Consequences of Test Anxiety on Adaptive Versus Fixed Item Testing

Tuulia M. Ortner and Juliane Caspers
Free University, Berlin, Germany

Abstract. We investigated the effects of test anxiety on test performance using computerized adaptive testing (CAT) versus conventional fixed item testing (FIT). We hypothesized that tests containing mainly items with medium probabilities of being solved would have negative effects on test performance for testakers high in test anxiety. A total of 110 students (aged 16 to 20) from a German secondary modern school filled out a short form of the Test Anxiety Inventory (TAI-G; Wacker, Jaunzeme, & Jakszat, 2008) and then were presented with items from the Adaptive Matrices Test (AMT; Hornke, Etzel, & Rettig, 1999) on the computer, either in CAT form or in a fixed item test form with a selection of items arranged in order of increasing item difficulty. Additionally, half of the students were given a short summary of information about the mode of item selection in adaptive testing before working on the CAT. In a moderated regression approach, a significant interaction of test anxiety and test mode was revealed. The effect of test mode on the AMT score was stronger for students with higher scores on test anxiety than for students with lower test anxiety. Furthermore, getting information about CAT led to significantly better results than receiving standard test instructions. Results are discussed with reference to test fairness.

Keywords: adaptive testing, CAT, test anxiety, fairness, matrices

Computerized adaptive testing (CAT) ranks among the most outstanding technical innovations in the domain of psychological assessment in the last century. When employing this technique, also called “tailored adaptive testing,” the testtaker’s individual trait level is iteratively estimated during the testing process. Presented items are selected on the basis of the current trait estimate, which depends on individuals’ responses. In the most informative case of CAT, each person is confronted only with items that have an individual success probability of .50 (Van der Linden & Glas, 2000; Wainer et al., 1990; Weiss, 1982). The use of CAT is based on item response theory (IRT; see Lord, 1980) and item characteristics have to be previously determined in calibration surveys. The advantage of CAT is mostly economic: Studies applying items from achievement tests have shown that measurement errors can be noticeably reduced by using CAT, even when using fewer items as compared to traditional fixed item testing (FIT; Thissen & Mislevy, 2000; Wainer & Eignor, 2000; Weiss & Kingsbury, 1984). As a result, the use of CAT has become common practice in the domain of achievement assessment (see Wainer, 2000), especially when large numbers of tests have to be conducted in a cost-effective manner. This is the case, for example, in military placement testing (Moreno, Wetzel, McBride, & Weiss, 1984; Rauch, Weber, & Wildgrube, 1993), in educational testing (Weiss, 2004; Weiss & Kingsbury, 1984), or in the assessment of health outcomes (e.g., Fayers, 2007; Gibbons et al., 2008).

Whereas the psychometric and technical aspects of CAT have been investigated and described in detail (Chen, Ankenmann, & Chang, 2000; Hau & Chang, 2001; Meijer & Nering, 1999; Van der Linden & Glas, 2000; Wang & Vispoel, 1998), few studies have dealt with CAT’s psychological effects on testtakers (Tondiandel & Quinones, 2000). For the domain of achievement testing, early publications highlighted several positive mechanisms that occur through the application of CAT. Weiss and Betz (1973) expected CAT to be “more ‘fair’ to minority group members in that the range of item difficulties is less likely to result in frustrating or negative experiences” (p. 59). High- as well as low-ability testtakers were supposed to be similarly motivated if they were able to solve about half of the items correctly. This is indeed plausible because – as is true for conventional fixed item tests – testtakers with lower ability are more often in the situation of not being able to answer an item correctly than testtakers with higher ability. Betz (1975) reported higher motivation by employing a pre-version of a tailored CAT strategy (so called stradaptive testing) compared to using fixed item tests. Similar results indicating improvement of test-taking motivation by CAT and even a reduction of average test score differences across ethnic groups were later reported by Pine, Church, Gialluca, and Weiss (1979). However, if different motivation was a result, it did not produce effects on test results; Studies applying adaptive and fixed item versions of the same tests on different samples have shown largely comparable estimates of testtakers’ abilities (Legg & Buhr, 1992; Overton, Harms, Taylor, & Zickar, 1997).
In recent years, more studies have addressed the question of potentially negative effects of CAT on test-takers' motivational and emotional experiences. It was shown that test-takers tend to dislike certain features of adaptive tests, for example, the inability to review or skip items as is possible on traditional paper-pencil tests (Tonidandel & Quinones, 2000; Vispoel, Rocklin, & Wang, 1994). Some studies have again addressed the question of whether working on CAT increases or decreases test-takers' motivation. These more recent studies highlighted an unfamiliar experience in CAT: to be able to solve only about half of the items and being confronted with a higher level of item difficulty more rapidly in the CAT testing process than in FIT. Tonidandel, Quinones, and Adams (2002) experimentally manipulated the difficulties of initial and subsequent items and revealed a significant negative relationship between the average probability of solving items on a computerized adaptive test and self-reported motivation after being tested. Häusler and Sommer (2008) also compared a classic maximum-information item-selection algorithm with two experimental versions employing intermittently easier motivator items. They found a smaller decline in self-confidence when using the version with easier items. In contrast to the earlier positions described above, recent results for an adaptive test assessing concentration showed that test-takers working on the adaptive version reported significantly lower motivation than test-takers who worked on the conventional version of the same test (Frey, Hartig, & Moosbrugger, 2009).

Some recent studies have dealt with the effects of test anxiety on performance in CAT. Test anxiety refers to a set of emotional, physiological, and behavioral responses that accompany persons' concerns about possible negative consequences or failure on an exam, a test, or another evaluative situation (Sieber, O'Neil, & Tobias, 1977). Persons with high test anxiety tend to view test situations in particular as personally threatening (Sarason & Sarason, 1990), and, on the other hand, high levels of trait test-anxiety are generally related to stronger experiences of state anxiety in evaluative situations (Paulman & Kennely, 1984). Test anxiety is generally known to reduce intellectual performance (Hembree, 1988; Seipp, 1991). Harmful effects on test performance have been explained as occurring when highly test-anxious persons divide their attention between self-relevant and task-relevant variables, in contrast to the low test-anxious persons who focus their complete attention on the task (Wine, 1971).

In general, the experience of failure (success) has been shown to result in higher (lower) state test anxiety (Bradshaw & Gaudry, 1968). In line with this finding, higher levels of state anxiety have generally been reported if the self-perceived probability of success on a task is low compared to a situation where success probability is high (Feather, 1965). In view of the method of item presentation, the application of CAT makes the influence of test anxiety likely: Test-anxiety-related behavior is assumed to be evoked if a person believes that his or her intellectual, motivational, or social capabilities are taxed or exceeded by the given demands (Sarason & Sarason, 1990). Comparing CAT to FIT, item difficulty increases more quickly, and, accordingly, success probability also decreases more quickly, and thus might increase state anxiety, especially for individuals possessing high levels of trait test anxiety.

With reference to these considerations, Ponsoda, Olea, Rodriguez, and Revuelta (1999) found lower levels of state anxiety when applying a CAT consisting of easier items than when applying a CAT including individually tailored medium-difficulty items. Tonidandel et al. (2002) tested the hypothesis of whether perceived performance on an adaptive test would mediate the relationship between objective test difficulty and test anxiety. Testing undergraduates, they increased the relevance of the test situation by telling the participants that their scores would be shared with other persons currently taking the test. In their study, anxiety was not related to the number of questions answered correctly. Using items asking for knowledge about the preceding semester, the authors explained their results as the test-takers' performance being subjectively unrelated to any important outcome. They assumed that the test situation failed to provoke any anxiety.

The purpose of the present study was to assess the fairness of CAT when examining persons with high versus low trait test anxiety. In line with Helms (2004), we understand fairness to be the identification of systematic variance attributable to the test-takers' psychological characteristics that are "irrelevant to measurement of the construct of interest" (p. 87) and the removal of this variance from test scores. The term psychological refers to any psychological characteristics such as attitudes, behaviors, and emotional characteristics in this definition (Helms, 2006). We assessed test anxiety in a group of students, then gave a CAT matrices test to half of the sample, and gave the same test as a FIT consisting of items of increasing difficulty to the other half of the sample. We investigated whether test anxiety had differential effects on the test outcome depending on the test mode.

With reference to the study of Tonidandel et al. (2002), the individual relevance of the test situation was expected to be high because we tested young students. Their current situation was one in which they were making choices about focusing on a particular job or course of study. Additionally, they were explicitly told that the test would assess a component of general intelligence, which was explained to be a good predictor of future success in life. We hypothesized comparable results (person parameters) between high-anxious and low-anxious persons when the FIT version was applied, but more of an impact of test anxiety on results for the CAT version (significant test mode × test anxiety interaction).

Second, we investigated the question of whether informing test-takers about the mechanisms and procedures of item selection in adaptive tests would have an impact on test results. In standard instruction texts employed in computerized assessment, there is no information given about the
mode of item selection, order of items, or difficulty. Half of the testtakers confronted with the CAT form were, therefore, informed about procedures and the individualized technique of item selection before taking the test. We assumed that being informed about the CAT technique would reduce negative effects resulting from divided attention on test-irrelevant cognitions and emotions. Thus, testtakers would have more resources at hand for task performance. We, therefore, hypothesized generally better test performance on CAT under the condition in which test information was provided compared to performance on CAT under standard instructions.

Materials and Method

Participants

A total of 110 students (70 young women aged 16 to 20, \( M = 17.58, SD = 1.13 \), and 40 young men aged 16 to 20, \( M = 17.47, SD = 1.25 \)) from a German secondary modern school were tested. They volunteered after being informed that a test of general intelligence and a test of occupational interests would be given. They were informed that individual feedback would be provided. Data sheets and test data were pseudonymized by creating an individual code for each person. Results and a debriefing were provided in a closed envelope for all participants by naming the individual code some weeks after.

Materials

The Adaptive Matrices Test (AMT; Hornke, Küppers, & Etzel, 2000) is a speech-free test assessing reasoning based on figural items. Item design is based on a rationale defining difficulty as including three components: (a) required operations, (b) realization of the required operation in the matrices structure, and (c) the direction of solution (Hornke, 2002). In order to prove its validity, Hornke et al. (2000) showed that the construction rules explained more than 60% of the variance in the difficulty parameter. The AMT fulfills the criteria of the Rasch model (Rasch, 1960). Because of time constraints given for testing at the school, we employed a short form of the AMT as the adaptive version with an average of 15.46 items (13 minimum) out of a total pool of 293 different items, resulting in a medium reliability of .71 for this sample (stopping rule: SEM = 0.63). The second part of the sample was presented a FIT version of the AMT consisting of 16 items with increasing difficulty (item parameters are \(-3.58, -3.19, -2.57, -2.29, -1.71, -1.35, -0.86, -0.54, -0.37, -0.05, 0.22, 0.41, 0.57, 1.49, 1.56, 2.2\)). Reliability was slightly lower using this version (\(\alpha = .66\)).

A short form of the Test Anxiety Inventory – German version (TAI-G; Hodapp, 1991), a questionnaire consisting of 15 items (Wacker et al., 2008), was employed to assess four dimensions of test anxiety: worry (5 items; \(\alpha = .86\)), excitement (4 items; \(\alpha = .85\)), lack of confidence (3 items; \(\alpha = .82\)), and interference (3 items; \(\alpha = .77\)). The TAI-G has been shown to assess more trait-related stable individual differences than situational effects (Keith, Hodapp, Schermelleh-Engel, & Moosbrugger, 2003). The total score was applied to test the main hypothesis. For further analysis, subscales were used.

Procedure

All testings were conducted in groups of two to four persons in a quiet room within the school. Participants were welcomed by a female test administrator (JC) and general information was given about the duration and parts of the testing. Half of the students (using random sampling) were presented the CAT version of the matrices test and the other half worked on the FIT version. To balance test order effects, half of each group filled out the test anxiety questionnaire before the matrices test and half of each group filled it out afterward. Before the test was started, half of the sample who worked on the adaptive version (\(n = 28\)) received information about how adaptive tests work (“The items are chosen dependent on your answers. If you answer items correctly, the difficulty of the following items increases. If you fail, you will be confronted with a choice of easier items.”). The other half of the sample (\(n = 28\)) received standard instructions without information on adaptive testing. Participants were not informed about the existence of different conditions. Including instructions and the occupational interest inventory, each testing took about 50 min. After completion, testtakers were thanked and debriefed. They received their results a few weeks later in a closed envelope signed with their individually created code.

Statistical Analysis

First, referring to the main hypothesis investigating the effect of test anxiety on adaptive versus fixed item testing, we employed a hierarchical moderated regression analysis (Aiken & West, 1991). More specifically, we tested whether test anxiety moderated the effect of test mode on the test results. We entered test anxiety as a predictor in Block 1. Thereafter, we enter test anxiety and test mode as predictors of the AMT score in Block 2. To examine test anxiety as a potential moderator, we then added the interaction of test anxiety and test mode in Block 3. Second, to address the question of the effects of being informed about the principles of CAT, we additionally calculated a 2 (high vs. low test anxiety, dichotomized by median split) × 2 (information) ANOVA with the AMT score as the dependent variable, but only included results from the adaptive test presentation. Third, to further identify the particular form of test anxiety involved in the effect, we calculated a multiple
linear regression with the four subscales of the TAI-G (worry, excitement, lack of confidence, and interference) as predictors and the AMT score from the adaptive condition as the dependent variable.

**Results**

Within the four groups, the following average person parameters were obtained for the AMT (see Table 1): students working on the fixed item version with low test anxiety $M (SD) = -0.98 (0.92)$ versus high test anxiety $M (SD) = -0.88 (0.73)$; students working on the adaptive version with low test anxiety $M (SD) = -0.79 (0.89)$ versus high test anxiety $M (SD) = -1.40 (0.94)$. Employing the moderated regression approach, a statistically significant main effect of test anxiety on AMT scores was revealed in Block 1, $\beta = -0.19$, $p \leq .05$, $\Delta R^2 = .04$. In Block 2, including test anxiety and test mode, neither of the variables reached significance, test anxiety: $\beta = -0.19$, $p = .05$; test mode: $\beta = -0.08$, $p = .40$. In Block 3, a significant interaction of test anxiety and test mode was revealed ($\beta = -0.20$, $p \leq .05$, $\Delta R^2 = .04$, see Table 2 for a summary). As can be seen in Figure 1, the effect of test mode on the AMT score was stronger for students high in test anxiety than for students low in test anxiety. This corroborates the first hypothesis that test anxiety leads to lowered intellectual performance only under the condition of adaptive test presentation.

The 2 (high vs. low test anxiety) × 2 (information) ANOVA addressing effects of being informed about principles of adaptivity within the adaptive version revealed a significant main effect of test anxiety, $F(1, 53) = 7.07$, $p \leq .01$, partial $\eta^2 = .12$, as well as the hypothesized significant main effect of information, $F(1, 53) = 5.32$, $p \leq .05$, partial $\eta^2 = .09$. No significant interaction effect between test anxiety and information was revealed.

In the stepwise multiple linear regression analysis including the AMT score from the adaptive versions as the dependent variable and all four test anxiety subscales as independent variables, a significant effect of the lack of confidence subscale was revealed, $\beta = -0.37$, $t(54) = -2.96$, $p \leq .01$. No other predictor had a significant effect on performance.

**Discussion**

The aim of the present paper was to investigate the influence of test anxiety when assessing intellectual performance by employing either an adaptive form of a matrices test or a fixed item form with ascending item difficulty. Based on previous literature, we hypothesized that test anxiety would have a stronger effect on a person’s performance on the adaptive test than on the fixed item test. Indeed, we found a significant interaction of test mode and test anxiety on test results with test anxiety moderating the relationship between test mode and result. The given result indicates that the particular mode of item presentation might impair persons suffering from test anxiety: Students with higher
test anxiety had similar average results compared to students with lower test anxiety when working on a fixed item matrices test that consisted of items of increasing difficulty. When confronted with an adaptive matrices test, testtakers with high test anxiety had lower test scores compared to persons with low test anxiety. In other words, adaptive item presentation may lead to a bias that produces a disadvantage for testtakers with higher test anxiety.

The given finding indicates a fairness problem of CAT. A general requirement of fair assessment procedures is to provide comparable chances for all examinees to demonstrate acquired knowledge and skills that are relevant to a test’s purpose (Willingham & Cole, 1997). Independent of their ability, highly anxious testtakers have been shown to be at a disadvantage through the application of a particular assessment method (CAT), whereas they had similar chances compared to others when another assessment method was applied (FIT). The phenomenon we found emerged as a medium-sized effect with practical implications: If achievement tests are employed as part of a selection procedure, at least in situations where the outcome is relevant for (younger) testtakers, implementation of CAT may discriminate against highly anxious persons. Our results, therefore, contradict early assumptions generally supporting higher fairness for CAT. However, the superior fairness of CAT was postulated earlier with reference to individuals’ ability levels; persons with high as well as with low abilities were supposed to experience a similar amount of frustration by being able to solve only about 50% of the items. We cannot contribute to the answer to this question, as this would require a different experimental design.

Nevertheless, CAT is still “the king’s road” from a technical point of view for measurement: No other procedure allows for results with such measurement precision in such an economic way. The second main result of our study shows that there is definite hope for CAT: Informing testtakers about the mechanisms and procedures employed in item selection in adaptive testing led to higher scores than presenting standard instructions without that information. A very frugal intervention, that is, simply to inform testtakers about CAT – that items are selected dependent on the given answers, and that more correct items lead to the presentation of more difficult items – may at least serve as an economic way to reduce unfavorable effects of adaptive item presentation on intellectual performance.

Analyzing the effects of different facets of test anxiety on performance during the adaptive testing process revealed a significant contribution by only the Lack of confidence subscale. The relevance of this scale – addressing a cognitive component of the multifaceted construct of test anxiety – is well in line with the assumed increase of an earlier and longer experience of failure when using CAT, when only about 50% of the items can be solved from the beginning: Testtakers with higher confidence might more easily maintain their performance even under more threatening testing conditions, in contrast to persons possessing lower levels of confidence. This result is also well in line with recent data presented by Frey et al. (2009), who investigated motivational effects of an adaptive presentation of a concentration test compared to a standard version. They found that reduced motivation under CAT was mostly a result of lowered subjective estimation of the testtaker’s probability of success. This lowered subjective estimation of success may be even more salient if persons generally lack confidence.

For a deeper understanding, further studies should investigate the mechanisms underlying the given results. For example, they might analyze whether state confidence during or after testing increases as a function of given information about adaptivity. Furthermore, studies could investigate whether the given effects can be influenced if students are well experienced with CAT (e.g., through frequent testings at school). Being attuned to adaptive testing procedures might have effects similar to being informed about the mechanism, in contrast to students who have experienced only FIT procedures.

In general, results indicate that future research on test performance should expand the view from universal effects to effects on a subgroup level, integrating persons’ individual characteristics beyond ethnic or gender groups in research on test bias. Research in the domain of test fairness as well as the future application of psychological assessment would benefit from such a differentiated view.

Acknowledgments

We would like to thank Eva Weißkopf and Jane Thompson for valuable comments on an earlier version of this article. We thank the Schuhfried GmbH, Austria, for allocation of software.

References

Effects of computerized adaptive testing on black and white students (Research Report 79–2). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.


Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Mea-


