Dear editor,

Thank you for the opportunity to resubmit our paper. The comments of the reviewers were very helpful. Below you find our replies to the authors’ comments. In this file, the original comments are written in a typewriter font, whereas my replies are in normal fonts.

Sincerely,

Niels Smits

**Reviewer 1**

I enjoyed reading the manuscript. It is clearly written and structured and tackles an important question in clinical assessment using adaptive and non-adaptive procedures. Nevertheless, in its current form the manuscript has some issues, mostly minor ones, that should be addressed prior to publication. In other words, the manuscript seems to be in need of some polishing in some places.

Reply: We are very happy the reviewer had mostly minor comments!

1. Abstract: Does the article really “scrutinize” the usefulness of computer adaptive testing in clinical psychology? I am aware that this is also a matter of taste, but I would prefer a less presumptuous “investigates” here.

Reply: In reply to this we removed the word ‘scrutinize’ from the abstract. In the present abstract the word ‘compares’ is being used.

1. Page 4 (Introduction): “Apart from an illustration by Waller and Reise (1989), it [CAT] has never been used in the clinical psychology.” This is not quite true, as there are some applications in the more recent literature (e.g. Fliege, et al., (2009), International Journal of Methods in Psychiatric Research, 18, 23-36). A more accurate review of the literature would be beneficial here.

Reply: Unfortunately, we did not mean [CAT] as filled in by the reviewer in the above sentence, but [CCT]: whereas many clinical studies on CAT exist, only a single one on CCT exists. Our previous text was apparently unclear. Therefore, we explicitly replaced `it’ with `CCT’ as can now be read on p. 4 of the new manuscript. Moreover, the reference of the paper by Fliege et al., 2009 was added to the paper.

1. Page 5 (Introduction): The purpose of the study should be outlined in more detail and more specifically.

Reply: In reply to this we changed the last paragraph of the introduction (p.5 of new manuscript) to be more detailed and specific: ‘*’This study has two purposes. The first is to investigate whether it is more appropriate to use CCT instead of CAT when clinical decision making is an important test goal. The second is 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 are employed on artificially generated clinical data; clinical score distributions are systematically manipulated, and effects on the quality of clinical measurements and decisions are studied.*’’

1. Page 6: The motivation to use the Graded Response Model (GRM) should be discussed briefly.

Reply: In reply to this we provided a motivation for using GRM on p. 6 of the new manuscript: ‘’*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 (PCM, Muraki, 1992) or the Graded Response Model (GRM, Samejima, 1969). We chose the GRM because of its ease of understanding (see, e.g., Mellenbergh, 1995).*’’

1. Page 9-10. Bayesian estimation methods, EAP. I would prefer if the original paper by Bock & Mislevy was referenced for the EAP estimation.

Reply: We changed the reference as requested (see p. 10 of the revised manuscript).

1. Page 12 (Methods): I suggest dividing the description of the criterion variables in two subsection (‘fidelity coefficient’) and (‘external accuracy’). This way, the results sections mirrors the presentation of the methods.

Reply: We changed the headings accordingly. See p. 12 of the revised manuscript.

1. Page 16 (Discussion): The presentation of types I and II errors is quite detailed in the results section of the manuscript. Therefore, I suggest discussing the relevance of these findings in more detail here.

Reply: We followed the suggestions of the reviewer, and the second half of the first paragraph of the Discussion now reads: ““*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.”.*

**Reviewer 2**

In this simulation study, the investigators sought to compare the utility of computerized classification testing (CCT) and computerized adaptive testing (CAT) methods for assessing mental health outcomes. The specific focus of this paper was to compare the ability of CCT and CAT to facilitate clinical diagnosis and decision-making. The application and relative strengths of CCT and CAT in clinical psychological assessment is an important topic which has received little attention in the field. Strengths of this paper include examination of the effects of prevalence rates of a clinical diagnosis on measurement and diagnostic accuracy of CAT and CCT, the use of training and test datasets, and the basis of the simulation on an existing item bank. The investigators employed an innovative approach to simulating the diagnostic criterion variable. The comments below reflect technical aspects of the simulation that may require clarification or elaboration in the text.

Reply: We thank the reviewer for his very kind words and constructive comments!

**Major Comments**

1. On p. 5, the authors describe the procedures used to simulate a diagnostic criterion variable, first by simulating a random variable with a fixed correlation with the latent trait and then dichotomizing this random variable based on a cutoff value of 1.28. The authors selected a value for p (the biserial correlation between the measure and the diagnosis) of 0.60 based on previous research. The simulation procedures described in the text suggests that this correlation held between the latent trait and the diagnostic latent variable (prior to dichotomization) but not between the latent trait and the diagnostic status (following dichotomization). If this is not the case, then this should be clarified in the text. Otherwise, the correlation between the CCT/CAT measure and the dichotomized diagnostic variable will clearly be lower than the desired value of 0.60. Further, the use of different prevalence values may also affect the correlation between the latent trait measured by CAT/CCT and the diagnostic criterion. It would be useful to know what the average correlations were between the full measure (using all 40 items), CAT/CCT measures and the diagnostic criterion in each of the conditions.

Reply: The reviewer is correct that the correlation of 0.60 is the one held between the latent trait and the diagnostic latent variable (prior to dichotomization) but not between the latent trait and the diagnostic status (following dichotomization). Consequently, the text needs no clarification on this point. However, to deal with the question concerned with the correlation between the latent trait estimate and the dichotomous diagnosis variable, we did two things. First, we decided to use these point-biserial correlation as an `external accuracy’ criterion variable (referred to as ‘predictive utility’) in the study; see pp. 12-13,15-16, and p. 17 of the new manuscript. Second, we discussed the effect of dichotomizing quantitative variables on correlations, and gave a reference to Cohen (1983) on pp. 15-16.

1. On p. 6, it is described that parameters for 40 items were simulated for the item bank. What is the rationale for this bank size, especially given that the original item bank consisted of 29 items?

Reply: In reply to this we motivated the choice for using 40 items (p. 6-7 of new manuscript)): ““*The CATs encountered have different item bank sizes: Smits et al. (2011) used the smallest bank, which consists of 20 items, whereas Fliege et al. (2005, 2009) used the largest bank, which consists of 64 items. In the present simulation it was chosen to use an item bank with 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*.”

1. As indicated on p. 8 (second paragraph), item parameters were estimated using the training dataset from each generated data file. Item parameters were estimated in this way "in order to avoid imprecise results in estimation due to sampling error" (p. 8). Estimated item parameters therefore reflected the simulated disease prevalence rate. Such a match between prevalence in an item validation sample and subsequent test samples may or may not reflect clinical practice, as is evident with the development of items banks used for the general population (e.g., PROMIS) and specific clinical populations (e.g., ROM), it would have been interesting to examine the effect of disease prevalence differences in the calibration and testing samples. This point should be acknowledged as a study limitation and perhaps discussed as an avenue for further research.

Reply: The first part of this comment is based upon an unfortunate misunderstanding of what we intended to say. By `sampling error’ we meant the statistical term also often referred to as ‘estimation error’: the error which results from using samples rather than populations (not of using estimates from one population for use in another). **However, the issue the author addresses is a very interesting one!** As suggested by the reviewer we added this issue in the last paragraph of the Discussion (p. 19): ““*Finally, in the current study the training (calibration) set and test (application) set were drawn from the same population distribution. This may be at odds with clinical practice; for example, mental health institutions may choose to employ an item bank developed for the general population (e.g., from 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.*”

1. At the top of p. 11, the authors state: "Because the means of the j distributions differed between prevalence conditions, the CATs started at different locations in terms of the true j scale (about 0.00, 0.76, and 1.28, respectively)." Were these the actual values used as the starting point for the CAT, or were they rescaled in accordance with the calibrated item pool? That is, while these mean values are appropriate for true j, it does not seem that items at these locations would not be optimally targeted to the calibration j scale given the use of marginal maximum likelihood and its assumption of an N(0,1) distribution.

Reply: The starting values of 0, .76, and 1.28 were the values in terms of the original generating latent trait scores. For CCT and CAT in the simulation, the actual value of the starting point is equal to zero, which is, by definition, the average in the train set because of marginal maximum likelihood estimation. To deal with this comment, clarifying the issue, we changed the paragraph in question, which now reads:” *As mentioned in the Population and Calibration Scales section, the average ˆj in each training data set was zero. As a consequence, the item with the highest information at this initial latent trait value was chosen as the first item for all simulees in the test set. Note that in terms of the true (generating) j scale, because the j 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)*”.

1. On p. 16, the statement: "...the potential advantage of CCT over CAT for classification is not so sizable in clinical psychology." To say that the present simulation study findings can generalize across all applications of CAT and CCT in clinical psychology is overly generalizing. Alternative item selection methods (e.g., Kullback-Leibler), cut points more distant from the population mean, other types of clinical decisions, need to be further investigated.

Reply: The reviewer is right that our conclusions were rather bold, and we did two things to deal with this. First, we qualified our conclusion (p. 17-18 ) : ““*The present study shows another* ***possible*** *consequence of quasi-traits: the potential advantage of CCT over CAT for classification* ***may*** *not be so sizable in clinical psychology.*” Second, in the final paragraph of the Discussion, we discuss the limitation of the current study, and all factors mentioned by the reviewer are addressed. See p. 18 of the new manuscript.

1. On p. 16 (second paragraph) the question is posed: " What then, is the utility of CCT in clinical assessment? In addition to the reduced complexity in implementing CCT, I think it is also worth noting that not all clinically-relevant variables reflect the presence of a latent construct. CCT has been used without an accompanying IRT model for purposes of classification or predicting a discrete outcome.

Reply: In reply to this comment, we added the suggestion of the reviewer and provided two examples of non-IRT based CCT methods (p. 18 of new manuscript): “*Because the purpose of CCT is classification, instead of a measurement model, a prediction model may 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, in press).*”

1. The findings presented in this paper seem particularly relevant with respect to proposed changes to version V of the American Psychiatric association's Diagnostic and Statistical Manual, which includes the use of dimensional measures for diagnosis (Helzer, 2008). A statement to this effect could be added to the Introduction or Discussion section of the paper.

Reply: In reply to this we added the reference to the change in DSM in the introduction (p. 3): “*In addition, because the new version of 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.”*

**Minor Comments/Edits:**

1. Perhaps the title could be changed to: "A comparison of computerized classification testing and computerized adaptive testing in Clinical Psychology"

Reply: We changed the title accordingly.

1. On p. 5, the authors state: "In this paper it is studied whether it is appropriate to use CCT instead of regular CAT when clinical decision making is an important test goal." Wouldn't CCT be the de facto choice in this instance? It seems to me that this question should be stated the other way around--whether is appropriate to use CAT instead of CCT when clinical decision making is an important test goal.

Reply: We changed the sentence as suggested (p. 5 of new manuscript).

1. Is there a reason why the quasi-diagnosis variable and its relationship to the primary measure was based on observed correlations between depression and depression diagnosis, whereas the item parameters for the simulated item bank were based on a measure of anxiety? If so, perhaps this could be more explicitly stated.

Reply: To deal with this comment, we stated more explicitly that we used literature both on CATs for both anxiety and depression (p. 6 of the new manuscript): “*To that end, we studied the estimated IRT models of several papers on CATs developed for assessing anxiety and depression (e.g., Fliege et al., 2005; Forkmann et al., 2009; Gardner et al., 2004; Smits et al., 2011).*”

1. On p. 20 in the reference by Reise and Waller (2009), remove the second comma after "Clinical Psychology".

Reply: We removed this error.

1. On p. 21 in the reference by Waller and Reise (1989), remove the second comma after "Psychology" from the journal title.

Reply: We removed this error.

1. For figure 2, it would be helpful to show the values on the x and y axes corresponding to the vertical and horizontal lines (corresponding to mean theta for the horizontal lines and 1.2 for the diagnostic variable).

Reply: We changed the figure accordingly. See Figure 2, p. 29 of the new manuscript.