Journal of Computerized Adaptive Testing

Volume 1 Number 2

February 2013

A Comparison of Computerized Classification Testing and Computerized Adaptive Testing in Clinical Psychology

Niels Smits and Matthew D. Finkelman

DOI 10.7333/1302-0102019

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 1, Number 2, February 2013
DOI 10.7333/1302-0102019
ISSN 2165-6592

A Comparison of Computerized Classification Testing and Computerized Adaptive Testing in Clinical Psychology

Niels Smits, *VU University*
Matthew D. Finkelman, *Tufts University*

The appropriateness of computerized classification testing (CCT) and computerized adaptive testing (CAT) as methods for efficient administration of self-report questionnaires in clinical psychology are compared, when classification is an important test goal. Simulated data sets were used to compare the two methods and to study the effect of latent trait distributions and number of items administered on the quality of clinical measurements and decisions. CAT and CCT outcomes were very similar. The implications of these findings for assessment in clinical psychology are discussed.

Keywords: *computerized classification testing, computerized adaptive testing, item response theory, clinical psychology, disease prevalence*

In clinical psychology, self-report questionnaires for measuring attributes associated with common mental disorders such as anxiety and depression are often used, both in research settings and clinical practice. Examples of these are the Center for Epidemiologic Studies Depression Scale (Radloff, 1977), Beck's Depression Inventory (Beck, Steer, & Carbin, 1988), the Hamilton Anxiety Scale (Hamilton, 1959), and the Mood and Anxiety Symptom Questionnaire (Watson & Clark, 1991). In addition, because the new version of the Diagnostic and Statistical Manual of Mental Disorders (DSM), to be published in 2013, will incorporate dimensional measures into the existing classification system (Helzer et al., 2008), self-report questionnaires will become even more important in the future.

In the last decade, computerized adaptive testing (CAT) has become a popular method for efficient administration of the items of clinical scales. For example, the Patient Reported Outcomes Measurement Information System (PROMIS; Cella et al., 2007) is currently developing CATs for the measurement of emotional distress (Pilkonis et al., 2011) that allow for monitoring the mental health of medical patients. In addition, both German (Fliege et al., 2005, 2009; Walter et al., 2007) and Dutch (Roorda, 2011) CATs have been or are being developed for measuring depression and anxiety in similar populations. Finally, in Routine Outcome Monitoring (ROM; Carlier et al., 2010), a method devised to collect data on the effectiveness of treatments in mental health institutions, the efficiency of CATs has been studied as well (Smits, Zitman, Cuijpers, den

Hollander–Gijsman, & Carlier, 2012). The adaptive testing algorithms used in these CATs are based on item response theory (IRT) and are driven by increasing the precision of the resulting measurements.

Although interested in measurement, many clinical psychologists attach more value to diagnostic accuracy. Especially in clinical practice, self-report questionnaires are used to select examinees with a high probability of pathology. As in most selection situations, prominence is given to the utility of a measure with reference to predicting an external criterion (Weitzman, 1982), commonly a diagnosis by a clinician. Such a judgment entails the assignment of an examinee to one of two categories, "healthy" or "diseased," and the self-report measure is thus used for classification decisions (Cronbach & Gleser, 1965). Therefore, adaptive algorithms for classification might seem more appropriate than standard CATs.

Such adaptive procedures for classification have been developed in educational settings with the purpose of separating "masters" from "non-masters" (e.g., Parshall, Spray, Kalohn, & Davey, 2002). These algorithms are similar to standard CAT, but instead of optimizing measurement precision, they optimize the classification accuracy in the neighborhood of a cut score relevant for decision making. In a setting with clinical scales, such computerized classification testing (CCT) has been called "clinical decision adaptive testing" (Waller & Reise, 1989). Although it has been suggested that CCT be used instead of CAT in clinical psychology (Embretson & Reise, 2000; Smits, Cuijpers, & van Straten, 2011), apart from an illustration by Waller and Reise (1989), CCT has never been used in the clinical field. A thorough study of the usefulness of CCT, in contrast to CAT, in clinical psychology is therefore needed.

Of considerable importance in this context is the appropriateness of existing item banks for different populations of examinees. In ability or achievement testing, it is known that item banks designed for the average ability can provide much less information in subpopulations with more extreme abilities (see Gorin, Dodd, Fitzpatrick, & Shieh, 2005). Because item banks designed for use in clinical psychology are used both in the general population (such as in PROMIS) and clinical populations (such as in ROM), they might provide very different amounts of information. Consequently, the magnitude of the potential advantage of CCT over CAT, if any, might be very dissimilar in these different populations. In a study of the usefulness of CCT in clinical psychology, the effect of score distributions should, therefore, be included.

This study had two purposes. The first was to investigate whether it is more appropriate to use CAT instead of CCT when clinical decision making is an important test goal. The second was to determine if the magnitude of a potential difference between CCT and CAT is moderated by the distribution of clinical scores. In a simulation study, both CAT and CCT procedures were employed on artificially generated clinical data; clinical score distributions were systematically manipulated, and effects on the quality of clinical measurements and decisions were studied.

Method

Overview of Procedures

This simulation study was intended to mimic the development and use of adaptive tests typically encountered in clinical psychology and related disciplines (e.g., Fliege et al., 2005; Walter et al., 2007). First, the scores on a bank of items of a large sample from a relevant population were obtained to estimate the item parameters. Next, adaptive algorithms were developed using these estimates. Finally, the constructed procedures were employed in a simulated adaptive administration of the item bank in a new sample from the same population. The choices made with reference to characteristics such as the size of the item bank, and the estimation method of the

person parameter were also similar to those encountered in the field. For the comparison of CAT and CCT, two variables were manipulated in the simulation: (1) the distribution of the latent trait in the population, and (2) the number of administered items.

Construction of Item Banks

The estimated IRT models of several CATs developed for assessing anxiety and depression were studied (e.g., Fliege et al., 2005; Forkmann et al., 2009; Gardner et al., 2004; Smits et al., 2011). As a starting point of item bank construction, the data (N = 766) of a 29-item anxiety bank of the PROMIS project (Pilkonis et al., 2011) were used. This is an accompanying data file in the LORDIF library (Choi, 2011; Choi, Gibbons, & Crane, 2011) of the R package, and a typical example of what was available in the reviewed literature. Since this bank has 5-point Likert scale items, a polytomous IRT model had to be chosen. The CATs developed in clinical settings either use the partial credit model (Muraki, 1992) or the graded response model (GRM; Samejima, 1969). The GRM was chosen because of its ease of understanding (see, e.g., Mellenbergh, 1995). The GRM parameter estimates were used to construct a population distribution from which item parameters were drawn for the simulation. To that end, the vector of means and matrix of covariances of the five parameter estimates were calculated over the 29 items. Inspection of these estimates suggested a truncated normal distribution for each of the five types of parameters. The estimates were idealized (see Table 1) and used as the input for a truncated multivariate normal distribution, which, in addition to a mean vector and a covariance matrix, requires a minimum and maximum value for each parameter. Item sets resulting from this distribution had item information and test information functions that were similar to those of Pilkonis et al. (2011) and therefore to those typically found in the field (also see, e.g., Smits et al., 2011). Figure 1 shows the item information functions of a randomly chosen data set from the simulation. The CATs encountered in the literature had different item bank sizes: Smits et al. (2011) used the smallest bank, which consisted of 20 items, whereas Fliege et al. (2005, 2009) used the largest bank, which consisted of 64 items. In the present simulation, an item bank was used that had a size approximately in between these two extremes. For each data set, 40 items were drawn from the population item parameter distribution as input for the generation of item scores.

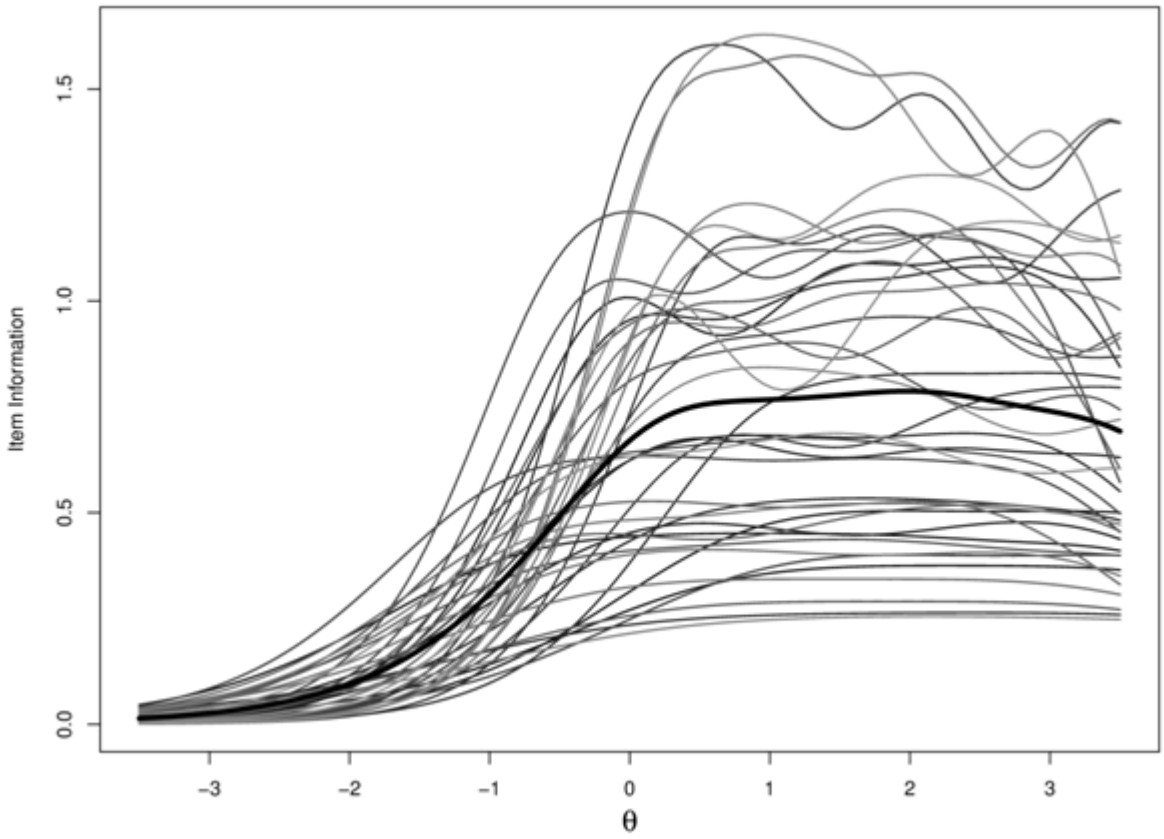
Table 1. Characteristics of the Truncated Multivariate Normal Population Distribution of GRM Parameters

	Statistic			Parameter Covariance Matrix				
Parameter	Min.	Mean	Max.	a	b_1	b_2	b_3	b_4
a	1	2	3	0.40				
b_1	−1	0	1	0.10	0.25			
b_2	0	1	2	0.05	0.20	0.25		
b_3	1	2	3	−0.05	0.17	0.20	0.25	
b_4	2	3	4	−0.10	0.14	0.17	0.20	0.25

Simulated Data Generation

The data generation procedure began by selecting latent trait (θ) values for the simulees. These values were drawn from normal distributions, which had different mean values in three populations (0.00, 0.76, and 1.28, respectively; the standard deviation was 1 in all cases). In all three populations, the same single critical θ cut score of 1.28 applied, above which simulees were

Figure 1. Item Information Functions of a Randomly Chosen Data Set in the 10% Prevalence Population (The thick line is the average item information)

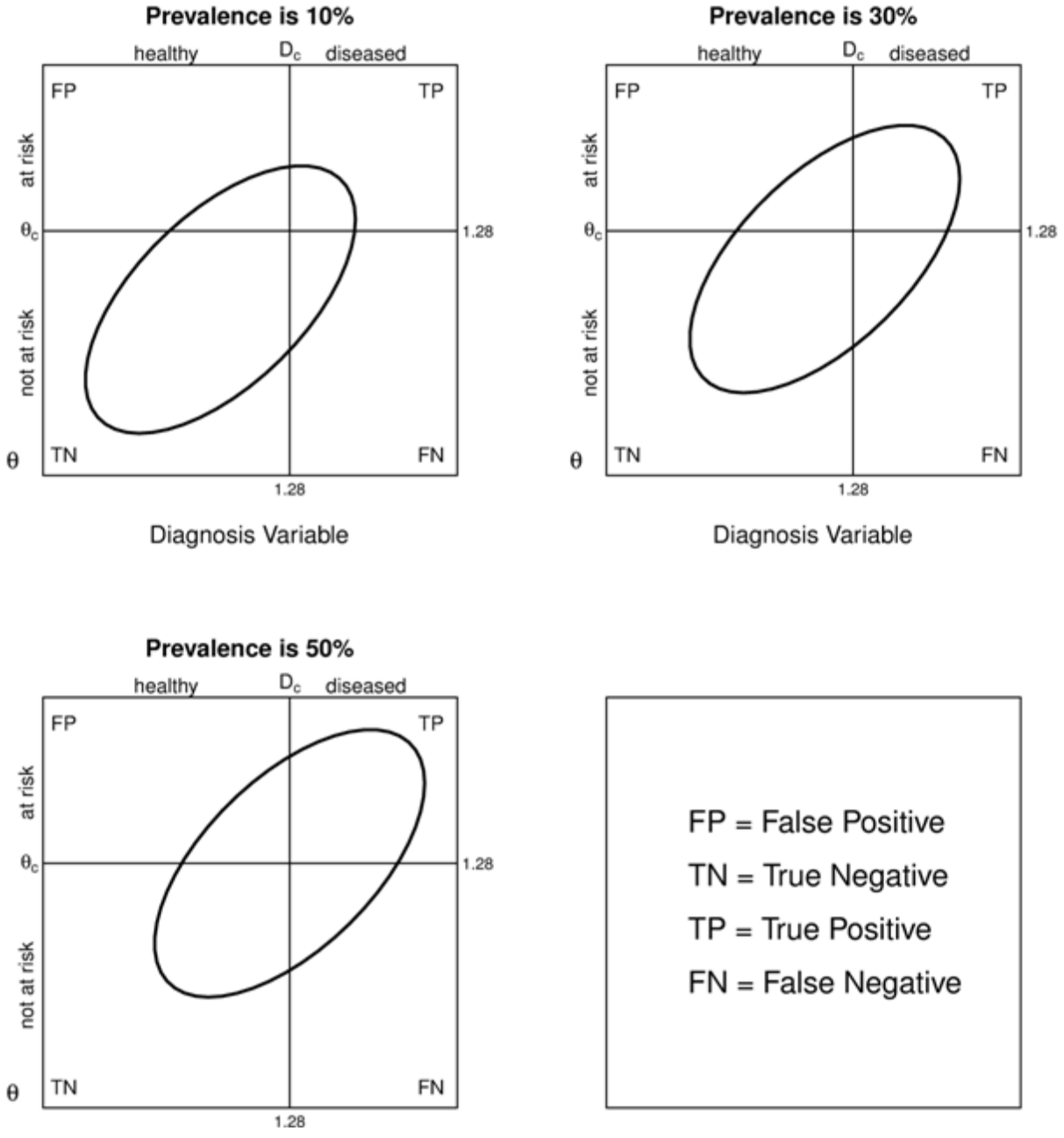


defined to be "at risk," and at or below which simulees were defined to be "not at risk" (this corresponds to the master/no master classification in mastery testing). Because the populations had different means, they had a different rate of at-risk as well. The resulting at-risk levels were 10%, 30%, and 50% (see Figure 2).

In mastery testing simulations, interest lies in the correspondence between decisions based on the test and those based on true θ; the criterion is therefore of an internal nature. By contrast, in clinical testing interest lies in the correspondence of test decisions with those based on an external criterion (e.g., Mellenbergh & van der Linden, 1979). Therefore, compared to simulation studies in mastery testing (e.g., Thompson, 2011), an additional criterion variable was simulated as well: a diagnosis on the basis of a clinical interview. Such a diagnosis is commonly seen as the "gold standard." Naturally, in the clinical field, a perfect relationship between test scores and this gold standard is never found.

To model this situation, another random variable was simulated that had identical population distributions (mean and standard deviation) as θ, and a fixed correlation, ρ, with θ. This diagnosis variable was dichotomized using a cut off of 1.28 (the 90% quantile value of the standard normal distribution): if the simulated diagnosis value was higher than this cut off, the simulee received a positive diagnosis (D = 1); if not, a negative diagnosis (D = 0). The value for ρ was chosen based on the data of Smits et al. (2011, 2012), two studies on CATs for assessing depression. For each study the biserial correlation was calculated between the latent depression esti-

Figure 2. Correct and Incorrect Classification Decisions for the Three Prevalences on the Population Level (D_c and θ_c are the cut scores on the diagnosis variable and θ, respectively)



mate and a dichotomous clinical diagnosis for depression. In both studies, this correlation was about 0.60; therefore this value for ρ was used in the simulation.

The resulting samples had three different prevalence percentages of diseased simulees: 10%, 30%, or 50%. The first value can be considered a low prevalence, and corresponds to what is often found in community populations, such as those for which the PROMIS project has been de-

veloping CATs for anxiety and depression. The third value, a prevalence of 50%, corresponds to what is often found in populations visiting mental health institutions, such as those monitored in ROM. The second value was chosen as an intermediate situation between these two extreme values. Note that in the population, the at-risk rate, i.e., the proportion of simulees having θ estimates above the cut score, and the prevalence were identical, as shown in Figure 2; in what follows, these two terms will therefore be used interchangeably. In addition, note again that at the population level, even when θ is known, classification errors with respect to the diagnosis variable (so-called "false positives" and "false negatives") are made. This simulation procedure, with the indicated ρ and prevalence levels, gave rise to area under the receiver operator curves with reference to predicting the clinical diagnosis using θ of about 0.80, a value that is often encountered in clinical measurement (see, e.g., Gardner et al., 2004; Smits et al., 2011). It was thus concluded that the simulated data were representative for assessment in clinical psychology.

For each data file a new set of item parameters and θs were randomly drawn and these were used as input for generating item scores on 40 5–category Likert scale items under the GRM. Mellenbergh (1995) described the GRM as a cumulative probability model: the K answer categories are divided into $K-1$ cumulative probabilities. The kth cumulative probability is the probability that category $k+1$ or higher is chosen. The GRM for a simulee with latent trait θ and responding to a five–category item j is of the form

$$P_{jk}^{*}(\theta)=\frac{\exp\left[a_j\left(\theta-b_{jk}\right)\right]}{1+\exp\left[a_j\left(\theta-b_{jk}\right)\right]},\ k=1,2,3,4. \tag{1}$$

The probability that the simulee would receive a particular category score k on item j can be obtained by subtracting adjacent values of $P_{jk}^{*}(\theta)$,

$$P_{jk}(\theta)=P_{j(k-1)}^{*}(\theta)-P_{jk}^{*}(\theta),\ k=1,2,3,4,5. \tag{2}$$

In addition, the corresponding item information function (see Figure 1) can be written as

$$I_j(\theta)=\sum_{k=1}^{5}\frac{\left[P_{jk}^{'}(\theta)\right]^2}{P_{jk}(\theta)}, \tag{3}$$

where $P_{jk}'(\theta)$ is the first derivative of $P_{jk}(\theta)$ (e.g., Dodd, de Ayala, & Koch, 1995).

In the simulation, the item parameters and θ value were entered in Equation 1, giving four cumulative probabilities for each item. Next, these were entered into Equation 2 to obtain five item category probabilities. Finally, a category score was randomly drawn from the resulting multinomial distribution.

For each data file, the scores of 1,500 examinees were simulated. The scores of a randomly drawn 1,000 simulees formed the training set, which was used to estimate the parameters; the remaining five hundred simulees were used as the test set for the adaptive procedures. One hundred replications of the adaptive testing procedures were conducted in each prevalence population to avoid imprecise results in estimation due to sampling error. Consequently, a total of three hundred data files were generated.

Parameter Estimation

In the training sets, the GRM threshold and discrimination parameter estimates of the items were obtained with the LTM library (Rizopoulos, 2007, 2006) in R (R Development Core Team, 2010), using marginal maximum likelihood (Bock & Aitkin, 1981). This estimation method assumes that θ follows a standard normal distribution. The GRM was run for the entire matrix of simulated item response data in each training set, i.e., for the 1,000 simulees and 40 items.

Population and Calibration Scales

Although the three prevalence populations differed in true θ means (0.00, 0.76, and 1.28, respectively), as a result of marginal maximum likelihood estimation, the item and person parameters were calibrated on the standard normal distribution scale (i.e., in each sample, the average $\hat{\theta}$ was 0.0, and the standard deviation was 1.0). To translate the original scale into the calibration scale, and vice-versa, a linear transformation was needed. For example, to express the original cut score of 1.28 in terms of the resulting calibration scales, it was necessary to subtract the original latent means from this cut score, which would give cut scores of 1.28, 0.52, and 0, on the calibration scale of the 10%, 30%, and 50% at-risk rate populations, respectively. In the analysis, at-risk classification of true θ was made using the original 1.28 cut off, whereas the classification of estimated θ was made using calibration scale cut offs. For clarity, however, in what follows procedures and outcomes are discussed in terms of the true (generating) θ scale.

Adaptive Testing Simulations

Adaptive testing algorithms generally have five basic components (Weiss & Kingsbury, 1984; Wainer, 2000): (1) a calibrated item bank, (2) a procedure for estimating θ, (3) a stopping rule, (4) an item selection method, and (5) a starting level of θ for the administration of the first item. The first three components were identical for CAT and CCT in this study. The calibrated item set (Component 1) resulted from estimating GRM item parameters in the training data set.

Two θ estimation methods (Component 2) are generally available in IRT: maximum likelihood (ML) and Bayesian (Embretson & Reise, 2000). The ML approach estimates θ as that value which has the highest likelihood of bringing forth the responses observed (Thissen, 1991). By contrast, Bayesian estimation uses, in addition to this likelihood, an a priori population distribution of the latent variable, such as the standard normal. Because of this prior distribution, Bayesian estimation can, and ML estimation cannot, provide an estimate for item response patterns consisting exclusively of either extreme lower or extreme higher categories. In clinical test applications, at least a small portion of responders is expected to score very low on the mental health measure, and their response patterns will therefore consist only of extreme lower category answers. In such applications, Bayesian procedures seem more appropriate than ML procedures. In this study, Bayesian expected a posteriori (EAP; Bock & Mislevy, 1982) was used, which assumed θ to follow a normal distribution, for both CAT and CCT. EAP is a method that has been used in many mental health CATs (see, e.g., Fliege et al., 2005, 2009; Walter et al., 2007).

CCT procedures and CAT procedures are quite different in their stopping criteria (Component 3) (Thompson, 2009). However, for an unequivocal comparison of these methods in this study, a uniform stopping criterion was needed. Therefore, the number of items administered was used as the stopping criterion. Both procedures were run for seven items. This number was based on the typical outcome that, on average, "about four to seven" administered items is deemed enough for mental health CATs (see, e.g., Fliege et al., 2005; Gardner et al., 2004; Smits et al., 2011; Walter et al., 2007). The measurement and classification outcomes were recorded after the

administration of each item, resulting in seven levels of the stopping rule.

The CAT algorithm selected items (Component 4) using maximum information under the estimated GRM for the current estimate of θ (Embretson & Reise, 2000; Wainer, 2000); the starting level (Component 5) was set to the average value of θ in the training data set, as is commonly done in mental health CATs (e.g., Fliege et al., 2005; Walter et al., 2007). As mentioned above, the average $\hat{\theta}$ in each training data set was zero. As a consequence, the item with the highest information at this initial value was chosen as the first item for all simulees in the test set. Note that in terms of the true (generating) θ scale, because the θ means differed between prevalence conditions, the CATs started at different locations (about 0.00, 0.76, and 1.28, respectively, for the 10, 30, and 50% prevalences).

The CCT algorithm selected items by using maximum information at the cut score of θ (Component 4, Thompson, 2009); the cut score was the starting level (Component 5) of the procedure, as well. Consequently, all simulees had the same sequence of administered items. In the training sets, the cut score was determined as follows. Because in the population the at-risk rate and prevalence were identical, the diagnosis variable could be used to determine the cut score (also see Waller & Reise, 1989, p. 1056). In each sample, the disease prevalence was estimated using the diagnosis variable, and one minus this proportion was used as input for the quantile function of the standard normal distribution to obtain the cut score. As a result, in terms of the original scale, all CCTs had cut scores of about 1.28 (due to sampling errors in the prevalence estimate, actual values fluctuated somewhat around 1.28).

In both algorithms, after each item, an update of $\hat{\theta}$ was obtained, and this estimate was used for an interim classification, flagging a simulee as "at risk" only when the estimate exceeded the cut score, and as "not at risk" when the estimate was equal to or lower than the cut score. It should be noted that this classification rule did not take into account the uncertainty associated with that estimate (i.e., its standard error).

A program, which comprised an alteration of, and additions to the code of the LTM library, was written in R to simulate the two adaptive procedures. First, the estimated item parameters and cut score within a specific training data set were stored. Next, the test set was used to examine the adaptive procedures employing these stored outcomes as input.

Criterion Variables

Two types of outcomes were studied. The first type of outcome was associated with the congruence between the true (generating) θ and observed $\hat{\theta}$. This type of outcome is stressed in methodological research on CAT and CCT, and is referred to as *internal accuracy* in the current study. The second type of criterion variable is associated with the congruence of the estimated θ and the diagnosis variable. This type of outcome, which is related to diagnostic accuracy, is what many clinical psychologists—heavily influenced by the medical field—focus on, and is referred to here as *external accuracy*.

Internal accuracy. The first measure of internal accuracy was the fidelity coefficient (Weiss, 1982), the correlation between true and estimated θ. The second was the proportion of correct decisions, an outcome often studied in CCT (e.g., Eggen, 1999; Waller & Reise, 1989), which is the rate of simulees for which true θ and $\hat{\theta}$ gave identical classifications. The third measure was the Type I error rate (see, e.g., Thompson, 2011; Weitzman, 1982), which is the rate of simulees that received a positive classification ($\hat{\theta}$ > cut score), but should have a negative classification, having a true θ below the cut score. The fourth measure was the Type II error rate (Thompson, 2011; Weitzman, 1982), the rate of simulees who had a negative classification ($\hat{\theta}$ < cut score),

but should have a positive classification on the basis of true θ.

External accuracy. The first external accuracy measure was the point-biserial correlation between $\hat{\theta}$ and the diagnostic variable, referred to as "predictive utility" (McDonald, 1999). The other two measures expressed the quality of the $\hat{\theta}$-based classifications in terms of the two conditional probabilities describing performance with reference to the diagnosis variable (see, e.g., Kraemer, 1992; Pepe, 2004). Sensitivity is the probability that a diseased person (D = 1) has a $\hat{\theta}$ higher than the cut score, i.e., is tested as such. Specificity is the probability that a healthy person (D = 0) has a $\hat{\theta}$ lower than the cut score, i.e., has a negative test outcome.

Data Analysis

The design of this study had three factors: between factor prevalence (10%, 30%, and 50%), within factors adaptive procedure (CAT and CCT), and number of items administered (1 to 7), producing a 3 × 2 × 7 mixed design with 100 replications on the between factor, yielding a total of 300 simulated data sets. The results of the study are presented in terms of the mean outcome statistics over the 100 replications under each of the three experimental conditions. The data analyses consisted mainly of studying scatterplots of mean outcomes.

Results

Internal Accuracy

Figure 3 presents the average fidelity coefficient as a function of prevalence, number of items administered, and adaptive procedure. As is to be expected, the congruency between true and estimated θ increased with the number of items administered. The fidelity coefficient increased with prevalence (i.e., the average of true θ), which resulted from the information function of the item bank peaking on the right-hand side of the scale; these banks provided more information for the high than low prevalence population. In addition, the fidelity coefficient was consistently higher for CAT than for CCT. This difference became smaller, however, as prevalence (i.e., average θ) increased. In the high (50%) prevalence condition, it hardly mattered for measurement precision if CCT instead of CAT was used.

Figure 4 shows the mean proportion of correct decisions. This proportion was highest in the low prevalence condition (the condition with lower true θs), which is to be expected because prediction is easier when the bulk of examinees is far from the cut off. Similarly, it can be argued that classification is easier if the distribution of two categories is very dissimilar than if it is about equal (e.g., giving every observation a not-at-risk classification gives better results in the former than in the latter case). In addition, the proportion of correct decisions increased with the number of items administered, and the rate of increase changed with prevalence (at-risk rate). The latter outcome can be explained from both a ceiling effect and items providing more information in the higher prevalence conditions (see fidelity coefficient results). The proportion of correct decisions was about equal for CCT and CAT in all conditions with an exception of the first item in the 30% prevalence condition, where CCT was higher; this small difference, however, seems of very little practical importance.

Figure 3. Fidelity Coefficients for Three Prevalence Rates

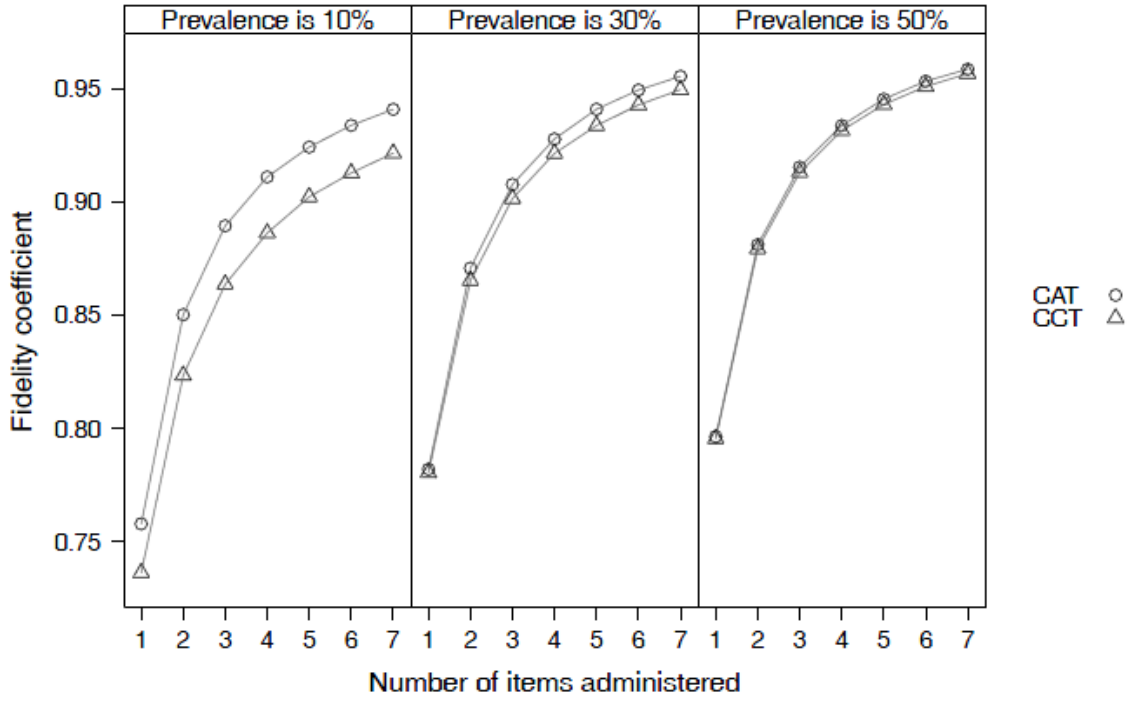
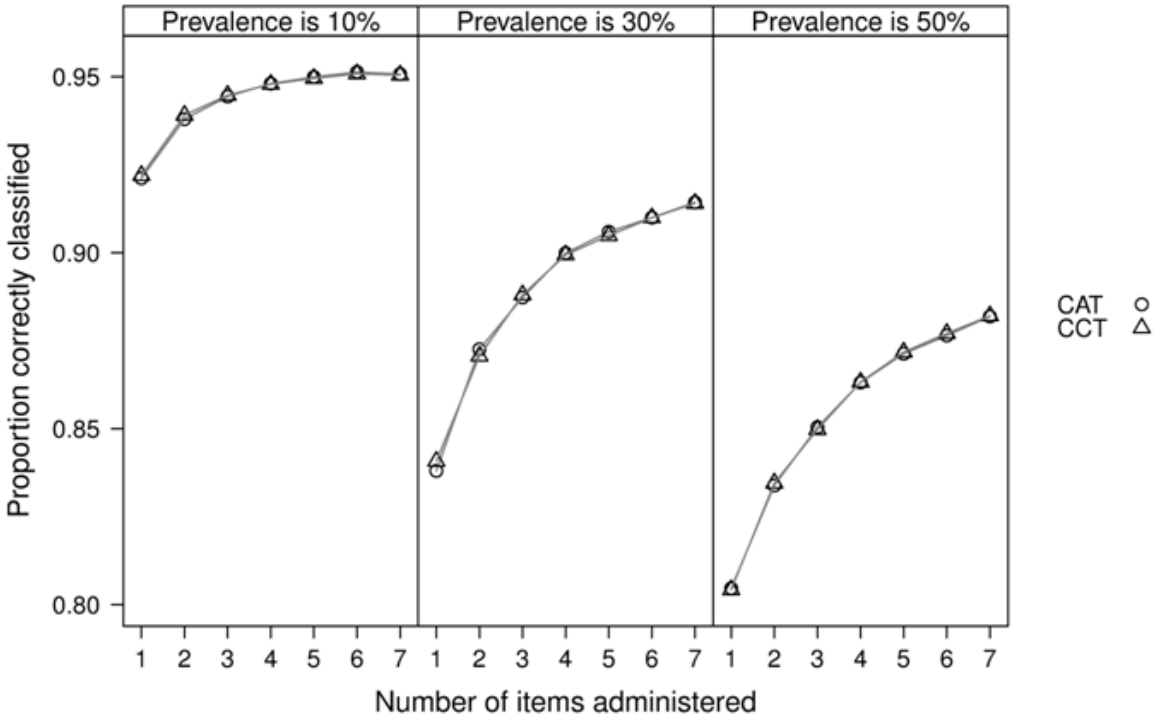


Figure 4. Proportion of Correct Decisions for Three Prevalence Rates



The results for the Type I error rate are presented in Figure 5a. Type I errors were more often made in high prevalence conditions (i.e., conditions with a higher at-risk rate). This is to be expected because the bivariate distributions of θ and $\hat{\theta}$ were very similar to those in Figure 2 (substitute θ for diagnosis on the *x*-axis, and $\hat{\theta}$ for θ on the *y*-axis): the proportion of simulees with a false at-risk classification relative to the rate of true not-at-risk simulees increased with prevalence. The effect of the number of items administered was ambiguous. In the high prevalence condition, the Type I error rate strongly decreased with the number of items, whereas in the low prevalence condition, this rate increased somewhat. Moreover, in the 30% prevalence condition, a decrease in Type I errors leveled off after the fourth item. This pattern of outcomes resulted from EAP estimation. The prior distribution in EAP draws the estimate toward the mean of θ; with more items administered, this shrinkage decreases. All this has little effect in the high prevalence situation because the cut score is about equal to the mean of θ, and although the distance of an estimate to this score might change due to an decrease in shrinkage, it tends to stay on the same side of the cut score, and therefore does not change its classification. By contrast, in the low prevalence situation, because the cut score is located on the right-hand side of the mean of θ, a reduction in shrinkage results in estimates crossing the cut score from left to right, leading to more Type I errors. More important, however, was the outcome that the rate of Type I errors was about equal for CCT and CAT in all conditions.

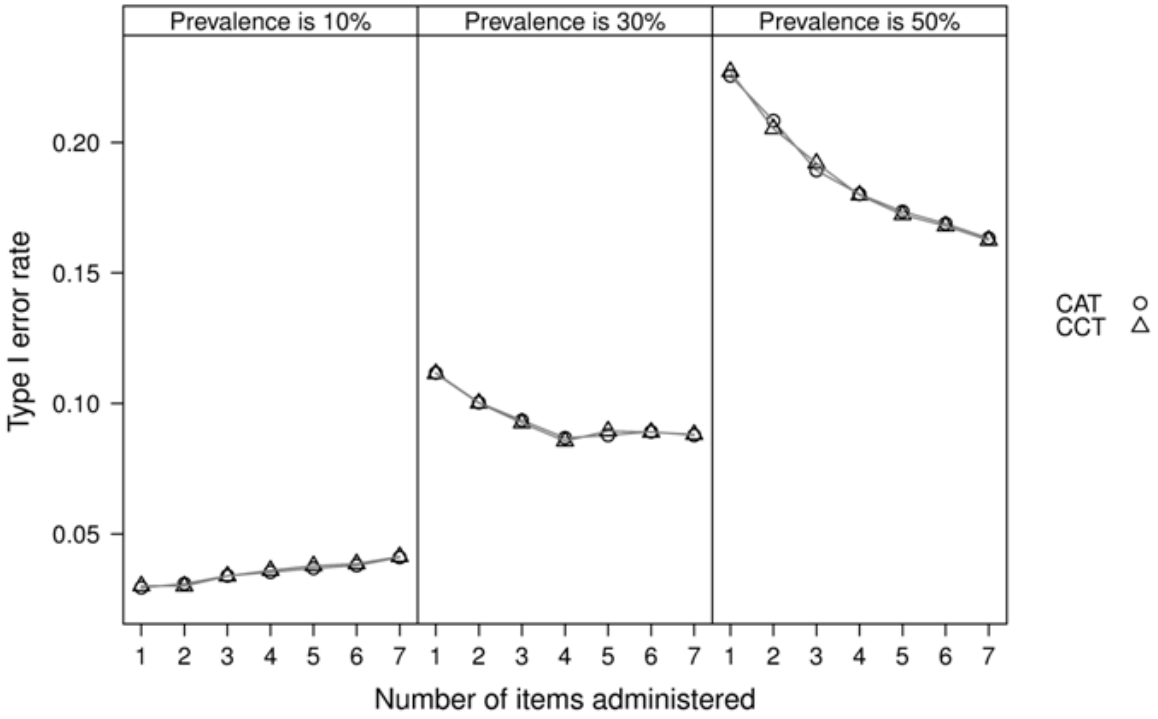
Figure 5b displays the mean Type II error rates. The Type II error rate was higher in the low prevalence condition (i.e., the condition with the lowest at-risk rate). This was anticipated because of the bivariate distributions of θ and $\hat{\theta}$ (compare Figure 2, with θ substituted for diagnosis, and $\hat{\theta}$ for θ): the proportion of simulees with a false not-at-risk classification relative to the ratio of true at-risk simulees decreased with prevalence. In addition, the Type II error rate decreased with the number of items administered; the rate of change decreased as prevalence increased. The Type II error rate was somewhat higher for CCT than for CAT after the administration of the first item; this difference, however, disappeared after the second item, and was absent in the 50% prevalence condition.

External Accuracy

Figure 6 presents the average of the utility of $\hat{\theta}$ for predicting the diagnostic outcome. These point-biserial correlations were lower with decreasing prevalence. In addition, note that all correlations were lower than the ρ of 0.60 (between true θ and the original continuous diagnostic variable) used for data generation. These outcomes are a typical result of dichotomizing: correlations lose size, and this loss is larger as splitting departs from the mean (e.g., Cohen, 1983): in the lower prevalence conditions, dichotomization was applied further from the mean, and therefore, predictive utility was lower. As the number of items administered increased, predictive utility increased, as well. CAT and CCT gave very similar outcomes, except for the lowest prevalence condition in which CAT had consistently lower values than CCT. An inspection of $\hat{\theta}$ distributions showed that, although mean differences between the two diagnostic groups were larger for CAT, this resulted from CCT having a somewhat smaller standard deviation than CAT. Note, however, that the differences were only in the third decimal place, and therefore seem of little practical importance.

Figure 5. Type I and Type II Error Rates for Three Prevalence Rates

a. Type I Errors



b. Type II Errors

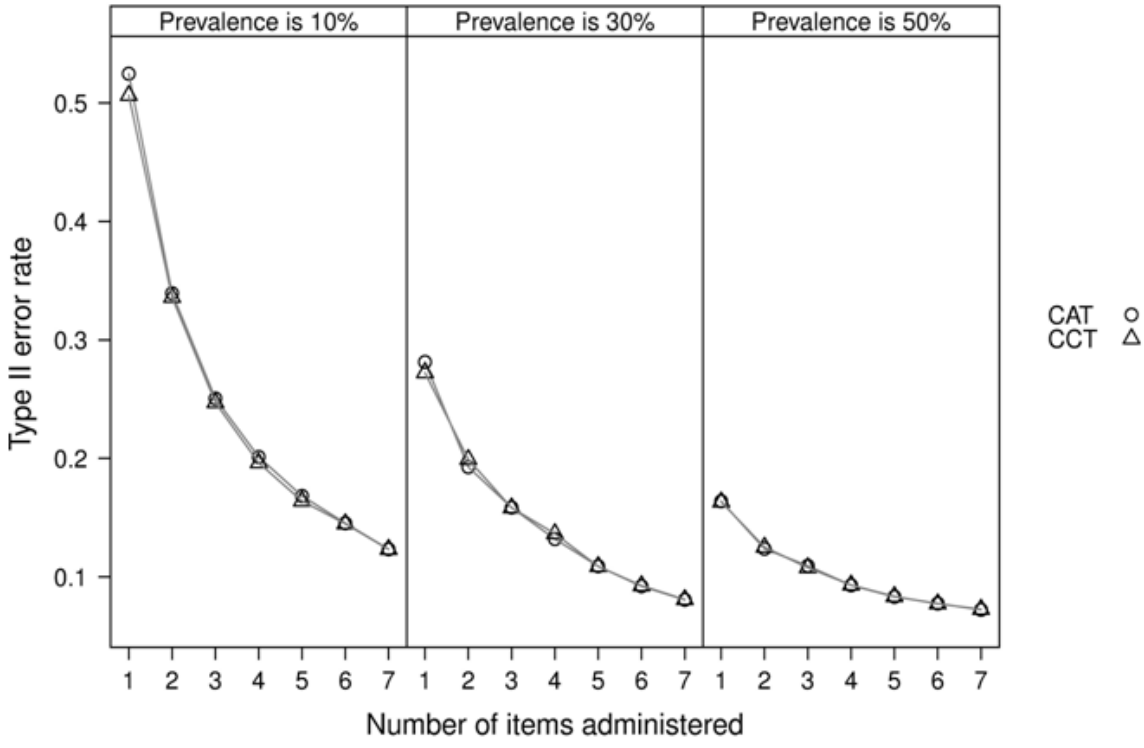


Figure 6. Predictive Utility Outcomes for Three Prevalence Rates

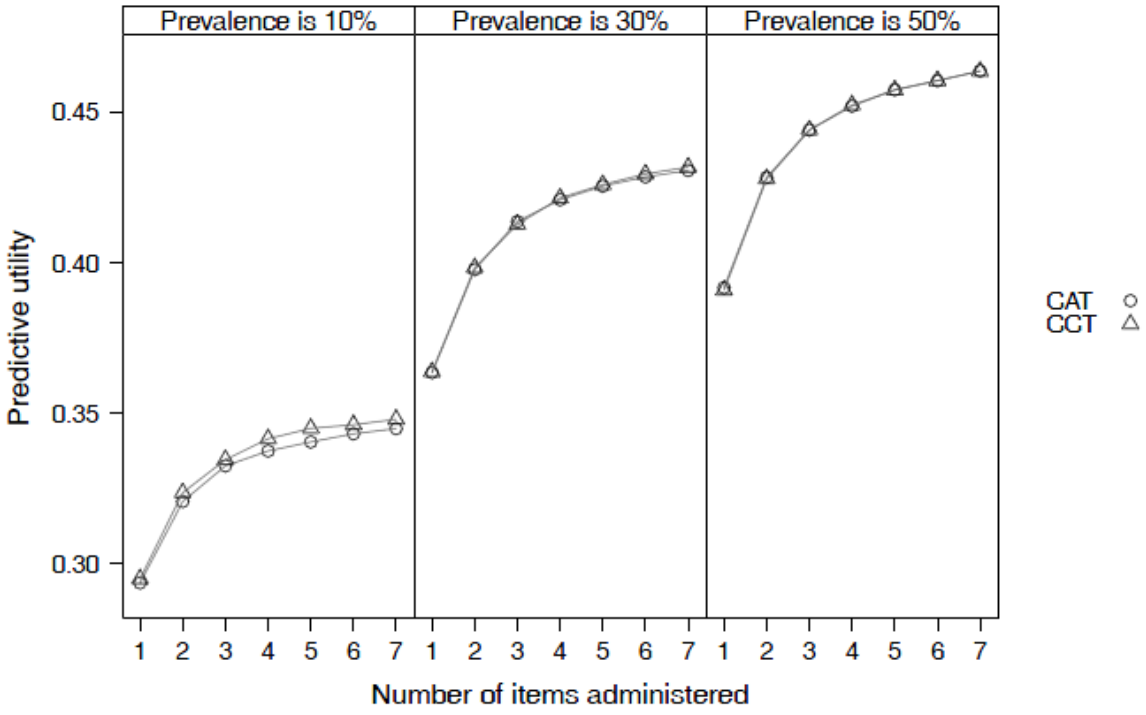
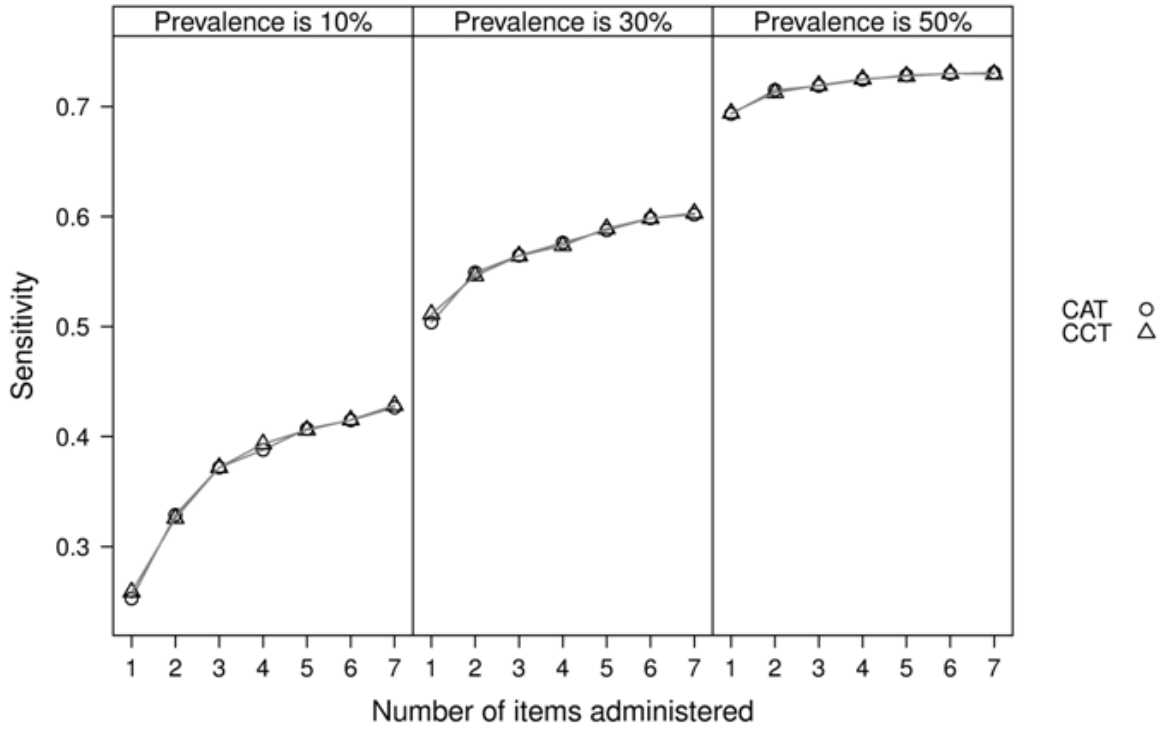


Figure 7a shows the mean sensitivity as a function of prevalence, number of items administered, and adaptive procedure. Sensitivity was higher with increasing prevalence, which is to be expected based on Figure 2: the rate of true positives relative to the sum of true positives and false negatives increases with prevalence. In addition, sensitivity increased with the number of items administered; the rate of change decreased as prevalence increased. Sensitivity was slightly higher for CCT than for CAT after the administration of the first item; this difference, however, disappeared after the second item, and was absent in the 50% prevalence condition.

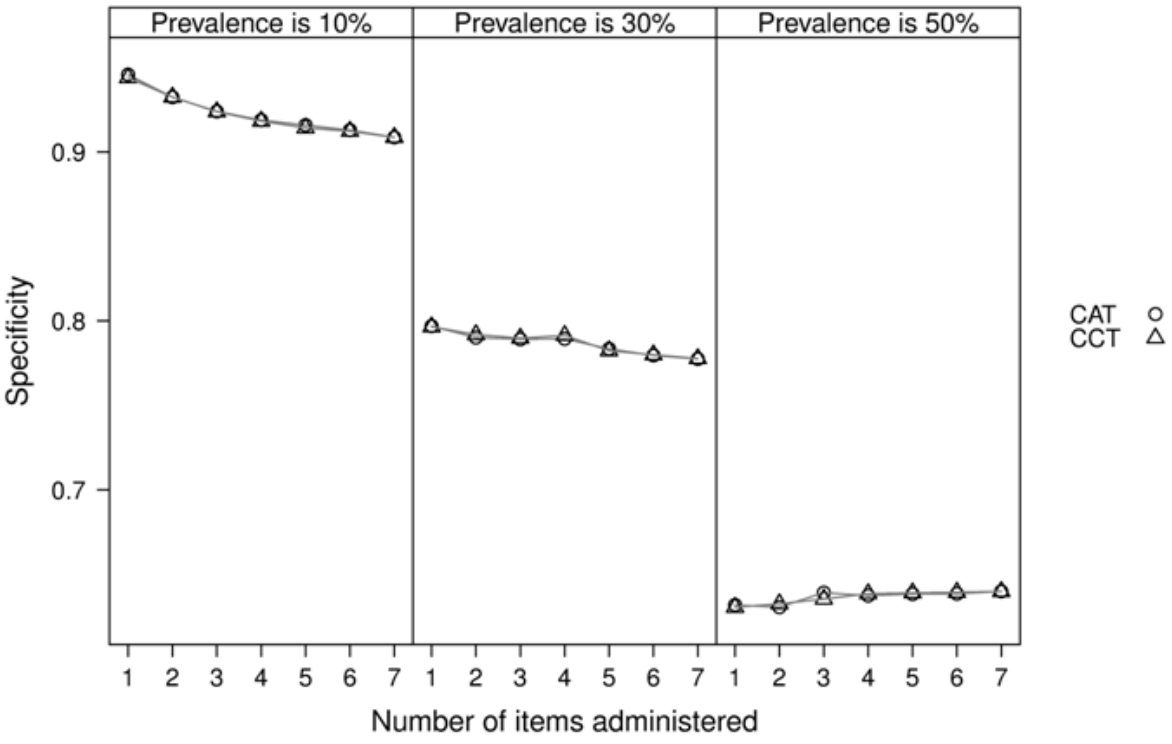
Figure 7b displays specificity outcomes. Specificity was lower in high prevalence populations, which was anticipated. Figure 2 shows that the rate of true negatives relative to one minus prevalence (true negatives plus false positives) decreases with prevalence. The effect of the number of items administered was ambiguous. In the high prevalence condition, specificity showed a mild monotone increase with the number of items, whereas in the two lower prevalence conditions, it decreased somewhat. This pattern of outcomes is similar to those of the Type I error rate results, and once more EAP estimation explains these outcomes. The decrease in shrinkage with more items administered had little effect in the high prevalence situation, but in the lower prevalence situations, some estimates crossed the cut score thus producing more false positives. The most important outcome, however, was that specificity was about equal for CCT and CAT in all three conditions.

Figure 7. Sensitivity and Specificity for Three Prevalence Rates

a. Sensitivity



b. Specificity



Discussion and Conclusions

Measurement precision, as expressed in the fidelity coefficients, was generally higher for CAT than for CCT. These differences nearly disappeared, however, in the 50% prevalence population. In addition, the utility of θ estimates for predicting the diagnostic outcome was very similar for CAT and CCT; an exception was the low prevalence condition, in which correlations were marginally higher for CCT than for CAT. Both the Type I and II error outcomes (internal at-risk classification) and the sensitivity and specificity outcomes (for predicting external diagnosis) were nearly identical for CCT and CAT. If there were any differences, they were of very little practical importance and/or disappeared after the administration of the second item. These outcomes seem to suggest that if classification is the test goal, it does not matter if CCT instead of CAT is used.

An explanation for the current outcomes can be the information that the items provided for the different populations. Although the means differed substantially in the three populations, they were located at or above the center of the θ scale. Therefore, even in the 10% and 30% prevalence conditions items that were informative for current $\hat{\theta}$ were informative for the cut score, as well. Inspection of the item information functions in Figure 1 illustrates this. For example, items that were informative for a θ of 0.0 (the average in the 10% prevalence condition) were generally informative for the cut score (1.28), as well. In their review of IRT and clinical measurement, Reise and Waller (2009) interpreted such information functions as reflecting the quasi-trait status of psychopathology constructs. In addition, they stated that "the existence of quasi-traits ... is consequential for many IRT applications" (p. 31). The present study shows another possible consequence of quasi-traits: the potential advantage of CCT over CAT for classification might not be so sizable in clinical psychology.

Given these outcomes, what is the utility of CCT in clinical assessment? For classification, both CCT and CAT can be applied because they had similar outcomes. An advantage of CCT over CAT is that item selection is not adaptive, which is more economical in terms of the complexity of the algorithm. Hence, if classification is the only goal of the assessment, and CCT exhibits similar performance to CAT, the former might be preferred in some applications. An additional benefit of CCT is that it is not restricted to employing IRT models, whereas CAT is. Because the purpose of CCT is classification, instead of a measurement model, a prediction model can be used. Therefore, because not all variables used in clinical assessment reflect the presence of a latent construct, only CCT is an option for computerized assessment of these variables. Recently, such non-IRT based CCT algorithms were applied to a health questionnaire (Finkelman, He, Kim, & Lai, 2011), and a depression inventory (Finkelman, Smits, Kim, & Riley, 2012).

Results from the simulation suggest that using EAP for estimating θ affects classification accuracy. Because of the shrinkage resulting from EAP, in populations with lower prevalence, Type I error rates and specificity decreased as more items were administered, which is, of course, an awkward outcome. Therefore, further investigation is needed to determine if it is more appropriate to use other estimation methods in clinical assessment.

Limitations

Although this study provided useful information on the relative utility of using CCT instead of CAT for classification, there are some limitations that have implications for future research. First, CCT and CAT used an identical stopping rule, i.e., number of items administered. This was necessary to prevent potential differences in outcomes to be ascribed to differences in number of items administered. As a result, it is difficult to relate the current outcomes to comparisons of

CAT and CCT in mastery testing (e.g., Eggen, 1999; Thompson, 2009, 2011). In those studies, the two methods generally showed similar classification results as well, but CCT needed fewer items. Therefore, future research should examine this potential advantage of CCT over CAT in clinical testing. Second, as alluded to earlier, due to its classification bias, EAP should be compared with other estimation methods, such as ML (combined with the stepsize method; Dodd, Koch, & De Ayala, 1989), and weighted ML (Wang & Wang, 2001) to see if these are more appropriate. Third, the current study used only a single item selection method in CCT (maximum information at the cut score), although other measures, such as Kullback-Leibler information (Eggen, 1999), are available. Further research including multiple item selection methods would show whether the current outcomes can be generalized to other methods. Fourth, this study used a single cut score, the value of 1.28 on the standard normal scale, and other cut scores, further or closer to the population mean could also be used; it would be instructive to see if other cut offs give similar results. Fifth, the focus was on one specific diagnostic decision, a healthy versus diseased classification, whereas other clinical classifications (e.g., three categories: low, moderate, and high risk; e.g., Eggen, 1999) used in the field as well should also be studied. Finally, in the current study the training (calibration) set and test (application) set were drawn from the same population distribution. This might be at odds with clinical practice; for example, mental health institutions might choose to employ an item bank developed for the general population (e.g., by PROMIS) for the assessment of their clinical populations (e.g., in ROM). It would be instructive to examine the effect of disease prevalence differences in the calibration and testing samples.

References

Beck, A. T., Steer, R. A., & Carbin, M.G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*(1), 77–100. CrossRef

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. CrossRef

Bock, R. D., & Mislevy, R. J. (1982).Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444. CrossRef

Carlier, I. V. E., Meuldijk, D., Van Vliet, I., Van Fenema, E., Van der Wee, N., & Zitman, F. (2010). Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *Journal of Evaluation in Clinical Practice*. CrossRef

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, *45*, S3–11. CrossRef

Choi, S. W. (2011). Package "lordif" [Computer program]. (Library of the R package)

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of Statistical Software*, *39*(8), 1–30.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7* (3), 249–253. CrossRef

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Dodd, B. G., de Ayala, R., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5–22. CrossRef

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13* (2), 129. CrossRef

Eggen, T. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249. CrossRef

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

Finkelman, M. D., He, Y., Kim, W., & Lai, A. M. (2011). Stochastic curtailment of health questionnaires: A method to reduce respondent burden. *Statistics in Medicine, 30*, 1989–2004. CrossRef

Finkelman, M. D., Smits, N., Kim, W., & Riley, B. (2012). Curtailment and stochastic curtailment to shorten the CES-D. *Applied Psychological Measurement, 36*, 632–658. CrossRef

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*, 2277–2291. CrossRef

Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. B., & Klapp, B. F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research, 18* (1), 23–36. CrossRef

Forkmann, T., Boecker, M., Norra, C., Eberle, N., Kircher, T., Schauerte, P., et al. (2009). Development of an item bank for the assessment of depression in persons with mental illnesses and physical diseases using Rasch analysis. *Rehabilitation Psychology, 54*, 186–197. CrossRef

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., et al. (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry, 4*, Article 13. CrossRef

Gorin, J., Dodd, B., Fitzpatrick, S., & Shieh, Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29* (6), 433–456. CrossRef

Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32* (1), 50–55. CrossRef

Helzer, J. E., Kraemer, H. C., Krueger, R. F., Wittchen, H.-U., Sirovatka, P. J., & Regier, D. A. (Eds.). (2008). *Dimensional approaches in diagnostic classification: Refining the research agenda for DSM-V*. American Psychiatric Publishing, Inc.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage Publications.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*, 91–100. CrossRef

Mellenbergh, G. J., & van der Linden, W. J. (1979). The internal and external optimality of decisions based on tests. *Applied Psychological Measurement, 3* (2), 257–273. CrossRef

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–177. CrossRef

Parshall, C., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer. CrossRef

Pepe, M. (2004). *The statistical evaluation of medical tests for classification and prediction*.

Oxford: Oxford University Press.

Pilkonis, P., Choi, S., Reise, S., Stover, A., Riley, W., Cella, D., et al. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, anxiety, and anger. *Assessment, 18* (3), 263–283. CrossRef

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer program]. Vienna, Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0)

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401. CrossRef

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Review of Clinical Psychology*, *5*, 27–48. CrossRef

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and Item Response Theory analyses. *Journal of Statistical Software*, *17*(5), 1-25.

Rizopoulos, D. (2007). The ltm package: Latent trait models under IRT [Computer program]. (Library of the R package)

Roorda, L. D. (2011). *Prosumerism and computerized adaptive testing.* Paper presented at the Annual Retreat of the Institute for Health and Care Research of the VU University Medical Center, Amsterdam.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded responses. *Psychometrika Monograph Supplement, 17*.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, *188* , 147–155. CrossRef

Smits, N., Zitman, F. G., Cuijpers, P., den Hollander–Gijsman, M. E., & Carlier, I. V. E. (2012). A proof of principle for using adaptive testing in Routine Outcome Monitoring: the efficiency of the Mood and Anxiety Symptoms Questionnaire–Anhedonic Depression CAT. *BMC Medical Research Methodology, 12,* 2. CrossRef

Thissen, D. (1991). *MULTILOG user's guide.* Mooresville, IN: Scientific Software.

Thompson, N. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement* , *69* (5), 778. CrossRef

Thompson, N. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research, & Evaluation*, *16* (4), 2. See http://pareonline.net/pdf/v16n4.pdf

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah NJ: Lawrence Erlbaum Associates.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology*, *57* , 1051–1058. Crossref

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for anxiety (Anxiety-CAT). *Quality of Life Research, 16,* 143–155. CrossRef

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement* , *25* (4), 317-331. CrossRef

Watson, D., & Clark, L. A. (1991). *The Mood and Anxiety Symptoms Questionnaire*. Iowa City: University of Iowa: Department of Psychology.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473. CrossRef

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to edu-

cational problems. *Journal of Educational Measurement*, *21*, 361–375. CrossRef
Weitzman, R. (1982). Sequential testing for selection. *Applied Psychological Measurement* , *6* (3), 337. CrossRef

Supplementary Data

The Supplementary Data file for this article contains the following data:

- Characteristics of the Anxiety items of Pilkonis al. (2011)
- Characteristics of the simulation: Details on the populations and cut scores used and items generated.
- Means and standard deviations for the data in Figures 3 to 7

This file can be requested from the Editor, djweiss@umn.edu.

Author Address

Niels Smits, Department of Clinical Psychology, Faculty of Psychology and Education, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, the Netherlands. Email n.smits@vu.nl.