Psychological Test and Assessment Modeling, Volume 55, 2013 (1), 92-104

Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit

Regine Asseburg[1] *& Andreas Frey*[2]

Abstract

Usually, it is assumed that achievement tests measure maximum performance. However, test performance is not only associated with ability but also with motivational and emotional aspects of test-taking. These aspects are influenced by individual success probability, which in turn depends on the ratio of individual ability to item difficulty (ability-difficulty fit). The impact of ability-difficulty fit on test-taking motivation and emotion is unknown and rarely considered when interpreting test results.

N = 9,452 ninth-graders in Germany (PISA 2006) completed a mathematics test and a questionnaire on test-taking effort (motivation) and boredom/daydreaming (emotion). Overall, mean item difficulty exceeded individual ability. Ability-difficulty fit was positively linear related with effort and boredom/daydreaming.

The results suggest that low ability students may not show maximum performance in a sequential achievement test. Thus, test score interpretation for this subsample may be invalid. As a solution to this problem the application of computerized adaptive testing is discussed.

Key words: achievement test, test-taking, effort, boredom, performance

[1] *Correspondence concerning this article should be addressed to:* Regine Asseburg, PhD, Leibniz Institute for Science and Mathematics Education at the University of Kiel (IPN), Germany; email: asseburg@ipn.uni-kiel.de

[2] Institute of Educational Science, Department of Research Methods in Education, Friedrich-Schiller-University Jena, Germany

Introduction

Achievement tests usually aim to measure maximum performance as an indicator of an underlying domain-specific ability (Cronbach, 1970). A valid interpretation of test scores requires an assumption that the observed test performance equates to maximum performance (Eklöf, 2010; Wise, 2009; Wolf & Smith, 1995; see also Messick, 1995). But do we really know?

Test-taking motivation and test performance

Test performance primarily depends on individual ability. However, the willingness to show maximum performance is associated with motivational and emotional aspects of test-taking, especially in the case of low-stakes testing (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Cole, Bergin, & Whittaker, 2008; Eklöf, 2008, 2010; Lau, Swerdzewski, Jones, Anderson, & Markle, 2009; O'Neil, Sugrue, & Baker, 1995; Sundre & Kitsantas, 2004; Wise & DeMars, 2010). Test-taking motivation is defined as "willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441). Consequently, test-taking motivation is often operationalized via self-reported effort (Asseburg, 2011). The perceived degree of effort in answering an item successfully, in turn, can be defined as "how much work is involved in a test item or the degree to which an item is 'mentally taxing'" (Wolf, Smith, & Birnbaum, 1995, S. 342).

A meta-analysis by Wise and DeMars (2005) revealed a statistically significant performance difference between motivated and unmotivated groups of test takers in 24 of 25 analyzed effect sizes (mean $g = 0.59$, with g being the mean performance difference between two corresponding groups, divided by the pooled within-group standard deviation; in some studies, motivation was operationalized via self-reported effort, in others via experimentally induced effort). That is, motivated test takers outperform unmotivated test takers by more than one half of a standard deviation. This difference does not seem to be simply due to a positive relationship between effort and the underlying ability construct. According to Nicholls (1984, p. 329), "there is … agreement that a test score does not reveal one's present capacity if one does not apply optimal effort during a test". Especially in low-stakes testing, the variability of effort is quite high (Sundre & Kitsantas, 2004). That is, test performance in low-stakes tests may be notably distorted by motivational effects. In the case of tests being low-stakes for individuals, but high-stakes for higher-level instances (e. g., large scale assessments like PISA or TIMSS), this result is alarming. Accordingly, AERA, APA, and NCME (1999) in their latest version of the Standards For Educational and Psychological Testing propose an interpretation of test results in due consideration of test-taking effort.

From a theoretical point of view, expectancy-value models of achievement motivation provide a good background for a deeper understanding of the relationship between effort and ability. These models predict that strong negative or positive discrepancies between individual ability and test difficulty affect test-taking motivation negatively or positively,

respectively (e. g., Eccles & Wigfield, 2002). Wolf et al. (1995) pointed out that in low-stakes testing students probably do not invest sufficient effort in solving too difficult, mentally taxing items. Students dislike tests with great item demand and feel demotivated which, in turn, results in lower test performance (Sundre & Kitsantas, 2004).

Test-taking emotion and test performance

Regarding emotional aspects, Kleine, Goetz, Pekrun, and Hall (2005) distinguish between positive and negative as well as activating and deactivating emotions before, during and after an achievement test. According to Pekrun's cognitive-motivational model, positive-activating emotions like enjoyment and negative-deactivating emotions like boredom are especially relevant antecedents of test performance (Pekrun, 1992; Pekrun, Goetz, Titz, & Perry, 2002). Boredom, as one example of a negative-deactivating emotion, is characterized by feeling unchallenged and perceiving one's own activities as meaningless (van Tilburg & Igou, 2011). It "involves feeling restless and unchallenged at the same time while thinking that the situation serves no purpose" (van Tilburg & Igou, 2011). Pekrun (2006) assumes that the relative match between task demands and individual ability is important for valuing an activity and, thus, avoiding boredom.

The present investigation

As reviewed above, there is strong evidence that test performance shown in low-stakes testing situations is not solely associated with test taker's ability but also with inter-individual differences in motivational and emotional aspects. Even though motivational and emotional aspects mostly explain a relatively small amount of the variance of test performance, they do play a role in whether or not maximum performance is shown.

Effort and boredom are negatively related (e. g., Acee et al., 2010; van Tilburg & Igou, 2011). Thus, a low level of effort as well as a high level of boredom may impair test performance. The level of both, effort and boredom, in a testing situation can be assumed to be a result of the difference between the test taker's ability and the difficulty of the items the test taker worked on (among other influencing factors like item format or item length). Thus, when confronted with items that are far too easy or far too hard, a test taker is likely to work on the test with reduced effort and to experience increased boredom. Especially in booklet-based testing, when test takers are randomly assigned to booklets with divergent mean item difficulties, systematic design-based differences in effort and/or boredom may arise. Both, reduced effort as well as increased boredom will be most likely associated with a test performance that lies below the individual's maximum performance. This may pose a serious threat to the validity of test score interpretation as maximum performance. Since even low-stakes test situations like PISA or TIMSS are often high-stakes at super-ordinate levels, a valid interpretation of test scores is of upmost importance.

The present investigation thus aims to analyze the relationship between the match of individual ability with the difficulty of the items a test taker worked on (that is, the "abil-

ity-difficulty fit") and effort or boredom, respectively, in a typical large-scale assessment context. Specifically, the following research questions are examined:

1. How are the differences between individual ability and mean difficulty of the items a test taker worked on distributed (ability-difficulty fit)?
2. Which relationship can be observed between the ability-difficulty fit and effort or boredom, respectively?

The results will provide an answer to the question if a large difference between the ability of the test taker and the difficulty of the items he or she worked on may pose a serious threat to the validity of ability estimator interpretation as maximum performance in low-stakes large-scale assessments.

Method

Participants

In the course of PISA 2006, a supplementary grade-based national sample was examined in Germany. The sample consists of N = 9,577 students from 204 schools. It is representative of the population of ninth-graders in Germany (49 % female; M = 15.7 years, SD = 0.57; Prenzel & Blum, 2007). One hundred and twenty-five students without responses on the achievement items were excluded from the analyses.

Procedure

On the second testing day of the PISA 2006 assessment in Germany, the students of the above sample were given a test measuring the attainment of the German national educational standards in mathematics. The students were allowed to use dividers and calculators. Testing time was limited to 120 minutes with a 10 minute break after 60 minutes. Then, another 10 minute break followed, before the students answered an extensive questionnaire including questions regarding test-taking motivation and emotion (35 minutes).

Materials

The *achievement test in mathematics* consisted of 313 items (response format: 49 % multiple choice, 18 % constructed response, 33 % open-ended). Because the item pool was too large to present all items to each student, one of 29 booklets was administered to each student. The items of the pool were assigned to the booklets by use of a balanced incomplete booklet design (e.g., Gonzalez & Rutkowski, 2010). On average, each item was presented to 1,980 students. Unlike the approach in Prenzel and Blum (2007), the item parameters in this study were derived from a one-dimensional scaling using the Rasch model. The one-dimensional scaling was used because a differentiation between

content domains is irrelevant in the present context. Because this study concerns differences between ability and item difficulty on the individual level, an individual estimator for ability was used (weighted likelihood estimate, WLE; Warm, 1989). The reliability of the weighted likelihood estimates for mathematics achievement is .91.

The test-taking motivation and emotion questionnaire included the two dimensions "test-taking effort" (3 items; e. g. "How much effort did you spend on the test?"; 4-point Likert scale with 1 [*none*], 2 [*little*], 3 [*much*], and 4 [*very much*]) and "boredom/daydreaming" (5 items; e. g. "I was bored."; 5-point Likert scale with 1 [*disagree completely*], 2 [*disagree somewhat*], 3 [*undecided*], 4 [*agree somewhat*], and 5 [*agree completely*]). Both scales aimed to measure states, explicitly referring to the current test situation. The items were adapted from the On-Line Motivation Questionnaire by Boekaerts (2002). The same items had also been used in the PISA 2003 assessment (Ramm et al., 2006). The scales show good to reasonable reliabilities (test-taking effort: Cronbach's $\alpha = .85$; boredom/daydreaming: Cronbach's $\alpha = .73$).

Statistical analysis

Descriptive analyses of the difference between mean item difficulty and individual ability (research question 1) are conducted with SPSS 19. The ability-difficulty fit was calculated as the difference between an individual student's WLE for mathematics achievement and the mean difficulty of all items this student had worked on. Items that were not reached individually, at the end of the booklet, were not taken into account, because these items can not influence test-taking motivation, of course. Including these items in the analysis would distort the results. That is, the ability-difficulty fit in the present study differs from the sometimes reported ability-difficulty fit in technical reports of large-scale assessments, which is based on all test items included in the booklets, regardless of which items the test takers actually worked on. The ability-difficulty fit can be built as described above because person parameters (here, WLEs) and item parameters (here, difficulties) are located on the same metric under the Rasch model.

Research question 2, which pertains to the relationships between ability-difficulty fit and effort or boredom, respectively, is analyzed by regression analyses with complex sample structure in Mplus 6.11 (Muthén & Muthén, 1998-2011). All analyses are conducted with weighted data. Statistical significance is defined by $\alpha = .05$.

Results

As expected, test-taking effort and boredom/daydreaming are negatively correlated ($r = -.44$, $p < .01$). The correlation between effort and performance in the mathematics achievement test is $r = .25$ ($p < .01$). Boredom/daydreaming and mathematic performance are correlated with $r = -.22$ ($p < .01$).

Descriptive distribution of the ability-difficulty fit (research question 1)

The first research question compares individual ability and mean difficulty of the items a student worked on. The distribution of the WLEs for mathematic ability is found to have a mean of 0 and a standard deviation of 1.07 (min = -5.21, max = 4.49). The mean item difficulty of all items in the pool is 0.41 with a standard deviation of 1.62 (min = -4.30, max = 6.49).

The ability-difficulty fit, which includes only the difficulties of the individually processed items, has a mean of -0.49 and a standard deviation of 1.07 (min = -5.80, max = 4.05). The mean is negative. That is, on average, mean item difficulty of the processed items exceeds individual ability.

Relationship between the ability-difficulty fit and test-taking effort or boredom/daydreaming, respectively (research question 2)

Does the ability-difficulty fit play a role in how much effort is made and/or boredom is experienced during test-taking? To answer the second research question, test-taking effort and boredom/daydreaming are regressed on ability-difficulty fit, taking into account the hierarchical sample structure (students nested in classes which are in turn nested in schools). Beforehand, descriptive analyses are conducted.

With a mean of 2.82 ($SD = 0.60$; min = 1, max = 4) the average effort spent on working on the test lies above the scale mean (between 2 [*little*] and 3 [*much*]). For boredom/daydreaming, a mean of 2.36 was found ($SD = 0.85$; min = 1, max = 5). Thus, the average reported intensity of boredom and daydreaming falls below the scale mean (between 2 [*disagree somewhat*] and 3 [*undecided*]). Figure 1 shows test-taking effort and boredom/daydreaming as a function of ability-difficulty fit (standardized solution). The relative frequencies of students falling into the categories underlying the figure are plotted, also.

As a general trend, it appears that test taking effort is higher the easier the items are for an individual student, and vice-versa. This relationship appears to break down for very large negative ability-difficulty fit values < -3 or very large positive ability-difficulty fit values ≥ 2.5, and thus for items being much too hard or much too easy compared to the individual ability level. However, since only few students (< 50) fall into these extreme categories of ability-difficulty fit, the results have low precision. The standard errors of the mean values in these extreme categories turn out to be approximately tenth as large as the standard errors of the mean value in the middle category. Thus, the uncertainty of the ability-difficulty fit of < -3 or ≥ 2.5, respectively, is too high for a proper interpretation.

With regard to boredom/daydreaming, a somewhat different picture is observed. Broadly speaking, the easier the items are for an individual student, the less boredom is reported. The relationship does not necessarily seem to be linear, though. Here too, however, very large or very low values of the ability-difficulty fit have very large confidence intervals.

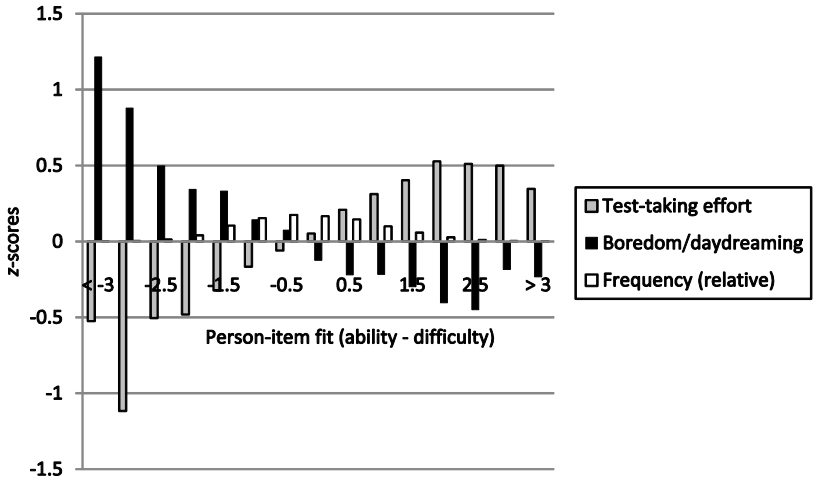

Figure 1:
Test-taking effort and boredom/daydreaming as a function of ability-difficulty fit (N = 9,452).

Regression analyses are conducted to obtain a clearer picture of this relationship and to test the statistical significance of the relationship between ability-difficulty fit and test-taking effort or boredom/daydreaming. Table 1 presents the results obtained from the linear regression analyses.

According to Hu and Bentler (1999), both models fit well to fairly well (CFI ≥ .90, SRMR ≤ .05). The results show a statistically significant positive linear relationship between ability-difficulty fit and effort and a statistically significant negative linear relationship between ability-difficulty fit and boredom.[3] Hence, both test-taking effort and

Table 1:
Simultaneous linear regression analyses (predicting test-taking effort and boredom/daydreaming, respectively, by ability-difficulty fit; N = 9,452)

	B	*SE(B)*	β	χ^2	*df*	CFI	SRMR	BIC
Model 1: predicting effort								
ADF	0.14	0.01	.28**	57.35	2	.99	.013	73,113
Model 2: predicting boredom								
ADF	-0.16	0.01	-.24**	536.08	9	.92	.035	172,271

Note. ADF = ability-difficulty fit. R^2 = .08 for model 1, R^2 = .06 for model 2, (ps < .01). **p < .01.

[3] Booklet specific regression analyses lead to similar results.

boredom/daydreaming depend on how well the average item difficulty matches individual ability. However, the amount of explained variance in both models is rather small with $R^2 = .08$ and $R^2 = .06$, respectively.

Discussion

We examined the relationship between ability-difficulty fit and test-taking effort or boredom/daydreaming, respectively, in a typical large-scale assessment of student achievement. Ability-difficulty fit varies over a broad range and shows a negative mean. For approximately two-thirds of the sample, the average difficulty of the items the students worked on exceeds their individual ability. One-third of the students have an individual ability exceeding the average difficulty of processed items. Thus, on average, the test was too difficult for the students, which means not maxing the test information out given the available testing time.

How do variations in ability-difficulty fit relate to psychological variables connected to test performance? Our analyses reveal that ability-difficulty fit is significantly related with test-taking effort and boredom/daydreaming. Students tend to report greater effort and less boredom/daydreaming the easier the processed items are for them, individually. In this regard, apparently, items cannot be easy enough in order to foster preferable levels of the examined psychological states.

Implications for testing

What conclusions can be drawn from these results? First, it should be remembered that the absolute values of reported effort and boredom/daydreaming are on a satisfying level: The empirical mean of reported effort lies above the scale mean (near the response category “much”), whereas the empirical mean of reported boredom falls below the scale mean (near the response category “disagree somewhat”). That is, on average, students claim to be spending considerable effort on working on the test items and do not feel very bored. At first sight, this is good news for everybody dealing with low stakes large-scale assessments of student achievement, and it supports results from similar studies (e. g., Eklöf, 2008). Nevertheless, it must be assumed that students whose ability estimates fall clearly below the difficulty of the items they worked on do not show maximum performance in low-stakes large-scale assessments. For students with a difference of -2 logits between their ability level and mean item difficulty, the regression analyses predict that the performance is reduced due to a lack of effort by .56 logits and due to boredom/daydreaming by .48 logits. This effect is roughly one half of the standard deviation on the mathematics achievement scale and is therefore of a relevant magnitude. Thus, the results indicate that the test seems to fail to measure “what students know and can do” and seems to merely measure student performance conditional on the difference between their ability level and the difficulty of the processed test items. The interpretation of the derived test scores as maximum performance is, thus, not valid for all students.

These results are particularly interesting with regard to booklet-based testing: Though the ability-difficulty fit is based on the mean difficulty of the items an individual student actually worked on, to a certain degree it also represents differences between the mean booklet difficulties. For students who finish one booklet completely, the difficulty within the ability-difficulty fit is a constant term. Consequently, for these persons, the relationship between ability-difficulty fit and effort or boredom/daydreaming, respectively, simply equates to the relationship between ability and effort or boredom/daydreaming, respectively. Since the students are randomly assigned to the booklets, for those persons who reach all items, the relationship between ability-difficulty fit and effort or boredom/daydreaming represents the differences between the booklet difficulties. Usually, the majority of students reaches all items of a booklet. Thus, in order to avoid systematic differences in effort or boredom/daydreaming, which – in turn – are related to ability, it seems beneficial to create booklets with similar mean item difficulty.

Another possible solution to the issue lies in the application of computerized adaptive testing (CAT) in large-scale assessments. In CAT, the items in a test are individually selected, depending on the test taker's previously shown response pattern. That means, having given a wrong answer causes the selection of an easier item to be presented next. By contrast, after giving a correct answer a more difficult item is selected for presentation (e. g., Wainer, 2000). Usually, in CAT the mean probability of answering an item correctly is at around 50 percent, at least if the Rasch model is used as measurement model. Thus, no test taker is coerced to work on items that are far too easy or far too hard (providing that the ability estimation converges rapidly towards the true ability and that there are sufficient items in the item pool to cover the whole range of test takers' abilities). Furthermore, an adaptive test algorithm with 50 percent success probability for each individual fosters measurement efficiency substantially, because those items provide maximum information given the ability level of the test taker. As a result, a much shorter CAT allows for the same measurement precision as a sequential test – an aspect that is especially valuable in large-scale assessments (see Frey & Seitz, 2011; Frey, Seitz, & Kröhne, 2013). The implementation of CAT in large-scale assessments of student achievement like PISA has already been discussed (OECD, 2006).

Beyond that, the individually adapted item selection in a CAT possibly presents an opportunity to remedy the threat to test fairness and validity of the interpretation of low ability persons' test results as maximum performance in sequential testing. Extreme over- as well as under-challenging could be prevented. With regard to our results, however, we recommend increasing the mean success probability from 50 to 70 percent for example to account for the interconnection of ability-difficulty fit and effort or boredom, respectively. Such a modification of the CAT algorithm can be made without a considerable loss of measurement efficiency (Bergstrom, Lunz, & Gershon, 1992).This could offer an opportunity simultaneously to facilitate high measurement efficiency as well as performance-enhancing test-taking motivation and emotion.

Limitations

The strength of the relationship between effort and test performance varies across countries (Barry et al., 2010). In German samples, this relationship seems to be quite weak, whereas in samples from the United States strong relationships are found. On the one hand, this limits the generalizability of our results with regard to other countries. On the other hand, it points out how important it may be to keep an eye on motivational and emotional aspects of test-taking, especially in an international context, to guarantee the cross-national comparability of test results. An interesting future study would be to compare the relationship of ability-difficulty fit and motivational/emotional variables between countries. This can be done, for example, for international large-scale assessments like PISA or TIMSS. Based on the results presented here, it can be assumed that differences may be observed, even for highly aggregated statistics like the country mean.

Moreover, future studies regarding the relationship between ability-difficulty fit and test-taking motivation may take the item format into account. The decrease of effort during test-taking, for instance, may be stronger for open-ended items than for multiple-choice items. Thus, analyses regarding the item format may give further useful clues for booklet-based test construction.

Finally, since the motivation and emotion questionnaires were administered after the mathematics test, the relationship between ability-difficulty fit and test-taking motivation and emotion may be mediated by perceived performance (see Tonidandel, Quinones, & Adams, 2002). By taking the perceived performance into account in future studies, even clearer results may arise.

Concluding remarks

Observed test performance primarily depends on the underlying ability construct. Psychological aspects like test-taking effort and boredom/daydreaming, however, are also associated with test performance. They can pose a serious threat to test fairness and validity of test score interpretation, especially in the case of a strongly negative ability-difficulty misfit. Therefore, the initiative of AERA, APA, and NCME (1999) to foster the consideration of motivational aspects of test-taking pinpoints an important issue. Design-related differences between booklet difficulties, for example, are connected with systematic differences in test-taking motivation and emotion (see also Frey & Bernhardt, 2012). This should be kept in mind when planning booklet-based testing or when analyzing booklet-based performance data. Hopefully, in the future, psychological aspects of test-taking will assume a more prominent role and will be considered by test developers and test administrators, in addition to psychometric and organizational aspects of test-taking.

References

Acee, T. W., Kim, H., Kim, H. J., Kim, J.-I., Chu, H.-N. R., Kim, M., Cho, Y., Wicker, F. W., & The Boredom Research Group (2010). Academic boredom in under- and over-challenging situations. *Contemporary Educational Psychology, 35*, 17-27.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education (AERA, APA, & NCME). (1999). *Standards for educational and psychological testing.* Washington, D. C.: AERA.

Asseburg, R. (2011). *Leistungsbereitschaft in Testsituationen. Motivation zur Bearbeitung adaptiver und nicht-adaptiver Leistungstests. [Motivation to exert effort in testing situations. Test-taking motivation in adaptive and sequential achievement testing].* Marburg: Tectum.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10,* 342-363.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441-462.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5*, 137-149.

Boekaerts, M. (2002). The On-Line Motivation Questionnaire: A self-report instrument to assess students' context sensitivity. *New Directions in Measures and Methods, 12*, 77-120.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82,* 300-310.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609-624.

Cronbach, L. J. (1970). *Essentials of psychological testing.* New York: Harper & Row.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109-132.

Eklöf, H. (2008). Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments,* 1, 9-21.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17,* 345-356.

Frey, A., & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling*, *54*, 397-417.

Frey, A., & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in PISA. *Educational and Psychological Measurement, 71,* 503-522.

Frey, A., Seitz, N. N., & Kröhne, U. (2013). Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Eds.), *Research on PISA* (pp. 103-120). Dodrecht: Springer.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *3*, 125-156.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Kleine, M., Goetz, T., Pekrun, R., & Hall, N. (2005). The structure of students‘ emotions experienced during a mathematical achievement test. *Zentralblatt für Didaktik der Mathematik, 37,* 221-225.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on General Education Program Assessments. *The Journal of General Education, 58,* 196-217.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User’s Guide.* Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance*. Psychological Review, 91,* 328-346.

OECD (2006). *The OECD Programme for International Student Assessment* [Electronic Version]. Retrieved August 19, 2011, from http://www.pisa.oecd.org/dataoecd/51/27/37474503.pdf.

O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the national assessment of educational progress mathematics performance. *Educational Assessment, 3*, 135-157.

Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology, 41,* 359-376.

Pekrun, R. (2006). The control-value theory of achievement motivations: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18,* 315-341.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students’ self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37,* 91-105.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Upper Saddle River (NJ): Pearson Education.

Prenzel, M., & Blum, W. (2007). *Entwicklung eines Testverfahrens zur Überprüfung der Bildungsstandards in Mathematik für den Mittleren Schulabschluss*.[Test development for the examination of the Educational Standards for the Intermediate School Leaving Certificate in mathematics]. Kiel: IPN.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., et al. (2006). *PISA 2003 Dokumentation der Erhebungsinstrumente* [PISA 2003 scale documentation]. Münster: Waxmann.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*, 6-26.

Tonidandel, S., Quinones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87,* 320-332.

van Tilburg, W. A. P., & Igou, E. R. (2011). On boredom: Lack of challenge and meaning as distinct boredom experiences. *Motivation and Emotion, 35(3).* [Online first version]. Retrieved August 5, 2011 from http://www.springerlink.com/content/y48616w246v28325/.

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 1-21). Mahwah (NJ): Lawrence Erlbaum Associates.

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

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*, 152-166.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15,* 27-41.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351.