Polytomous adaptive classification testing:
Effects of item pool size, test termination criterion and number of cutscores

Timo Gnambs and Bernad Batinic

University of Linz

Corresponding author:
Timo Gnambs
Department of Education and Psychology
Johannes Kepler University Linz
Altenberger Strasse 69
4040 Linz, Austria
Email: timo.gnambs@jku.at

Published in *Educational and Psychological Measurement*, doi:10.1177/0013164410393956

This is the pre-peer reviewed version of the article.

The definitive version is available at online.sagepub.com.

Abstract

Computer-adaptive classification tests focus on classifying respondents in different proficiency groups (e.g., for pass/fail decisions). To date, adaptive classification testing has been dominated by research on dichotomous response formats and classifications in two groups. This paper extends this line of research to polytomous classification tests for two- and three-group scenarios (e.g., inferior, mediocre, and superior proficiencies). Results of two simulation experiments with generated and real responses (*N* = 2000) to established personality scales of different length (12, 20 or 29 items) demonstrate that adaptive item presentations significantly reduce the number of items required to make such classification decisions, while maintaining a consistent classification accuracy. Furthermore, the simulations highlight the importance of the selected test termination criterion, which has a significant impact on the average test length.

Keywords: classification, personality, adaptive testing, stopping rule, item pool

Polytomous adaptive classification testing:

Effects of item pool size, test termination criterion and number of cutscores

By administering items in a sequential manner adaptive assessment procedures usually reduce the average length of computerized tests without significantly increasing their measurement error (cf. Fayers, 2007; Forbey, 2007; Reise, Ainsworth, & Haviland, 2005). A variant of such procedures are adaptive classification tests that focus on validly classifying respondents in two or more proficiency groups, for example, to differentiate between students who already master a specific subject and those who do not. So far, adaptive classification testing has been dominated by dichotomous achievement tests (e.g., Hambleton & Xing, 2006; Jodoin, Zenisky, & Hambleton, 2006; Vos & Glas, 2010). Many applied settings, however, administer instruments with polytomous response formats. For example, miscellaneous clinical symptoms (e.g., anxiety or depression) are frequently assessed with polyomous self-report scales. Comparably, occupational aptitude testing that increasingly relys on web-based screening procedures to eliminate obviously unqualified candidates from the recruitment process (cf. Nye, Do, Drasgow, & Fine, 2008) frequently incorporates personality scales with polytomous response formats (Ployhart, Weekley, Holtz, & Kemp, 2003). In both cases adaptive classifcation testing can reduce the burden placed on respondents by administering fewer items while still allowing for an efficient and precise classification of, for example, patients or applicants in groups with inferior vs. superior trait levels. However, up to now, reasearch on *polytomous* adaptive classification testing is scarce and addressed by a few studies only.

The purpose of this study is twofold. First, it extends existing findings concerning test efficiency on dichotomous adaptive classification testing (Finkelman, 2008, 2010, Wouda & Eggen, 2009) to polytomous items. It is demonstrated that frequently used personality scales, which contain as few as 12 items, can be further shortened by adopting a procedure of sequential item presentation. Second, a simulation experiment underpins the importance of choosing an appropriate stopping rule, as the test termination criterion significantly affects the average test length.

Adaptive classification testing

In conventional fixed length tests all examinees are usually presented with the same items. As a consequence, proficient examinees are frequently administered items that are too easy for them, and less proficient examinees receive too many difficult items. These items are not very informative and hardly contribute to an individual´s proficiency estimate. Furthermore, proficient candidates become quickly demotivated, and less proficient candidates frustrated. "Conventional tests are inefficient [...]. Examinee ability-item difficulty mismatches result in wasted testing time and may create fatigue, boredom, or carelessness [...]" (Mead & Drasgow, 1993, p. 450). In computerized adaptive testing (CAT), items are administered examinee-driven in a sequential order. The choice of the next item depends on an examinee´s interim proficiency estimate. In CATs, each examinee is administered different items and only as many items as required to reach a decision in terms of sequential testing (Wald, 1947). As a consequence, CATs usually lead to a significantly reduced test length (Hol, Vorst, & Mellenbergh, 2007, Reise & Henson, 2000). A special form of traditional CATs are computerized adaptive classification tests (CACT), which classify examinees in two or more groups. In contrast to a point estimate of an invididual´s proficiency (as is done in CATs), the goal of CACTs is the accurate classification of examinees in different proficiency groups. As soon as an unambiguous classification decision for an examinee is reached, the testing procedure is stopped. The construction of a CACT requires numerous decisions by the test developer that can affect its accuracy and average test length, such as the choice of proficiency estimator (Yang, Poggio, & Glasnapp, 2006), the item selection algorithm

(Thompson, 2009), or practical constraints like item exposure or content controls (Eggen & Straetmans, 2000). This study focuses on three of such elements that have been proven to be influential for dichotomous CACT (Finkelman, 2008, Thompson, 2007, Wouda & Eggen, 2009): (a) the size of the item pool, (b) the test termination criterion, and (c) the number of classification groups.

Item pool size

As the item selection algorithm has to select matching items for each proficiency level, adaptive tests usually require large item pools. The size of the item pool affects the test efficiency and also the classification accuracy (Lau & Wang, 1999). Larger pools usually contain more informative items around the cutscore and thus increase the overall test quality. Typical pools for dichotomous items usually contain more than 100 items, sometimes even more than 300 items. So far, the only study that explicitly compared the effect of the item pool size on polytomous CACTs reported a better classification accuracy and test efficiency for a larger item pool containing 266 items as opposed to a smaller 90 item pool (Lau & Wang, 1999). For the smaller pool the average test lengh increased about 47%, while simultanously producing up to a third more classification errors. Both item pools in this study, however, were rather large and, thus, are rather unrealistic for clinical or personality assessments in applied settings. In the context of polytomous CATs there are reports that item pools with as few as 25 to 30 items might be sufficient to reach reliable proficiency estimates (Dodd, Ayala, & Koch, 1995, Hol et al., 2007, Wang & Wang, 2001). Reise and Henson (2000) even demonstrated that the facets of the NEO PI-R, which contain eight items each, can be reduced to the half by applying an adaptive item presentation procedure. So far, no study has demonstrated yet if well-established personality scales, which form only short item pools, can benefit comparably from polytomous CACT.

Test termination criterion

A commonly used test termination criterion is the sequential probability ratio test (SPRT; Wald, 1947), a simple likelihood ratio test between two competing hypotheses (such as mastering vs nonmastering for an examinee). In the first step a cutscore θ_c is set on the latent proficiency scale to determine the classification groups. The choice of θ_c is typically based on the empirical distribution of the relevant proficiency in a reference sample or on subjective professional judgements of an expert group (see Cascio, Alexander, & Barret, 1988, for a review). In personnel selection, for example, a cutscore may be established from an incumbant group (e.g., current employees of an organization) and - based on various cost-benefit considerations - set at a value below the mean to exclude obviously candidates who do not possess a required miminum value of an elemental proficiency from the selection process (cf. SIOP, 2003). In the second step an indifference region δ is specified around that cutscore within which examinees cannot be properly classified. For an intermediate proficiency estimate, θ_k, after administering k items, the SPRT then tests the hypothesis H_0: $\theta_k = \theta_c + \delta$ vs. $H_1 = \theta_c - \delta$ by calculating the ratio between two likelihoods, $\lambda_k = L(\theta_c + \delta) / L(\theta_c - \delta)$. An evaluation of the ratio with regard to two decision points, $A = \alpha / (1 - \beta)$ and $B = (1 - \alpha) / \beta$, leads to one of three conclusions (Spray & Reckase, 1996, Wald, 1947): (a) if λ_k is less than or equal to A, then H_1 is accepted, (b) if the ratio is greater than or equal to B, then H_0 is accepted, or (c) if the likelihood ratio falls between A and B, then another item is administered.

The SPRT can be inefficient in cases where it administers another item, even though this observation cannot change an examinee´s classification decision. This is illustrated by Finkelman (2008, example 1): after presenting the *k*th item, the likelihood ratio statistic λ_k might be moderate enough to satisfy neither decision (a) nor (b). However, even if one administered all remaining items, the classification decision would be unlikely to change. For such cases, Finkelman (2008, 2010) has recently proposed stochastically curtailed versions of the SPRT (SCSPRT), which also halt further testing when the probability of a change of the classification decision is rather unlikely or even impossible. The SCSPRT extends the SPRT by specifying two additional stopping rules. Given *k* observations, the SCSPRT also halts further testing and accepts H_1, if the probability of keeping the current classification decision after presentation of the remaining items is higher than a predefined threshold, γ. Conversely, the SCSPRT stops and accepts H_0 if this probability exceeds γ‘. This curtailed version of the SPRT can lead to significantly reduced average test lengths in both two-group (Finkelman, 2008, 2010) and three-group classifications (Wouda & Eggen, 2009), while maintaining a consistent classification accuracy. So far, the SCSPRT has not yet been evaluated in the context of polytomous CACT.

Number of cutscores

The goal of typical CACTs is the classification of examinees in one of two groups, such as failing vs. passing. In some cases, however, multiple classification decisions are of interest; for example, when identifying job applicants with inferior, mediocre, and superior proficiencies. Multiple cutscores put an additional strain on the item pool. The proficiency area, that is, where a test requires the greatest number of items to make a decision, is near the cutscore (Spray & Reckase, 1996). When using more than one cutscore, the item pool has to be large enough to provide a reasonably large number near all cutscores. So far, research on polytomous CACT with multiple cutscores is scarce and supported by a single study only. Thompson (2007) compared various design features of polytomous CACT, including the shape of the item bank, the choice of test termination criterion, item selection procedure and number of cutscores (two vs. three-group classifications). Compared to the other factors studied, the latter resulted in considerably longer tests; three-group classifications required nearly four times as many items as comparable two-group classifications. The study, however, used a rather artifical pool of 60 simulated items. So far, no study has established yet if CACT provides an advantage with regard to test efficiency for personality scales that comprise of only a small set of items for the classification of examinees in more than two groups.

Overview of studies

Two Monte Carlos studies evaluated the impact of polytomous CACTs on established personality tests for two dependent variables: (a) the average test length (ATL) and (b) the percentage of correct classifications (PCC). The experimental design manipulated three independent variables: (a) the length of the scales and thus the available size of the item pool (k_1 = 12, k_2 = 20, and k_3 = 29 items respectively), (b) two test termination rules, SPRT (Spray & Reckase, 1996) and SCSPRT (Finkelman, 2008), and (c) the number of cutscores, resulting in two-group classifications to identify individuals with inferior proficiencies and three-group classifications that distinguish examinees with inferior, mediocre, and superior trait levels. As smaller item pools are likely to have fewer matching items around the cutscore, PCC is assumed to be larger for short scales, particularly for three-group classifications. In line with previous results for dichotomous items (Finkelman, 2008, Wouda & Eggen, 2009), it is assumed that the SCSPRT will outperform SPRT and result in a lower ATL, while maintaining a consistent PCC.

This yielded a completely crossed 3 (item pool size) x 2 (test termination criterion) x 2 (number of cutscores) ANOVA design. All simulations were programmed in R (R Development Core Team, 2009).

Study I

Method

Simulees

Proficiency estimates for the simulees were selected at 21 equidistant points within [–4, 4]. For each estimate, responses to the items of the three scales (see instrument section) according to the graded response model (Samejima, 1969) were simulated for 2000 simulees, thus generating a total sample size of N = 42000.

Instruments and item parameter estimation

The choice of instruments was motivated by their test length, thus resulting in different item pool sizes for the CACT simulations. Hence, we selected three established scales in personality research for our analyses: *Conscientiousness* was assessed with 12 items from the German version of the NEO-FFI (Borkenau & Ostendorf, 1993), *achievement motivation* was operationalized with a 29-item scale[1] by Schuler and Prochaska (2001), and *generalized opinion leadership* was measured with 20 items by Batinic, Gnambs, Appel, and Wiesner (submitted). All items were answered on five-point response scales.

Item parameter estimation

The item parameters for the conscientiousness and achievement motivation scale were estimated from a random sample of N = 1500 prospective students (883 women), who provided a series of cognitive and self-report measures as part of a voluntary study orientation program (see Bergmann, 2008, for details). Item calibration of the generalized opinion leadership scale was conducted with the norm sample (N = 1575, 848 women) presented in Batinic et al. (submitted).

CACTs require known item parameters. The item parameter estimation process involves three steps: 1) checking the assumption of unidimensionality for the item set, 2) identifying the appropriate response model and estimation of the item parameters, and 3) adjuging the fit of the items to the selected response model.

Dimensionality. To assert that an individual´s response probability is a function of one latent trait, the dimensionality of the items sets was analyzed by factor-analyzing the polychoric correlation matrices of the data sets. Parallel analysis and a visual comparison of the second eigenvalues confirmed that the three data sets were truly unidimensional. In addition, the ratios of the first and second eigenvalue were 5.2 (conscientiousness), 3.30 (achievement motivation), and

9.13 (opinion leadership). Furthermore, Reckase (1979) recommended that the dominant first factor accounted for at least 20% of the items variance for acceptable item calibration. For the three items sets, the first factor explained 42% (conscientiousness), 30% (achievement motivation), and 52% (opinion leadership) of the variance of items, indicating an adequate latent factor.

Model selection. To determine the optimal response model for the data set, four polytomous IRT models were fitted to responses of the two samples with *ltm* (Rizopoulos, 2006) and compared on the basis of Schwarz's Bayesian information criterion (BIC; cf. Kang, Cohen, & Sung, 2009): (a) generalized partial credit model (GPCM; Muraki, 1992), (b) GPCM with equal discrimination parameters, (c) graded response model (GRM; Samejima, 1969), and (d) GRM with equal discrimination parameters for all items. On the basis of the BIC criterion, the GRM was deemed the optimal response model for all three scales. The resulting item parameters for the three scales are summarized in table 1.

| Insert table 1 around here |

Item fit. To assess the fit of the items to the response model, the adjusted chi-square statistic to degree of freedom ratio (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001) for item pairs and triplets were inspected. Ratios exceeding 3.5 indicate severe model misfit. The opinion leadership scale displayed moderate misfit, with about 20% of all margins exeeceding the specified threshold. The graded response model, however, is rather robust when the number of deviant items is smaller than the remaining item set (Sinar & Zickar, 2002). The two- and three-way margins for the conscientiousness and achievement motivation scale exhibited no item misfit (less then five percent of pairs and triplets exceeded the specified threshold).

Altogether, the graded response model represented an appropriate response model for the three scales in this study.

CACT simulation

A simulee´s latent proficiency was derived by the weighted maximum likelihood estimator (Warm, 1989). Cutscores were selected at the 25th and 75th percentile of the proficiency distribution in the calibration samples: θc = {-0.48, 0.68} for conscientiousness, θc = {-0.64, 0.65} for opinion leadership, and θc = {-0.61, 0.70} for achievement motivation. The item sequence for a simulee was determined by maximizing Kullback-Leibler information at the cutscore, which selects items based on their ability to discriminate between simulees near the cutscore (cf. Thompson, 2009). For the three-classification case, Kullback-Leibler information was maximized at the cutpoint nearest to the current θ estimate (cf. Wouda & Eggen, 2009). Regarding the indifference region, δ, previous simulations demonstrated that an increase in δ results in significantly longer tests (Eggen, 1999; Lau & Wang, 1999). As our simulations used rather small item pools – 12 items in one case – δ was set at a rather high value of 0.2, the upper limit studied by Eggen (1999), to increase the probabilities of reaching a classification decision without administering the complete scales. Following Finkelman (2008), the error rates α and β were set at .05, resulting in $A = 1 / 19$ and $B = 19$. For the SCSPRT, the probability thresholds γ and γ', which indicate early test termination, were given values of .95. The SPRT for three-group classifications followed the generalized Sobel and Wald (1949) procedure proposed by Eggen and Straetmans (2000).

| Insert figure 1 around here |

Results

The ATLs of the tests (see figure 1) were highly influenced by the test termination criterion. The difference in mean number of items that were administered between the SPRT and the SCSPRT varied along the latent proficiency scale for conscientiousness within Δk = [3.04, 8.26], for opinion leadership within Δk = [1.00, 8.88], and for achievement motivation within Δk = [2.99, 18.20]. The number of cutscores had virtually no effect on these results, with the exception of a comparable increase in ATL around the second cutpoint. Considering the latent proficiencies, the SCSPRT demonstrated to be particularly effective around the cutscores. As the item pools obviously had difficulties in classifying simulees near the cutscores, the traditional termination criterion, SPRT, continued to administer items, even though they were insufficient to improve the classification decisions. Thus, the SPRT administered the complete scales around the cutscores. The SCSPRT, however, prevented the continued item presentation, which resulted in considerably shorter test lengths: ATL near the cutscore reached 5.02 (conscientiousness), 9.92 (opinion leadership), and 9.84 (achievement motivation) for two-group classifications and 5.41, 10.17, and 10.93 for three-group classifications. Hence, the SCSPRT cut the longer scales to about the half or even the third of their original length, while maintaining a comparable classification accuracy as the SPRT.

| Insert figure 2 around here |

The respective classification accuracies of the CACT simulations were only marginally affected by the test termination criterion (see figure 2). For two-group classifications, the classification accuracy reached an average PCC of .96 (conscientiousness), .98 (opinion leadership), and .97 (achievement motivation). For three-group classifications, the PCC was slightly lower, with .93, .96, and .95. The more conservative stopping rule, SCSPRT, had virtually no effect on the PCC and lead for all scales, in both two- and three-group conditions, to an average decrease in PCC of less than .01. Although, generally, PCC seemed quite high, it was highly dependent on the proficiency level. PCC was very high at the more extreme proficiencies, but decreased considerably around the cutscores. For the longer scales, PCC fell to about .74 around the cutscores, and for the short conscientiousness scale, PCC was even as low as .60. However, as seen before, the CACTs using SPRT administered the complete scales near the cutscores. Hence, the decreased PCCs are not a result of the adaptive presentation mode itself, but rather a reflection of the overall quality of the administered scales and their limited measurement precision, that is their increased measurement error, at the cutscores.

Study II

The second study extends the previous results in two important aspects. First, we compared polytomous CACTs to fixed length tests as Reise and Henson (2000) reported that some adaptive tests do not outperform tests with a fixed length when administering a selection of, for example, the most discriminating items to all respondents. Hence, the first aim of the second simulation was a comparison of adaptive classification tests with tests containing a fixed number of items. Second, we replicated the results from study I with empirical responses. Although the use of simulated data is common practice in psychometric research, there is no guarantee that generated pseudo-samples are indeed representative of real data (cf. Micceri, 1989; Steiger, 1977). In practice, item responses can be influenced by numerous factors, for example, the current mood, response sets, or less than perfect fit of the applied item response model. Hence, the second goal of the study was a confirmation of the adpative tests´ advantages with regard to their test length under more naturalistic conditions.

Method

Simulees and participants

The proficiencies for the simulees were randomly drawn from normal distributions with means and standard deviations that were derived from the proficiency distributions estimated in the calibration samples (see previous study): *M* = .097 (*SD* = .860) for conscientiousness, *M* = .003 (*SD* = .957) for opinion leadership, and *M* = .045 (*SD* = .972) for achievement motivation. For each scale, responses to the items according to the graded response model (Samejima, 1969) were simulated for 50000 proficiencies.

The empirical responses stem from two independent samples. The first sample includes *N* = 4110 (2354 women) prospective students with a mean age of *M* = 19.24 (*SD* = 1.14) from Bergmann (2008). From the available data set of the years 2003 to 2009, *N* = 2000 students were randomly selected for the real-data simulation including the conscientiousness and achievement motivation scale. A second sample of *N* = 2000 (1347 women) members of a commercial market research panel (mean age *M* = 28.00, *SD* = 11.13) provided measures of opinion leadership as part of an anonymous web-based survey.

Item parameters and test procedures

The CACT simulations were conducted analogously to study I with the same item parameters. In addition, a series of fixed length tests was created for each scale by ranking the k_j items according to their discrimination indices and subsequently administering the items in this order to all respondents. This resulted in k_j fixed length tests for each scale comprising of $i_j = 1 \dots k_j$ items. Each fixed length test was created by selecting the first i_j administered items of the scale (see Hol et al., 2007, for a similar procedure).

| Insert table 2 around here |

Results

In line with the previous study, the CACT simulations demonstrated a considerable reduction in the number of administered items for most conditions, particularily for two-group classifications with the modified test termination criterion, SCSPRT. Using the traditional stopping rule, SPRT, the simulation administered the full 12-item conscientiousness scale in nearly 60 percent of all cases and resulted in a rather large ATL of 10.24 items (see bar chart in figure 3). Although the longer scales yielded larger savings of about 8.58 (opinion leadership) and 11.52 items (achievement motivation) and, thus, reduced these scales to about 57% of their original length, the complete scales were still administered to about a quarter to a third of all simulees. In contrast, the more conservative test termination criterion, SCSPRT, lead to considerably shorter tests, particularly, for those cases near the cutscores that could not be classified unambigously. Hence, testing was stopped for all simulees before the complete scales were administered (see bar charts in figure 3). As a consequence, the SCSPRT used only about 21% (achievement motivation) to 31% (opinion leadership) of the available item sets. These results for the generated response sets were closely mirrored by the real data simulations, which displayed comparable test reductions (see table 2). Again, the SPRT had difficulties in classifying simulees around the cutscores and thus reduced ATLs only marginally (see table 2). SCSPRT, by contrast, was more parsimonious and lead to ATLs of 4.64 (conscientiousness), 8.56 (opinion leadership), and 8.82 (achievement motivation) for two-group classifications. Hence, the superiority of the modified test termination criterion for polytomous CACTs could be equally demonstrated with simulated and empirical responses.

Although the SCSPRT resulted in significantly shorter tests, it had only a marginal impact on PCCs. The difference in PCCs between SPRT and SCSPRT varied from .01 to .03, with higher differences for the short conscientiousness scale and three-group classifications (see table 2). Furthermore, in terms of their classification accuracies the adaptive procedures even proved superior to comparable fixed-length tests (see bold lines in the lower charts of figure 3). As expected, with an increase in the number of administered items the PCCs of the fixed length tests gradually rise until they reach the maximum classification accuracy of the complete scales (horizontal line in figure 3). However, the adaptive procedures (dashed and dotted lines in figure 3) do not function significantly worse than tests that administer the complete scales. Although they use less items their PCCs are comparable to the full scales´ accuracies. While the SPRT reaches similar PCCs to the complete scales, accuracies for the SCSPRT fall on average one to two points below those of the full scales. No fixed length tests that contained only a subset of items resulted in comparable or even higher PCCs than the adaptive procedures. Hence, simply selecting the most discriminating items to create short fixed length versions of a test did not result in classification accuracies that are comparable to those of the adaptive procedures.

| Insert figure 3 around here |

Discussion

Polytomous CACTs resulted in considerably shorter tests, while maintaining comparable classification accuracies. The presented simulations demonstrated for both two- and three-group decisions that the ATLs of the administered instruments were reduced to, in the best of cases, about 30 to 40 percent of the entire scale. This was confirmed with simulated data sets and also with empirical responses. Difficulties mainly arose for proficiencies near the cutscores. Depending on the overall measurement precision of the instruments, the PCCs increased considerably in this region. This was, however, not specific to the adaptive presentation mode, but was an inherent weakness of the instruments themselves. Classifications that were based on the entire scales resulted in virtually identical error rates as classifications with CACTs using SPRT.

Hence, in the area around the cutscores, increased error rates are to be expected in any case. Adaptive procedures cannot change this. CACTs, however, reach comparable classification results with a considerably reduced number of items on average.

The actual ATLs of CACTs were highly influenced by three factors: (a) the length of the original scale and thus the available size of the item pool, (b) the test termination criterion, and (c) the number of classification groups.

Item pool size

The number of available items that can be administered to an examinee primarily affected the classification accuracies around the cutscores. Shorter scales have fewer items that are informative near the cutscores and thus cannot classify examinees in this region properly. To account for the limited classification precision in this area the adaptive tests administered more items of the scales. In many cases, however, these additional items were little informative and did not improve the classification accuracy. Hence, for the shortest scale in this study, which contained 12 items only, PCC fell about 10 percent points below that of the two longer scales. This was even worse for three-group classifications, where informative items around two cutscores are needed. By contrast, the two longer scales hardly differed in terms of PCC. Even ATLs were comparable for both longer scales, at least when considering the SCSPRT. This mirrors previous results for polytomous CATs (Dodd et al., 1995, Hol et al., 2007), concluding that item pools with as few as 20 to 30 items can be sufficient to reach reliable proficiency estimates. The item pools in this study, however, used items from constructed fixed length scales, which were not developed for adaptive test administrations. These scales usually contain items that are informative across a broad span on the latent proficiency scale. Item pools that have been explicitly constructed for adaptive classification purposes, however, typically include more items around the cutscores and less items at the tails of the proficiency distribution. Hence, for appropriately constructed item pools the effects of CACTs on ATLs is likely to be even more pronounced.

Test termination criterion

The choice of the stopping rule had considerable implications for ATLs. As the CACTs had difficulties in classifying examinees near the cutscores, the traditional termination criterion, SPRT, administered almost all items to these examinees. This, however, did hardly improve the classification accuracy. By contrast, the SCSPRT resulted in early test termination for these cases and lead to considerably shorter ATLs. This mirrors previous results for dichotomous CACTs (Finkelman, 2008, Wouda & Eggen, 2009), Furthermore, the modified test termination criterion was also superior to shortened fixed length tests, that administered the same, i.e. the most discriminating, items to all respondents. Such fixed length tests resulted in significantly more misclassifications than adaptive tests with SCSPRT. Hence, fixed length classifiction tests could not approximate the error rates of adaptive tests with examinee-driven item selection and test termination.

Cutscores

The cutscores mark the regions on the proficiency scale where the most informative items are required. Hence, the region around the cutscores also exhibited the most misclassifications. Depending on the location of the cutscores on the proficiency scale, the cutscores have a huge impact on the overall PCC of a sample. The current simulations specified the cutscores at the 25th and 75th percentile of the proficiency distribution and resulted in PCCs around .92 (achievement motivation) for two-group classifications and a considerably lower PCC of about .83 for three groups. As the error rates in a sample depend on the selected cutscores, cutscores, for example, closer to the median of the proficiency distribution are expected to increase the overall PCC. Furthermore, PCCs are influenced by the quality of the item bank; error rates will increase the fewer items are located near the cutscores. For traditional personality scales, this would most likely be the case for more extreme cutscores set near the tail of the proficiency distributions as these scales usually have less items discriminating well at the extremes. In terms of test lengths, the results are in line with Thompson (2007). Three-group decisions required about 40 to 50 percent more items and thus increased the ATLs of the instruments considerably. However, for CACTs using SCSPRT this translates in about three items only.

Conclusion

Computerized adaptive testing has become increasingly popular during the last two decades (Reise et al., 2005). Consequently, many tests including admission tests, such as the well-established Graduate Management Admission Test (GRE) have been converted to adaptive versions (Rudner, 2010). Adaptive tests are a means to construct shorter measurement instruments without sacrificing measurement precision. Traditionally, a representative sample of items from a long scale are selected to form a short version. As the example of the 12-item conscientiousness scale demonstrated in this paper, such short scales tend to produce more misclassifications near the cutscores than longer scales. The presented adaptive procedure, however, can make use of the entire scale, while administering only as much items to an examinee as needed to reach a classification decision. This leads to a classification accuracy comparable to that of the full scale, with a considerably reduced number of items on average. This is particularly true for the modified test termination criterion proposed by Finkelman (2008), which reduces the average test length to about one-third of the entire scale. Although these results were demonstrated in two independent studies with simulated as well as empiriclal responses, the generalizability of the findings might be affected by the choice of the three personality scales that were used as item banks for the simulations. While the use of empirical derived instead of artificially generated item parameters, rendered rather realistic conditions for the simulations, the study´s results have to be interpreted in light of the specific item locations of the item banks and might not be readily generalizable to other instruments. In practice, CACTs typically operate with explicitly constructed item pools that optimize item information and thus, include more items near the cutscores. Hence, to derive more general conclusions about the effects of the test termination criterion on CACTs future research should experimentally vary characteristics of the item pool (e.g., the number of items near the cutscore). For appropiately designed item banks the effect of the modified test termination criteria, SCSPRT, on ATL may be even more pronounced.

References

Batinic, B., Gnambs, T., Appel, M., & Wiesner, A. (submitted). *Assessment of generalized opinion leadership*.

Bergmann, C. (2008). Beratungsorientierte Diagnostik zur Unterstützung der Studienentscheidung studierwilliger Maturanten [Feedback orientated diagnostics to assist prospective student´s choice of study]. In H. Schuler & B. Hell (Eds.), *Studierendenauswahl und Studienentscheidung* (pp. 67–78). Göttingen: Hogrefe.

Borkenau, P., & Ostendorf, F. (1993). *NEO-FFI - Fünf-Faktoren-Inventar* [Five factor inventory]. Göttingen: Hogrefe.

Cascio, W. F., Alexander, R. A., & Barret, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology, 41*, 1-24.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*, 523.

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

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249–261.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, *60*, 713–734.

Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. *Quality of Life Research*, *16*, 187-194.

Finkelman, M. D. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *33*, 442–463.

Finkelman, M. D. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement*, *34*, 27–45.

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 Computerized Adaptive Version. *Psychological Assessment*, *19*, 14-24.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education*, *19*, 221-239.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2007). Computerized adaptive testing for polytomous motivation items: Administration mode effects and a comparison with short forms. *Applied Psychological Measurement*, *31*, 412–429.

Jodoin, M., Zenisky, A., & Hambleton, R. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*, 203-220.

Kang, T., Cohen, A. S., & Sung, H. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*, 499-518.

Lau, C. A., & Wang, T. (1999, April). *Computerized classification testing under practical constraints with a polytomous model.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449–458.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied*

Psychological Measurement, *16*, 159-176.

Nye, C. D., Do, B., Drasgow, F., & Fine, S. (2008). Two-Step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment*, *16,* 112–120.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56,* 733–752.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved on august 2nd, 2009 from http://www.r-rproject.org

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistic, 4*, 207-230.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, *14,* 95–101.

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

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17,* 1-25.

Rudner, L. M. (2010). *Implementing the Graduate Management Admission Test computerized adaptive test.* In W. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 151-166). New York: Springer.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34,* 100–114.

Schuler, H., & Prochaska, M. (2001). *LMI - Leistungsmotivationsinventar* [Achievement motivation inventory]. Göttingen: Hogrefe.

Sinar, E. F., & Zickar, M. J. (2002). Evaluating the robustness of graded response model and classical test theory parameter estimates to deviant items. *Applied Psychological Measurement, 26*, 181-191.

SIOP. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green: Society for Industrial and Organizational Psychology.

Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics, 20*, 502–522.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*, 405–414.

Steiger, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics, 5*, 1055-1098.

Thompson, N. A. (2007). *A comparison of two methods of polytomous computerized classification testing for multiple cutscores*. Unpublished doctoral dissertation, University of Minnesota.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*, 778–793.

Vos, H. J., & Glas, C. A. W (2010). Testlet-based adaptive mastery testing. In W. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 389-408). New York: Springer.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

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

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Wouda, J. T., & Eggen, T. J. H. M. (2009). *Computerized classification testing in more than two categories using stochastic curtailment* [Measurement and Research Department Reports No 2009-5]. Arnhem: Cito.

Yang, X., Poggio, J. C., & Glasnapp, D. R. (2006). Effects of estimation bias on multiple-category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement*, *66*, 545–564.

Footnotes

[1] Originally, the scale consists of 30 items. However, one item had to be excluded for the present analyses, as it yielded a deficient item discrimination parameter and thus would be rather uninformative for the proficiency estimations.

Table 1.

Summary of item parameter estimates

		a	b_1	b_2	b_3	b_4
	Mean	1.55	-0.75	0.97	1.94	3.65
Conscientiousness	Minimum	0.93	-1.63	-0.26	0.70	2.21
	Maximum	2.73	0.29	2.64	3.56	4.90
	Mean	1.78	-2.43	-0.61	0.64	2.37
Opinion leadership	Minimum	1.16	-3.10	-1.35	0.22	2.07
	Maximum	2.07	-1.64	-0.01	1.24	3.18
	Mean	1.16	-2.30	-1.08	0.25	1.16
Achievement motivation	Minimum	0.76	-4.58	-2.98	-1.64	0.76
	Maximum	1.92	-0.93	0.42	1.75	3.33

Notes. *a* … Discrimination parameter, *b* ... Threshold parameters

Table 2

Average test length and percentage of correct classifications (Simulation 2)

	Simulated responses				Empirical responses	
	SPRT		SCSPRT		SPRT	SCSPRT
	ATL	PCC	ATL	PCC	ATL	ATL
Two-group classification						
Conscientiousness	10.24	.90	3.26	.88	9.98	3.07
Opinion leadership	11.42	.94	6.23	.93	11.54	6.08
Achievement motivation	17.48	.93	5.96	.92	17.72	5.82
Three-group classification						
Conscientiousness	11.63	.80	4.66	.77	11.35	4.32
Opinion leadership	16.15	.88	8.57	.86	16.91	9.02
Achievement motivation	25.34	.85	8.84	.83	24.84	8.37

Notes. N_{sim} = 50000, N_{emp} = 2000, ATL ... Average test length, PCC ... Percentage of correct classifications, SPRT ... Sequential probability ratio test, SCSPRT ... Stochastically curtailed SPRT

Figure Captions

Figure 1. Average test length of CACT simulations (simulation 1): Black lines represent two-group classifications, grey lines represent three-group classifications; solid lines use SPRT, dashed lines use SCSPRT; vertical lines mark the cutscores

Figure 2. Classification accuracy of CACT simulations (simulation 1): Black lines represent two-group classifications, grey lines represent three-group classifications; solid lines use SPRT, dashed lines use SCSPRT; vertical lines mark the cutscores

Figure 3. Distribution of test lenghts and average classification accuracies for adaptive two-group classifications in comparison to fixed length tests (simulation 2)

Figure 1.

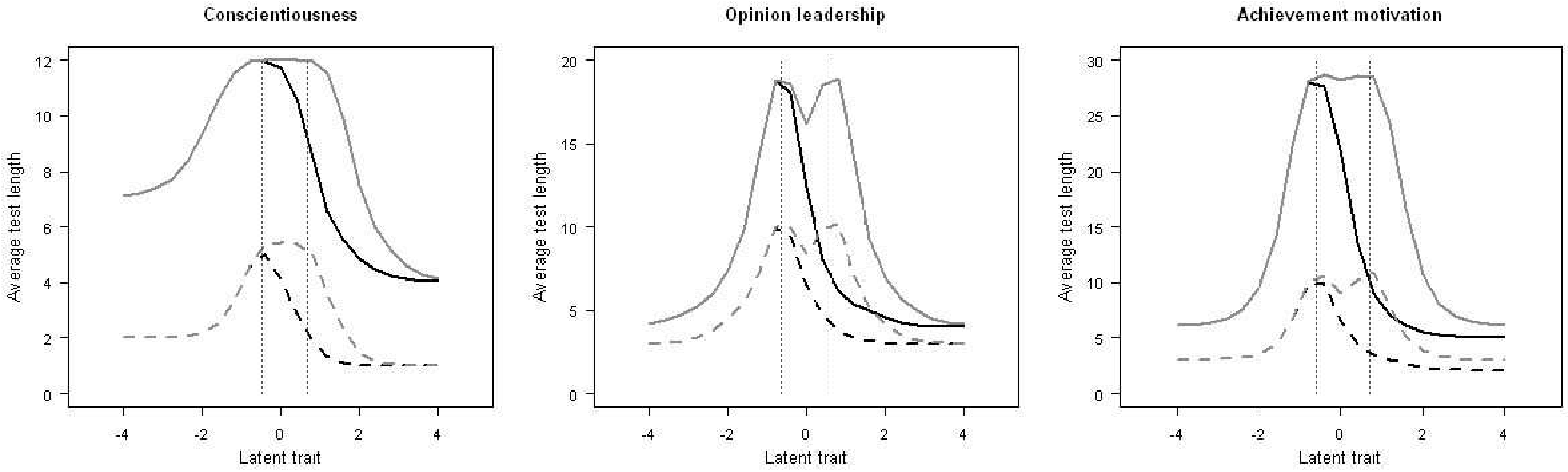

Conscientiousness
Opinion leadership
Achievement motivation
Average test length
Latent trait
-4
-2
0
2
4
0
2
4
6
8
10
12
5
15
20
25
30

Figure 2.

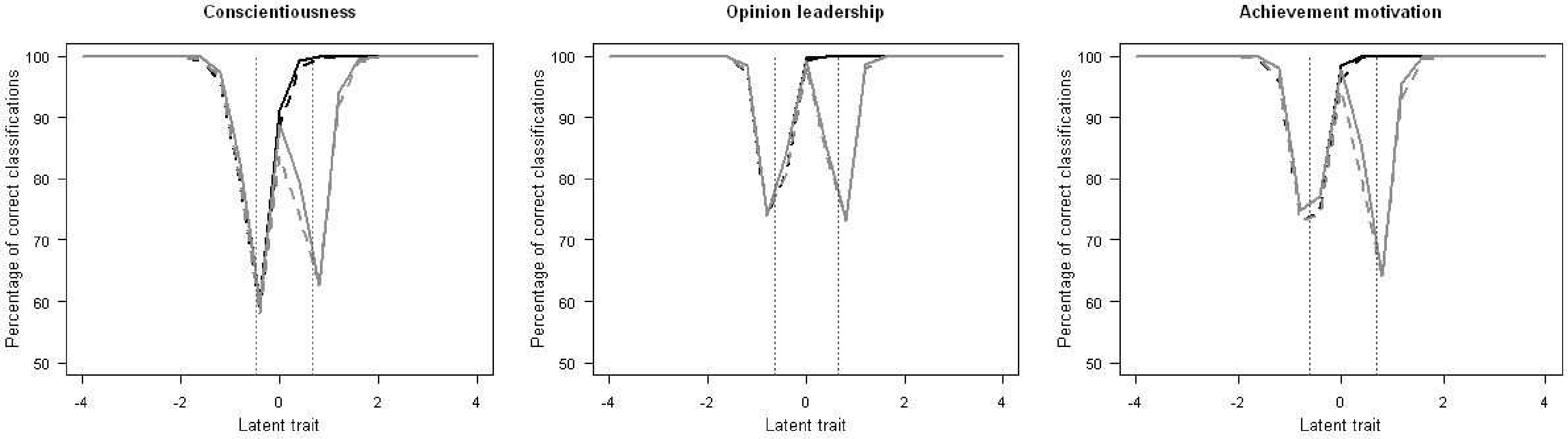
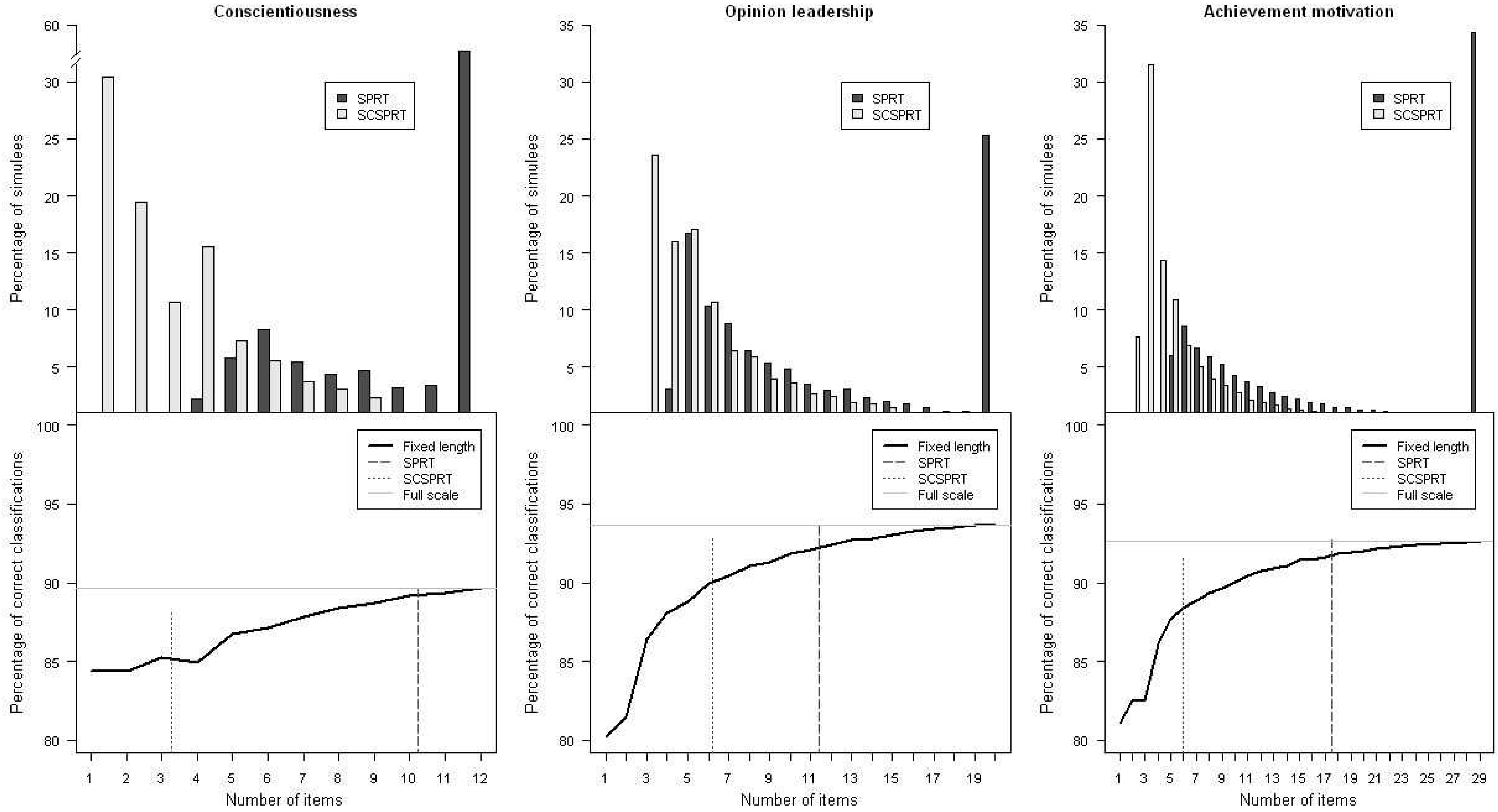

Conscientiousness
Opinion leadership
Achievement motivation
Percentage of correct classifications
100
90
80
70
60
50
-4
-2
0
2
4
Latent trait

Figure 3.



Conscientiousness
Opinion leadership
Achievement motivation
SPRT
SCSPRT
Percentage of simulees
60
35
30
25
20
15
10
5
Fixed length
SPRT
SCSPRT
Full scale
Percentage of correct classifications
100
95
90
85
80
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
23
25
27
29
Number of items