .31 (PERC = 25), .21 (PERC = 10), and .17 (PERC = 5). Although one could be tempted to blame these low values on weak item discrimination, for high item discrimination corresponding proportions were indeed higher but were not impressive: $P_{.9}(CD) = .66$, .52, .42, and .33, as PERC values became smaller. Thus, for short tests

Table 3
Intervals (θ_l, θ_u) and Proportions of Consistent Classification $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$ for Polytomous-Item Tests (GRM), Different Test Lengths, Discrimination Levels, and PERCs

	Low discrimination				High discrimination			
J	θ_l	θ_u	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$	θ_l	θ_u	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$
PERC = 50 ($\theta_c = 0$)								
6	−0.64	0.64	.52	.52	−0.42	0.42	.67	.67
8	−0.55	0.55	.58	.58	−0.36	0.36	.72	.72
10	−0.49	0.49	.62	.62	−0.32	0.32	.75	.75
12	−0.45	0.45	.65	.65	−0.29	0.29	.77	.77
20	−0.35	0.35	.73	.73	−0.23	0.23	.82	.82
40	−0.24	0.24	.81	.81	−0.16	0.16	.87	.87
PERC = 25 ($\theta_c = 0.675$)								
6	0.04	1.31	.69	.38	0.26	1.09	.80	.55
8	0.12	1.23	.73	.44	0.32	1.03	.83	.60
10	0.18	1.17	.76	.49	0.36	0.99	.85	.64
12	0.22	1.13	.78	.52	0.38	0.97	.87	.67
20	0.33	1.02	.84	.62	0.45	0.90	.90	.74
40	0.43	0.92	.89	.72	0.52	0.83	.93	.81
PERC = 10 ($\theta_c = 1.285$)								
6	0.64	1.92	.82	.27	0.86	1.70	.90	.45
8	0.73	1.84	.85	.33	0.92	1.64	.91	.51
10	0.79	1.78	.87	.38	0.96	1.60	.93	.55
12	0.83	1.73	.89	.42	0.99	1.57	.93	.58
20	0.94	1.63	.92	.52	1.06	1.51	.95	.66
40	1.04	1.53	.95	.64	1.12	1.44	.97	.75
PERC = 5 ($\theta_c = 1.645$)								
6	1.01	2.28	.89	.22	1.23	2.06	.94	.39
8	1.09	2.20	.91	.28	1.29	2.00	.95	.45
10	1.15	2.14	.92	.32	1.33	1.96	.96	.50
12	1.19	2.10	.93	.36	1.35	1.94	.96	.53
20	1.30	1.99	.95	.47	1.42	1.87	.97	.62
40	1.40	1.89	.97	.59	1.49	1.80	.98	.71

Note. All item locations at θ_c. GRM = graded response model; PERC = percentage of individuals in diagnostic category.

($J = 6$), *CC* proportions in *D* were nearly always smaller than .50, a result which is due to random measurement error having a great impact on classification. It can be verified in the tables that results did not rapidly become better for $J = 8$, 10, and 12 and that polytomous scoring did not boost proportions relative to dichotomous scoring (cf. Tables 2 and 3).

For medium ($J = 20$) and long ($J = 40$) tests, proportion $P_{.9}(CD)$ was considerably larger than for smaller *J*. Often it was far over .50, and when items had high discrimination, approximately three quarters of the group with $\theta \geq \theta_c$ were classified in *D* by at least 90% of the test repetitions. For example, for PERC = 50 and dichotomous, highly discriminating items, we found that $P_{.9}(CD) = .82$ ($J = 20$) and $P_{.9}(CD) = .87$ ($J = 40$), and for PERC = 5 we found corresponding probabilities of .58 and .71.

Results for Smaller Values of π

Lowering π to .8, .7, and .6 (results not tabulated here) resulted in an increase of $P_\pi(CD)$ relative to $\pi = .9$, but for short tests and small PERCs these proportions remained small. For example, for $\pi = .8$ and dichotomous-item tests consisting of items with low discrimination, $P_{.8}(CD)$ was at most .50 for combinations of short tests and small PERC values. These results mean that for less than 50% of the respondents with $\theta \geq \theta_c$ classification in group *D* was correct for at least 80% of the test repetitions. For smaller $\pi = .6$, and .7, proportions $P_\pi(CD)$ were higher than .50.

For dichotomous item tests and high item discrimination, setting $\pi = .8$ resulted in $P_{\pi}(CD)$ greater than .50 in all conditions. In particular, $P_{.8}(CD)$ was greater than .77 for PERC = 50, and greater than .67 for PERC = 25. For short tests ($J \leq 12$) and PERC ≤ 10, however, it was necessary to lower π to .7 or .6 for obtaining $P_{.7}(CD)$ and $P_{.6}(CD)$ of at least .70.

Location of Items and Spread of Item Difficulties

Table 4 provides proportions $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$ for dichotomous-item tests, in which items have varying spread of item difficulties (Δ); and Table 5 provides similar results for polytomous-item tests. The three predictions about the influence of the item difficulties on the *CC* proportions were all confirmed, but the effects were small.

In particular, for dichotomous-item tests, proportions $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$ decreased little with increasing spread in item difficulty (Δ). One may compare results for equal item locations (i.e., $\Delta = 0$; in Table 2 with those for varying item locations [Table 4]). For low item discrimination, differences between the $P_{.9}(CD)$s for equal item locations (i.e., $\Delta = 0$) and varying item locations were small (varying from .00 to .06). For high item discrimination, differences ranged from .00 to .11. For polytomous-item tests, differences between items having the same locations and items having varying item locations showed minor differences (largest absolute difference equal to .01).

Results for different mean item locations are not tabulated. In general, different mean item difficulties produced small differences in the proportions $P_{.9}(C\bar{D})$s and $P_{.9}(CD)$.

Table 4
Proportions $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$ for Dichotomous-Item Tests (1PLM), Different Test Lengths, Discrimination Levels, PERCs, and Spread of Item Locations (Δ)

	Low discrimination				High discrimination			
	$\Delta = 0.50$		$\Delta = 1.00$		$\Delta = 0.50$		$\Delta = 1.00$	
J	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$
				PERC = 50 ($\theta_c = 0$)				
6	.44	.44	.40	.40	.62	.62	.55	.55
8	.51	.51	.48	.48	.68	.68	.62	.62
10	.56	.56	.53	.53	.72	.72	.67	.67
12	.60	.60	.57	.57	.74	.74	.70	.70
20	.69	.69	.67	.67	.80	.80	.77	.77
40	.78	.78	.77	.77	.86	.86	.84	.84
				PERC = 25 ($\theta_c = 0.675$)				
6	.62	.30	.58	.26	.76	.49	.71	.41
8	.68	.37	.65	.33	.81	.55	.76	.49
10	.72	.42	.69	.39	.83	.60	.80	.54
12	.75	.46	.73	.43	.85	.63	.82	.58
20	.81	.57	.80	.54	.89	.71	.87	.67
40	.87	.68	.86	.66	.92	.79	.91	.76
				PERC = 10 ($\theta_c = 1.285$)				
6	.77	.20	.74	.17	.87	.38	.84	.31
8	.82	.26	.80	.23	.90	.45	.87	.38
10	.84	.31	.83	.28	.91	.50	.89	.44
12	.86	.35	.85	.32	.92	.54	.91	.48
20	.90	.46	.89	.44	.94	.63	.93	.58
40	.94	.59	.93	.57	.96	.73	.96	.69
				PERC = 5 ($\theta_c = 1.645$)				
6	.85	.16	.83	.13	.92	.33	.90	.25
8	.88	.22	.87	.19	.94	.40	.92	.33
10	.90	.26	.89	.23	.95	.44	.94	.38
12	.91	.30	.91	.27	.95	.48	.94	.42
20	.94	.41	.94	.38	.97	.58	.96	.53
40	.96	.54	.96	.52	.98	.69	.97	.65

Note. 1PLM = one-parameter logistic model; PERC = percentage of individuals in diagnostic category.

Table 5
Proportions $P_{.9}(C\bar{D})$ and $P_{.9}(CD)$ for Polytomous-Item Tests (GRM), Different Test Lengths, Discrimination Levels, PERCs, and Spread of Item Locations (Δ)

	Low discrimination				High discrimination			
	$\Delta = 0.50$		$\Delta = 1.00$		$\Delta = 0.50$		$\Delta = 1.00$	
J	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$	$P_{.9}(C\bar{D})$	$P_{.9}(CD)$
				PERC = 50 ($\theta_c = 0$)				
6	.52	.52	.52	.52	.68	.68	.67	.67
8	.58	.58	.58	.58	.72	.72	.72	.72
10	.62	.62	.62	.62	.75	.75	.75	.75
12	.65	.65	.65	.65	.77	.77	.77	.77
20	.73	.73	.73	.73	.82	.82	.82	.82
40	.81	.81	.80	.80	.87	.87	.87	.87
				PERC = 25 ($\theta_c = 0.675$)				
6	.68	.38	.68	.37	.81	.55	.80	.55
8	.73	.44	.73	.44	.83	.60	.83	.60
10	.76	.49	.76	.48	.85	.64	.85	.64
12	.78	.52	.78	.52	.87	.67	.87	.67
20	.84	.61	.84	.61	.90	.74	.90	.74
40	.89	.72	.89	.71	.93	.81	.93	.81
				PERC = 10 ($\theta_c = 1.285$)				
6	.82	.27	.82	.27	.90	.45	.90	.45
8	.85	.33	.85	.33	.91	.51	.91	.51
10	.87	.38	.87	.38	.93	.55	.93	.55
12	.89	.42	.88	.41	.93	.58	.93	.58
20	.92	.51	.92	.51	.95	.66	.95	.66
40	.95	.64	.94	.63	.97	.75	.97	.75
				PERC = 5 ($\theta_c = 1.645$)				
6	.87	.22	.89	.22	.94	.40	.94	.40
8	.91	.28	.91	.28	.95	.45	.95	.45
10	.92	.32	.92	.32	.96	.50	.96	.50
12	.93	.36	.93	.36	.96	.53	.96	.53
20	.95	.46	.95	.46	.97	.62	.97	.62
40	.97	.59	.97	.58	.98	.71	.98	.71

Note. GRM = graded response model; PERC = percentage of individuals in diagnostic category.

For high item discrimination, different mean item locations had more effect on the proportions than different degrees of spread (Δ), in particular for large J. These effects were found across all PERC values.

Some Results for (θ_l, θ_u) and Corresponding T_l, T_u Intervals

The (θ_l, θ_u) intervals were shorter as test length and item discrimination increased and for polytomous items relative to dichotomous items, but their length was constant for different cut-scores (PERCs) and all J items located at the cut-score, whereas everything else remained constant (see Tables 2 and 3). This result contradicts the prediction that intervals are shorter as the cut-score is more extreme. For example, in Table 2, for $J = 6$ dichotomous items with low discrimination, one finds that the (θ_l, θ_u) intervals shift to the right of the scale as θ_c shifts to the right (i.e., as PERC becomes smaller) but that the length of each of these intervals equals approximately 1.48. This constant length is due to all J items being located at θ_c for all values of θ_c. To find the cut-score on the true-score scale, T_c, and the unreliability interval (T_l, T_u), we insert ($\theta = \theta_c$) and $b_j = \theta_c$ ($j = 1, \ldots, J$) in the 1PLM; this yields probabilities equal to .5 and, consequently, $T_c = J/2$ (Equation 7). Thus, for this setup of the computational study, T_c is always located at the middle of the true-score scale and (T_l, T_u) intervals are always located around the middle of this scale.

Next, we argue that these (T_l, T_u) intervals have the same

length. For different cut-scores θ_c, we saw that, except for small rounding errors, the length of (θ_l, θ_u) intervals was the same. A shift of θ_c and the (θ_l, θ_u) interval and the J Rasch IRFs that are located at θ_c cause an equal shift of the test response function (Figure 2). Figure 2 can be used to infer that the true scores T_l and T_u are not affected by such a shift and, as a result, that the length of the (T_l, T_u) interval is the same for different θ_c values. Unlike the results for θ_l, θ_u intervals, however, the length of (T_l, T_u) intervals remains constant even for varying item discrimination. Figure 2 can also be used to see what happens when the test response function becomes steeper (which is due to a higher value of item discrimination a for all J items) and everything else remains constant. Such an increase produces a shorter θ_l, θ_u interval, but it does not affect the (T_l, T_u) interval.

What does change when item discrimination increases is the distribution of T. Thus, the same (T_l, T_u) intervals for different levels of discrimination may have different impacts on CC proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$, and this is revealed by Tables 2–5. Some noteworthy results are given in Table 6, in which values for T were obtained by using Equations 7 and 8. The last column reveals that polytomous-item tests produce (T_l, T_u) intervals that are shorter relative to scale length than those produced by dichotomous-item tests. Useful as this may be, for classification problems as studied here one needs to consider CC proportions $P_\pi(C\bar{D})$ and $P_\pi(CD)$ to be able to evaluate the impact of such differences. Tables 2–5 show that for fixed test length, differences in CC proportions between dichotomous-item and polytomous-item tests were not impressive.

Discussion

This study has dealt with a phenomenon that is familiar, at least at an intuitive level, to everyone who has tested individuals. In particular, if someone's score on a short test, questionnaire, or inventory is close to the cut-score, we feel uncertain about the decision: admit or reject, pass or fail? More information would be helpful and, moreover, fairer to the patient or to the student. For test performance that is clearly below or above the cut-score, this concern is not felt as explicitly. The situation described here has been formalized in this study.

The results of this study show that for scales consisting of 6–12 items, random measurement error exercised an unduly large influence on CC, even when items had the best quality encountered in test practice: That is, items had good discrimination power and locations at the cut-score θ_c, where they contribute maximally to estimating θ by means of maximum-likelihood methods. For longer tests, the results were much better but became more worrisome as the cut-score was more extreme (i.e., the PERC was smaller), a result well-known from personnel selection (e.g., Wiggins, 1973). Tests consisting of polytomous items did not substantially improve CC.

The main conclusion is two-fold. First, even if items have high quality, short tests must be used only for making decisions about people who are located outside the unreliability interval for that test. Tables 2–5 can be used to find the intervals for the conditions that correspond the best with the test at hand. This implies that the

Table 6
Intervals for True Scores, Low Discrimination, Dichotomous-Item Tests (1PLM), Polytomous-Item Tests (GRM), and Different Test Lengths

J	T_l	T_u	$T_u - T_l$	$\frac{(T_u - T_l)}{\max(X_+)}$
		Dichotomous-item tests		
6	1.49	4.51	3.02	.50
8	2.22	5.78	3.56	.45
10	3.02	6.98	3.96	.40
12	3.81	8.19	4.38	.37
20	7.16	12.84	5.68	.28
40	16.00	24.00	8.00	.20
		Polytomous-item tests		
6	8.54	15.46	6.92	.29
8	12.02	19.98	7.96	.25
10	15.56	24.44	8.88	.22
12	19.10	28.90	9.80	.20
20	33.64	46.36	12.72	.16
40	71.26	88.74	17.48	.11

Note. Items maximally informative about θ_c. 1PLM = one-parameter logistic model; GRM = graded response model.

test cannot be used for all those people who are located in the unreliability intervals, unless one is prepared to make many incorrect decisions. In a particular diagnostic category, this may easily concern half of the group, as the computational study has shown. This is a situation one likely wants to avoid in many classification problems.

The second conclusion is that one needs a long test (or a composite of several short tests) if one wants the test score to produce an acceptable *CC* that satisfies a required certainty level expressing the importance of the decision. Test length is easier to manipulate than any of the other factors included in our study. For a particular classification problem, the cut-score is often fixed given properties of the test and the nature of the diagnostic categories. Dichotomous items are not easily transformed into polytomous versions of those same items. The difficulty of items often can be predicted only within global intervals. Finally, highly discriminating items often have limited meaning or validity, and moderately or even poorly discriminating items are more often the rule than the exception. To identify membership in a particular diagnostic category, it is often easier to construct a larger number of items—either dichotomous or polytomous—than to try to tailor their difficulties exactly to the cut-score and hope that their discrimination will be the highest possible in practice. And even if all of this succeeds, this study has demonstrated that a short test simply will not suffice for large numbers of people.

This study corroborated our hypothesis that test scores based on short scales contain too much measurement error to make decisions with enough certainty for the majority of respondents. Thus, some final remarks are in order.

First, even experts in test theory may find it difficult to believe that a number of well-chosen items, albeit a limited number, select so few people for (non) treatment (i.e., low *CC*) with a sufficiently high certainty level (π). The problem becomes more serious with shorter tests, despite the use of highly discriminating items that are located at the cut-score (i.e., providing a great amount of statistical information).

Second, this study has made clear that one needs *CC* proportions like $P_{\pi}(C\bar{D})$ and $P_{\pi}(CD)$ to evaluate consistency of decision making on the basis of short tests. The information function or the standard error of maximum-likelihood estimate $\hat{\theta}$ conditional on θ does not provide the information necessary for this evaluation. Exactly how Cronbach's alpha and other reliability estimates are related to consistent decision making is a topic for future research.

Third, especially in clinical and medical practice there is a tendency to work with short scales to alleviate the burden on patients who are too confused or too ill to answer large numbers of questions. Understandable as these practical considerations are, they cannot make a short test produce higher *CC*.

More research is needed that directly links the use of test scores for classification—in clinical, medical, but also job selection contexts—to classification consistency. Utility of outcomes may be included as a variable that affects the choice of certainty level π and the classification consistency. Apart from that, we predict that the conclusion will be that either long, high-quality tests (i.e., containing at least 20 and preferably 40 items; for many tests, this is not an excessive test length) are needed or that decision making should be based on many small pieces of information, each of which covers a unique aspect of the construct to be measured or the criterion to be predicted. The collection of small pieces of information requires that patients and clients need to be bothered several times but only for a short period each time. This could provide a compromise between practical demands set by the clinical or medical reality and psychometric demands to ensure consistent classification.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27,* 319–334.

Cooke, D. J., Michie, C., Hart, S. D., & Hare, R. D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist–Revised (PCL:SV): An item response theory analysis. *Psychological Assessment, 11,* 3–13.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Denollet, J. (2005). DS14: Standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic Medicine, 67,* 89–97.

Donders, J. (2001). Using a short form of the WISC–III: Sinful or smart? *Child Neuropsychology, 2,* 99–103.

Edwards, P., Roberts, I., Sandercock, P., & Frost, C. (2004). Follow-up by mail in clinical trials: Does questionnaire length matter? *Controlled Clinical Trials, 25,* 31–52.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15,* 269–294.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 2,* 189–198.

Goring, H., Baldwin, R., Marriot, A., Pratt, H., & Roberts, C. (2004). Validation of short screening tests for depression and cognitive impairment in older medically ill patients. *International Journal of Geriatric Psychiatry, 19,* 465–471.

Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10,* 19–28.

Huynh, H. (1976). On the reliability of domain-referenced testing. *Journal of Educational Measurement, 13,* 253–359.

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics* (Vol 1, 3rd ed.) New York: Hafner.

Knight, J. R., Goodman, E., Pulerwitz, T., & DuRant, R. H. (2000). Reliabilities of short substance abuse screening tests among adolescent medical patients. *Pediatrics, 105,* 948–953.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer.

Koppes, L. L. J., Twisk, J. W. R., Snel, J., van Mechelen, W., & Kemper, H. C. G. (2004). Comparison of short questionnaires on alcohol drinking behavior in a nonclinical population of 36-year-old men and women. *Substance Use and Misuse, 39,* 1041–1060.

Kosinski, M., Bayliss, M. S., Bjorner, J. B., Ware J. E., Jr., Garber, W. H., Batenhorst, A., et al. (2003). A six-item short-form survey for measuring headache impact: The HIT–6TM. *Quality of Life Research, 12,* 963–974.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–198.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453–461.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing: Principles and applications.* Englewood Cliffs, NJ: Prentice Hall.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO-PI-R. *Assessment, 7,* 347–364.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81,* 93–103.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics, 25,* 391–415.

Stuss, D. T., Meiran, N., Guzman, A., Lafleche, G., & Willmer, J. (1996). Do long tests yield a more accurate diagnosis of dementia than short tests? A comparison of five neuropsychological tests. *Archives of Neurology, 53,* 1033–1039.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265–276.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263–267.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

Taylor, J., & Deane, F. P. (2002). Development of a short form of the Test Anxiety Inventory (TAI). *The Journal of General Psychology, 129,* 127–136.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39–49.

Traub, R. E., & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement, 4,* 517–545.

Van der Linden, W. J. (2005). *Linear models for optimal test design.* New York: Springer.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Waller, N. G., Putnam, F. W., & Carlson, E. B. (1996). Types of dissociation and dissociative types: A taxometric analysis of dissociative experiences. *Psychological Methods, 3,* 300–321.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

(*Appendix follows*)

Appendix

Finding Total Score Distributions and Unreliability Intervals

Recursion Formula for Expanding $\phi(X_+|\theta)$ for Polytomous Items

Let $\mathbf{P}_j$ be the (column) vector of the CRFs of item j; that is, $\mathbf{P}_j = (P^*_{j0}(\theta), \ldots, P^*_{jm}(\theta))$; see Equation 3.

Initialization Step

In the initialization step (indexed $s = 0$), the vectors containing the CRFs of items 1 and 2 are multiplied. This results in an $(m + 1) \times (m + 1)$ matrix $\mathbf{R}^0$ that contains the joint probabilities of all combinations of category scores on items 1 and 2; that is,

$$\mathbf{R}^0 = \mathbf{P}_1 \times \mathbf{P}'_2.$$

Matrix $\mathbf{R}^0$ is transformed into the discrete distribution $\phi(X_+|\theta)$ for the first two items. This is done as follows:

$$\phi(X_+ = x_+|\theta) = \begin{cases} \sum_{l=0}^{x_+} R^0_{l+1,x_+ +1-l} & \text{if } x_+ \leq m \\ \sum_{l=x_+-m+1}^{m+1} R^0_{l,x_+ +2-l} & \text{if } x_+ \geq m+1 \end{cases}.$$

The discrete distribution $\phi(X_+|\theta)$ can be expressed as a probability vector of length $2m + 1$, which will be denoted by $\mathbf{W}^0$ and defined as $\mathbf{W}^0 = (\phi(X_+ = 0|\theta), \ldots, \phi(X_+ = 2m|\theta))$.

Recursion Steps

The initialization step is followed by $J - 2$ recursion steps. Each step s ($s = 1, \ldots, J - 2$) proceeds as follows.

1. Multiply the CRFs of item $s + 2$ with the row vector $\mathbf{W}'$ obtained from the previous step; that is,

$$\mathbf{R}^s = \mathbf{P}_{s+2} \times \mathbf{W}'^{s-1}.$$

The matrix $\mathbf{R}^s$ has $(m + 1)$ rows and $[m(s + 1) + 1]$ columns.

2. Obtain the expanded distribution $\phi^s(X_+|\theta)$ from $\mathbf{R}^s$ as follows:

$$\phi^s(X_+ = x_+|\theta) = \begin{cases} \sum_{l=0}^{x_+} R^s_{l+1,x_+ +1-l} & \text{if } 0 \leq x_+ \leq m \\ \sum_{l=0}^{m} R^s_{l+1,x_+ +1-l} & \text{if } m+1 \leq x_+ \leq m(s+1) \\ \sum_{l=x_+-m(s+1)+1}^{m+1} R^s_{l,x_+ -ms+2} & \text{if } m(s+1)+1 \leq x_+ \leq m(s+2). \end{cases}$$

The discrete distribution is expressed in the new probability vector $\mathbf{W}^s$, which is now of length $m(s + 2) + 1$.

After all steps are accomplished, we have obtained the intended distribution $\phi(X_+|\theta)$ for the set of J items.

Bisectional Method for Determining Boundaries of the Unreliability Intervals

The boundaries θ_l and θ_u for a given level of π were obtained using interval bisection with $r = 12$ iterations.

Lower Bound

The lower bound θ_l was found as follows:

$$\theta_{l,r} = \theta_{l,r-1} + \Theta_{l,r},$$

with $\theta_{l,0} = \theta_c$ and $\Theta_{l,r}$ being the shifting parameter for obtaining values of θ_l with increased precision. For the first four iterations ($r = 1, \ldots, 4$):

$$\Theta_{l,r} = \begin{cases} -0.5 & \text{if } \phi(X_+ \geq T_c|\theta_{l,r-1}) < \pi. \\ 0 & \text{otherwise} \end{cases}$$

In the remaining iterations ($r = 5, \ldots, 12$),

$$\Theta_{l,r} = \begin{cases} \dfrac{-0.5}{2^{r-4}} & \text{if } \phi(X_+ \leq T_c|\theta_{l,r-1}) < \pi \\ \dfrac{0.5}{2^{r-4}} & \text{if } \phi(X_+ \leq T_c|\theta_{l,r-1}) \geq \pi \end{cases}.$$

Upper Bound

The upper bound θ_u was found as follows:

$$\theta_{u,r} = \theta_{u,r-1} + \Theta_{u,r},$$

with $\theta_{u,0} = \theta_c$ and $\Theta_{u,r}$ being the shifting parameter for obtaining values of θ_u with increased precision. For the first four iterations ($r = 1, \ldots, 4$),

$$\Theta_{u,r} = \begin{cases} 0.5 & \text{if } \phi(X_+ \geq T_c|\theta_{u,r-1}) < \pi. \\ 0 & \text{otherwise} \end{cases}$$

In the remaining iterations ($r = 5, \ldots, 12$),

$$\Theta_{u,r} = \begin{cases} \dfrac{0.5}{2^{r-4}} & \text{if } \phi(X_+ \geq T_c|\theta_{u,r-1}) < \pi \\ \dfrac{-0.5}{2^{r-4}} & \text{if } \phi(X_+ \geq T_c|\theta_{u,r-1}) \geq \pi \end{cases}.$$

The first four iterations are used to approximately locate θ_l and θ_u with a precision of .50, assuming that θ_l and θ_u are within two standard deviations from θ_c. If this assumption is unreasonable, more iterations are used to approximately locate θ_l and θ_u. The remaining eight iterations are used to obtain more precise values of θ_l and θ_u. In our study, the eight remaining iterations guarantee that the imprecision of θ_l and θ_u is smaller than $\dfrac{.5}{2^{r-4}} < .002$. More precise values of θ_l and θ_u can be obtained by using more iterations (r).

Received August 10, 2005
Revision received November 6, 2006
Accepted November 29, 2006 ■