Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 8 (2016), No. 2, 10–30

Ramona Lorenz

Does gender make a difference?

Gender-related fairness of high-stakes testing in A-level examinations in English as foreign language in the German state of North Rhine-Westphalia in the context of Educational Governance

Abstract

The shift from decentralized to centralized A-level examinations (Abitur) was implemented in the German school system as a measure of Educational Governance in the last decade. This reform was mainly introduced with the intention of providing higher comparability of school examinations and student achievement as well as increasing fairness in school examinations. It is not known yet if these ambitious aims and functions of the new centralized examination format have been achieved and if fairer assessment can be guaranteed in terms of providing all students with the same opportunities to pass the examinations by allocating fair tests to different student subpopulations e.g., students of different background or gender. The research presented in this article deals with these questions and focuses on gender differences. It investigates gender-specific fairness of the test items in centralized Abitur examinations as high school exit examinations in Germany. The data are drawn from Abitur examinations in English (as a foreign language). Differential item functioning (DIF) analysis reveals that at least some parts of the examinations indicate gender inequality.

Keywords

Educational Governance; High-Stakes testing; Centralized A-level examinations (Abitur); Gender; Differential item functioning (DIF)

Dr. Ramona Lorenz, Institute for Education Research and School Development (IFS), TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany
e-mail: ramona.lorenz@tu-dortmund.de

Geschlechterunterschiede im Zentralabitur?

Fairness von High-Stakes-Tests für Jungen und Mädchen im Fach Englisch in Nordrhein-Westfalen im Kontext der Educational Governance

Zusammenfassung

Die fast flächendeckende Implementation des Zentralabiturs in Deutschland als Maßnahme im Kontext der Neuen Steuerung im Bildungswesen ist eng mit dem Ziel verbunden, die Vergleichbarkeit von Schulabschlüssen und schulischen Leistungen insgesamt zu erhöhen und durch zentrale Prüfungen die Fairness von Leistungsfeststellungen im Sinne der Komparabilitätsfunktion zentraler Prüfungsformate zu sichern. Bisher ist jedoch nicht bekannt und untersucht, ob das Zentralabitur diesen Ansprüchen tatsächlich gerecht wird und die vorgenannten Funktionen so erfüllt, dass eine faire Leistungsmessung für unterschiedliche Schülersubgruppen gegeben ist. Auf der Grundlage von Daten zu differenzierten Schülerergebnissen zum Zentralabitur im Fach Englisch untersucht dieser Beitrag exemplarisch die geschlechtsspezifische Fairness von Abituraufgaben. Eine Differential-Item-Functioning-Analyse (DIF-Analyse) zeigt, dass zumindest ein Teil der eingesetzten Aufgaben auf eine geschlechtsspezifische Ungleichbehandlung durch die Aufgabenstellung hinweist.

Schlagworte

Educational Governance und Neue Steuerung im Bildungswesen; High-Stakes-Tests; Zentralabitur; Geschlechtsspezifische Unterschiede; Differential Item Functioning (DIF)

1. Introduction

In recent years, several reforms have been initiated in the German school system (e.g., Altrichter, Brüsemeister, & Wissinger, 2007; Altrichter & Maag Merki, 2010). One of these reforms is the implementation of centrally executed *Abitur* examinations in almost all German federal states. The *Abitur* is the highest school-leaving qualification in the German school system and at the same time the qualification to study at a German university. As each state has cultural sovereignty and thus supreme legislative and administrative power concerning cultural policy issues, 15 out of the 16 states conduct centralized *Abitur* examinations, with "centralized" in the German school system referring to the state level. While seven states had already a longer tradition of centralized *Abitur* examinations, the remaining eight states changed their assessment culture and implemented centralized *Abitur* examinations in recent years. This centralized examination format is meant to ensure equal opportunities and fairness for all students. North Rhine-Westphalia, as the federal state with the highest population in Germany, is one example where

this centrally executed school leaving examination was implemented as a measure of Educational Governance in 2007. The main research interest presented in this paper deals with these changes in the school system and focuses on data from this state.

In this context, national as well as international findings emphasize the necessity of analyzing centralized tests, which have to be treated as high-stakes tests, if they are applied as school exit examinations, and by this as a measure of Educational Governance. High-stakes tests, in contrast to low-stakes tests, have significant consequences with respect to the results (Featherston, 2011; Heubert & Hauser, 1999; Madaus & Russell, 2009; Nichols, Glass, & Berliner, 2005).

The role of these high-stakes tests is of high relevance to the individual student as well as to Educational Governance (much more important in its centralized form) for which it serves as a means of quality control and quality assurance in terms of providing fairness and equity in an educational system (e.g., Bellenberg, 2008). The current state of research underlines that the above-mentioned function of *Abitur* examinations has not yet been investigated. It is still unknown if and to what extent this examination format contributes to educational fairness. Moreover, it has to be explored if this function can be substantiated and verified for different student subgroups, e.g., for boys and girls in equal measure. So far, only a few researches have focused on the difficulty and fairness of centralized examination formats for different student subpopulations in Germany (e.g., Eickelmann, Kahnert, & Lorenz, 2013; Eickelmann, Kahnert, Lorenz, & Bos, 2011; Kahnert, 2014; Kühn, 2010; Lorenz, 2013; Lorenz, Kahnert, Eickelmann, & Bos, 2011; Maag Merki, 2012; Schräpler & Schmidtke, 2011).

The focus of the presented study lies on the centralized *Abitur* examinations in the school subject of English as a foreign language for German students and conducts research regarding the question of whether the newly implemented centralized examination format achieves the ambitious aim to provide fair examination tasks. Exemplarily, fairness will be examined in this article with respect to gender aspects. The special interest in gender is substantiated in two main aspects: Firstly, it is imperative to mention gender-specific findings regarding items favoring either boys or girls for which evidence has already been generated in large scale assessments (e.g., OECD, 2013). Secondly, there is an evident interest in conducting research on teaching, learning, and assessing foreign language learning in schools under gender aspects (Cheng, Watanabe, & Curtis, 2004; Kunnan, 2000).

2. High-stakes testing in the German school system from the perspective of Educational Governance – Theoretical background and implications for research

Abitur examinations in Germany are to be rated as high-stakes tests for students due to their function in the educational system. Passing the *Abitur* examinations equals the qualification to study at a German university. From the individual's perspective, the *Abitur* examinations and the obtained grades are of fundamental importance for those fields of studies in which the university capacities are limited; the *Abitur* grade point average therefore holds a key role in accessing the most popular study subjects. A thus increased relevance of centralized tests that holds high potential for determining future careers makes a critical assessment of the examination with respect to fairness and equal opportunity imperative.

Nevertheless, the execution of centralized examinations varies to a great extent among the federal states of Germany and also in comparison to other countries. Within the national context, the results of the final *Abitur* examinations count differently towards the grade point average of the *Abitur* certificate. For instance in North Rhine-Westphalia, 21 % of the grade point average is determined by the result of the *Abitur* examinations whereas in Brandenburg, a minimum of 4 % and at the other end of the scale, in Baden-Wuerttemberg, Lower Saxony, and Saarland, 27 % of the grade point average is determined by the centralized *Abitur* examinations (Aktionsrat Bildung, 2011). Thus, the allocation function (Fend, 2008) of this test type also varies a lot within the German school system. All in all, in an international context, the situation in Germany concerning high-stakes testing is characterized by a low level of standardization (Klein, Kühn, van Ackeren, & Block, 2009). Only a small part of a German student's school career, depending to a certain extent on the respective federal state's educational policies, is affected by high-stakes tests.

Due to their importance for both the national and international context, centralized *Abitur* examinations have a significant function within the domain of educational administration and governance: Firstly, they ensure equal opportunities in the access to and comparability of the school-leaving qualification results (e.g., Halbheer & Reusser, 2008). Secondly, they claim to be fair because they demand the same level of achievement from every student (e.g., Kühn, 2010). In terms of fairness, centralized examinations aim to ensure equal treatment and claim not to favor particular students or student groups with respect to the test setup or types of examination tasks and questions.

As an approach for educational research and especially in the light of the recent changes in the German school system, Educational Governance provides a theoretical and practical perspective to describe the complex structure and practices as well as the developments in the school system, also with respect to the anticipated effects (e.g., Creemers, Kyriakides, & Sammons, 2010; Maag Merki, Langer,

& Altrichter, 2014). In this approach, the educational system is understood to be subdivided into several hierarchical levels. Within these different levels, stakeholders are performing according to their own logic and interests; the coordination between these stakeholders on different levels can be specified in terms of Educational Governance (Altrichter & Maag Merki, 2010; Bache & Flinders, 2004; Bevir, 2009; Enderlein, Walti, & Zurn, 2010; Fend, 2011).

The implementation of centralized *Abitur* examinations can be described as a formal top-down innovation (Fend, 2011) initiated by the state administration as a means of Educational Governance which intends to raise comparability of learning results, their assessments and the fairness of measuring and certifying students' achievement in terms of equal requirement for all students in centralized examinations. The question deriving directly from the perspective of Educational Governance is how accurately these goals can be fulfilled and at that, how successful such implementations in schools are.

Taking a more detailed look at the innovation at school level, the shift from decentralized to centralized *Abitur* examinations firstly involves the development of the examinations: Before the changes came into effect, examination tasks were designed by teachers for their single school exit course in the subject they had been teaching their students. They were responsible for the development of the tests for the written examinations as well as the evaluation of the results. In order to ensure at least a demonstrable written certification of quality, the examination tasks were checked by the school administration on a regional level. Tasks were verified or revised by the ministry and their suitability was determined. With regard to the new forms of centralized *Abitur* examinations in North Rhine-Westphalia, the tests are now designed by an expert group from the ministry (instead of single teachers scattered around the state developing their own tasks for their classes). The centralized examinations are carried out by all students at the same time. The evaluation of the examinations is still carried out by the respective course teacher, which leads to the understanding of *Abitur* examinations as a half-standardized examination format. The evaluation of the students' answers is still not conducted centrally, but regulations in the form of predetermined criteria, also developed by experts in advance, are defined (for more detailed information see Lorenz, 2013). These criteria specify a maximum number of points that can be attributed to a correct answer. To increase the objectivity of this procedure, the course teacher evaluates the students' answers before a second teacher rates and re-evaluates the students' performance independently.

Hence it is important to analyze how the innovation of high-stakes centralized *Abitur* examinations set by the administrative level is actually implemented in practice. From the perspective of Educational Governance it is even more important to ask whether the centralized tests can provide fair assessment and increase the comparability of school-leaving examinations.

All in all, the approach of Educational Governance serves as an attempt to understand the changes and innovations in the German school system, especially the shift towards centralized examinations. Although some research exists on

both the national and international level, an evaluation of the aims of Educational Governance such as the increase of accountability and comparability of school examinations has thus far not been conducted. This includes the relationship between item construction in such examinations and the fairness for different types of students. The focus of this study lies on gender differences which have already been pointed towards in the linguistic domain. Previous research on high-stakes testing and gender differences will be presented in the following paragraph.

3. State of the art: The implementation of centralized high-stakes testing

Research concerning the implementation of high-stakes testing comprises different strands. Firstly, effects are observed at the classroom level. Secondly, the school level is considered. As a third strand, the effects on learning and the students' output can be determined. The results of the latter are controversial and do not allow for any consistent view on learning effects (Herman, 2004). Results at classroom and school level are summarized in the following paragraph, after which the importance of test fairness with regard to gender differences is addressed.

3.1 Intended and unintended effects of centralized examinations at the classroom and school level

Research concerning the implementation of centralized examinations has revealed that both at the school and the classroom level, both intended and unintended effects of the implementation of centralized high-stakes tests can be determined (e.g., Altrichter & Maag Merki, 2010; Hargreaves & Fullan, 2012; Koretz, McCaffrey, & Hamilton, 2001). Moreover, those findings are quite heterogeneous.

Nevertheless, a categorization of intended and unintended effects can be extracted. The so-called intended effects refer to the goals set in the context of the implementation of centralized school-leaving examinations. At the classroom level, such intended effects cover the pedagogical use of more demanding learning methods, higher standards of cognitive demand, ensuring that classroom practice meets the requirements of the curriculum, paying more attention to students' individual output and giving a stronger individual support (Amrein & Berliner, 2002; Bishop, 1999; Cheng et al., 2004; Hamilton, Stecher, Russell, Marsh, & Miles, 2008; Maag Merki, 2012; Madaus & Russel, 2009; Nichols et al., 2005). In addition to these findings, a great number of negative consequences has been exposed and could be captioned as unintended effects of centralized examinations. These effects, generally known as teaching-to-the-test effects, include a decline in pedagogical variance, a poorer performance of the teachers, a decreasing variety of topics resulting in a narrowing of the curricula, the aligning of instruction methods with the test format

and an excessive focus on test preparation instead of problem-oriented instruction (Amrein & Berliner, 2002; Bishop, 1999; Eickelmann et al., 2011; Hamilton et al., 2008; Lorenz, 2013; Lorenz, Eickelmann, & Dohe, 2013; Maag Merki, 2012; Madaus & Russel, 2009; Nichols et al., 2005).

3.2 Test construction with regard to fairness for different types of students

Another aspect of high-stakes tests, which has thus far not been thoroughly addressed, refers to a fair assessment of competences for different subgroups. A fair test must have the same level of difficulty for every person with the same ability, regardless of which subgroup they belong to (e.g., Li, Cohen, & Ibarra, 2004). In terms of test theory, a disadvantage due to the affiliation with a certain ethnic, sociocultural or gender-specific group has to be ruled out (Moosbrugger & Kelava, 2007).

As international large scale assessments have not only exposed differences between boys and girls in different domains, but also revealed that a substantial number of test items is not equally difficult for both genders, the test items of the centralized *Abitur* examinations have to be analyzed. About a third of the test items of large scale assessments show different levels of difficulty for boys and girls (Mullis, Martin, Fierros, Goldberg, & Stemler, 2000; Walther, Schwippert, Lankes, & Stubbe, 2008). Therefore, the question of fair high-stakes tests in the *Abitur* examinations for boys and girls will be analyzed in the article at hand. From the perspective of the aims with which the implementation of centralized *Abitur* examinations is associated, test fairness is a central topic to equality in the educational system.

Pedagogical and psychological research has highlighted gender differences in most analyzed domains. In general, this research demonstrates that girls perform better in linguistic domains (e.g., OECD, 2013; Stanat & Kunter, 2003). Differences in the linguistic performances between boys and girls take their root in the early childhood: The linguistic development of girls generally begins earlier (Bornstein, Hahn, & Haynes, 2004) and increases over the course of their school years (Bos, Bonsen, & Gröhlich, 2009; Lehmann, Gänsfuß, & Peek, 1999; Lenzen & Blossfeld, 2009). International large scale assessments such as the *Programme for International Student Assessment* (PISA) or the *Progress in International Reading Literacy Study* (PIRLS) have shown that girls attain higher levels in reading and orthographic skills and that these skills increase faster than it is the case with boys (Cole, 1997; Drechsel & Artelt, 2007; Hornberg, Valtin, Potthoff, Schwippert, & Schulz-Zander, 2007; Kampshoff, 2007; OECD, 2010).

When taking a closer look at foreign language skills, German national studies show that girls tend to outperform their male fellow students in 8th and 9th grade English classes (Hartig & Jude, 2008; Klieme, 2003; Nikolova & Ivanov, 2010). Their advantage can be confirmed in each analyzed subdomain, such as reading

and listening comprehension, writing or grammar. The largest difference in favor of the girls was found in writing skills and text recreation which are the objects of the *Abitur* examination analyzed in this article.

There is no general theory regarding the difference in performance between boys and girls, however, several attempts of explanation exist which include biological factors (Hirnstein & Hausmann, 2010; Stanat & Kunter, 2003; Strüber, 2008), psychosocial factors (Hornberg et al., 2007; Köller & Klieme, 2000), reading habits (Lehmann, 1994) and the teacher's influence (e.g., Ingenkamp, 1989; Jürgens, 2005; Kampshoff, 2007) to explain gender differences (for a more detailed overview refer to Lorenz et al., 2013).

Large scale assessments have not only pointed out gender-related differences in language skills, but also raised awareness of the need for a fair assessment for different subgroups. Secondary analyses have revealed that about one third of the test items show significant differential item functioning (DIF) for boys and girls (Mullis et al., 2000; Walther et al., 2008), which means that these two groups have a different likelihood of solving an item correctly. This finding leads to the question of whether the *Abitur* examinations can be considered fair in terms of differential item functioning for boys and girls in English as a foreign language.

4. Focus of the study and research questions: Researching the fairness of Abitur examinations for boys and girls

Due to the efforts to establish higher levels of fairness in the German school system the difficulties of the centralized examinations for student subpopulations are to be determined. The study focuses on the centralized *Abitur* examinations in the German federal state of North Rhine-Westphalia. The cultural sovereignty of the German federal states, which allows for an individual design of centralized school examination tasks in each state, requires this study to focus on one federal state only. At that, the gender differences observed in the context of the *Abitur* examinations in the school subject English as a foreign language will be examined with particular respect to intensive courses. Intensive courses are characterized by a higher number of lessons per week and are a component of the *Abitur* examinations.

The demonstrated inequality in performance skills in foreign languages between genders and findings from secondary analyses of large scale assessments about varying item difficulties for both groups give rise to the following questions:

- Do boys and girls perform equally in the centralized *Abitur* examinations in the school subject English as a foreign language?
- Are centralized *Abitur* examinations designed in a fair manner for both boys and girls?

It is assumed that there are hardly any differences between the skills of boys and girls because they have both chosen English as an intensive course and consequently for the *Abitur* examination. It could be assumed that their choice is based on their previous performance and marks in this subject. In addition, they had the same lessons and the same preparation for the examinations by their teachers.

5. Data of the current analysis

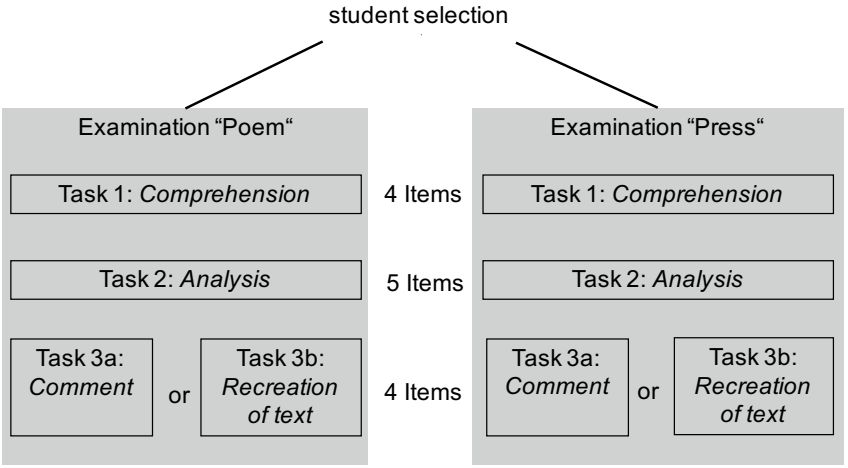
The analysis is based on a sample of 1,136 students who participated in the *Abitur* examinations of English as a foreign language in intensive courses of 2009 in the German federal state of North Rhine-Westphalia. The sample comprises a proportion of 60 % girls and 40 % boys which reflects the distribution in the state's student population. As the study at hand has an explorative character due to covering a relatively unexplored topic within the German school system it focuses on one subject only. Due to an inconsistent data collection for the subjects, the data basis does not allow for a concurrent interdisciplinary analysis.

Out of all 66,201 students taking *Abitur* examinations in North Rhine-Westphalia in 2009 about one third took examinations in English as a foreign language. This is the second most common subject in the examinations with a comparable amount of students only in Mathematics, German and Biology.

The data consists of the teachers' documents, which provide the evaluation results of the students. The evaluation is not executed centrally; it is conducted by the teachers and is based on predetermined evaluation criteria which are developed centrally together with the examination tasks. For each criterion a maximum number of points which is achievable by the students is defined. These criteria are the items of the following analysis.

In the subject English as a foreign language the students can choose between two examination tasks. The presented results of this paper are therefore related to the two *Abitur* examination tasks between which the students could choose in 2009 in North Rhine-Westphalia. The first examination task was based on a poem about the civilization process in the USA. The second examination task deals with a comment from the international press. It is a criticism of the social and political development of Great Britain and the passive behavior of the citizens.

Figure 1: General set-up of Abitur examinations (English as a foreign language in North Rhine-Westphalia, Germany)



Each examination task consists of three tasks. Tasks 1 and 2 have to be answered by all students who have chosen this examination task. For Task 3 students can choose between writing a commentary on the text and doing a recreation of the text, for example by retelling it from another point of view. Figure 1 illustrates the general set-up of the analyzed examinations.

Task 1 comprises four predetermined criteria, which are the analyzed items. Task 2 contains five items and Task 3a and 3b contain a further four items. This distribution of items is the same for both examinations. In total, this analysis therefore comprises 13 items per student. As there are two test manuals and as there are multiple tasks to choose from within each test, the data are treated as a multi-matrix design.

6. Methods

Descriptive analyses are considered to answer the first research question investigating the performance of boys and girls. The amount of points gained by boys and girls in the *Abitur* examinations – with regard to the single tasks – indicates differences between the genders in their English as a foreign language performance.

To answer the second research question whether the *Abitur* exams are fair for boys and girls, DIF analyses were conducted (Camilli & Shepard, 1994; Dorans & Holland, 1993; Holland & Wainer, 1993). The analysis of DIF provides information on the relative difficulty of an item for two or more groups.

> In IRT terms, a scale item displays DIF if examinees with the same latent-trait level have different probabilities of endorsing an item. In other words, in IRT terms, a personality or attitude item is biased if the IRCs [Item Response Curves] are not the same across two groups of examinees.
> (Embretson & Reise, 2000, p. 319)

DIF appears when people with comparable abilities from different groups have different probabilities to answer the item correctly. One group is defined as a "reference group", and the other groups as "focal groups". The basic idea is to estimate an Item Response Theory (IRT) model separately for all groups and then to transform the parameters to the metric of the reference group. This transformation allows checking for actual differences between the groups. Any remaining differences indicate that "equally able (or proficient) individuals, from different groups, do not have equal probabilities of answering the item correctly" (Angoff, 1993, p. 4). This interpretation is based on the assumption that the test as a whole only measures the relevant latent trait (cf. Koretz & McCaffrey, 2005).

The DIF analysis for this article is considered in the 1-Parameter model and is conducted using the statistical program R 2.13.0 (package difR; Venables, Smith, & R Development Core Team, 2011). DIF detection is performed by using Lord's chi-squared method (Lord, 1980). The level of significance for the DIF analyses is .05 and a generalized linear mixed model is conducted. The DIF effect is computed by multiplying the difference of the item difficulties for both groups by -2.35 (Penfield & Camilli, 2007). The outputs include estimates of the item difficulty for each item in each group based on the metric of the reference group. Girls were chosen as the reference group and boys as the focal group. This decision merely fixes the scale and has no further implications for the interpretation of the results.

7. Research results of the current study: Focus on gender-specific differences in Abitur examinations

7.1 Performance of boys and girls in the examination tasks

First of all it is worth mentioning that about 75 % of the students chose the examination task dealing with the comment from the international press. The allocation of the genders shows that significantly more girls chose the first examination task "Poem" while more boys chose the second one, i.e. "Press". Concerning the choice in Task 3 it can be summarized that more students chose to write a commentary (Task 3a) rather than a recreation of the text (Task 3b). But within this selection no gender-specific patterns can be detected.

With regards to the scores in the examinations it can be shown that girls have a significant but minor advantage in the examination "Press" while boys and girls get

the same amount of points in the examination "Poem" (see Table 1). All in all the difference is not very high.

Table 1: Performances of boys and girls in the Abitur examinations

	Task	Mean				Difference	Eta-squared
		Female (*N*)		Male (*N*)			
Examination "Poem"	1	9.38	(217)	9.42	(64)	-.04*	.000
	2	12.65	(217)	12.30	(64)	.35*	.001
	3a	9.79	(118)	9.55	(40)	.24	.000
	3b	12.55	(99)	13.25	(24)	-.70	.003
Examination "Press"	1	10.92	(471)	10.77	(384)	.15	.000
	2	15.83	(471)	14.29	(384)	1.54*	.020
	3a	10.37	(324)	9.95	(288)	.42*	.001
	3b	13.00	(147)	12.38	(96)	.62	.004

* $p < .05$

Considering the gender-specific scores for each task, it is obvious that girls gain more points. There is a significant difference favoring girls in one task in the first examination "Poem" and a significant difference in two tasks in the second examination "Press". Thus, a slight difference in the performance of boys and girls can be revealed.

7.2 Fair examinations for boys and girls?!

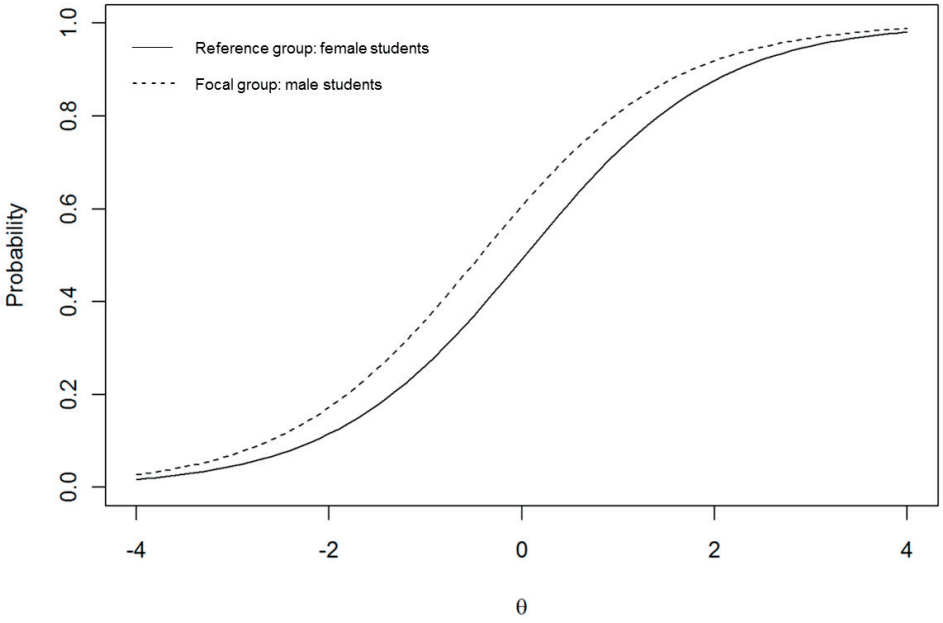
To determine the difficulty of the tasks for boys and for girls, it is necessary to have a closer look at the individual items of the questions. It is observed how many items are easier for girls and for boys. With the DIF analysis the items which favor one gender can be identified.

The results show no DIF item in the first examination "Poem". The test can be described as fair for boys and girls to the effect that no item can be solved significantly better by one group.

The second examination "Press" on the other hand contains three DIF items. To examine whether the items are easier for boys or girls, the item characteristic curves (ICC) are considered. If there is a significant difference between boys and girls, the item shows a different functioning for the two groups. Figure 2 illustrates that the first DIF item (Item 2) is easier for boys: With a difficulty of zero logits,

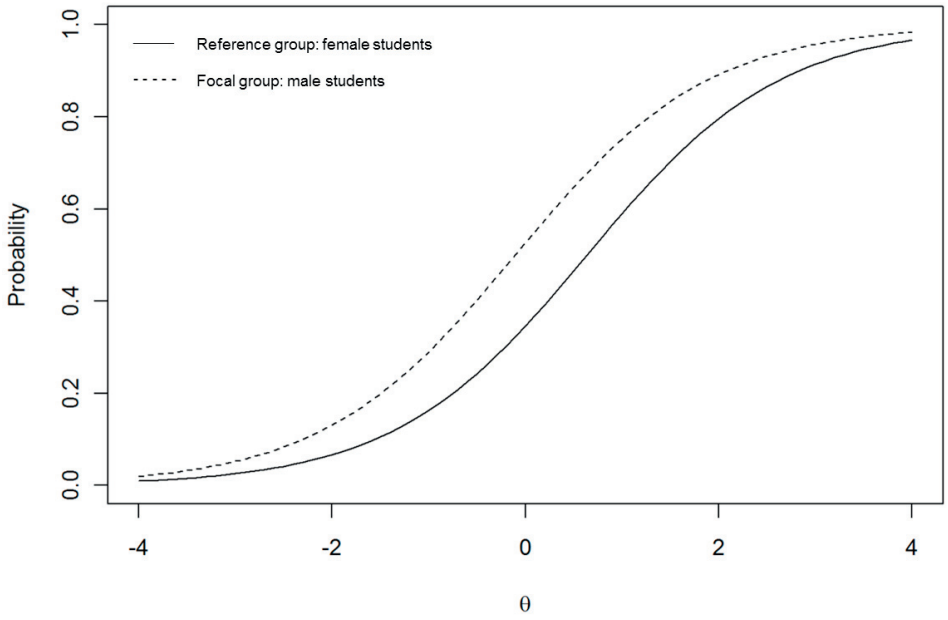
the probability that girls solve the item correctly is about 50 %, the boys on the other hand show a probability of 60 %.

Figure 2: ICC of Task 1, Item 2 (examination "Press") – terrorism and surveillance



Examining the item content more closely, the item task can be summarized as follows: The students were supposed to describe the author's point of view that terrorism and its consequences have caused intensified surveillance of the citizens. The students have to write a summary of the text to show that they have understood the author's argumentation. This topic may be more easily approachable for boys and thus meets the findings of previous studies about gender-specific reading habits and role models, which state gender specific interests in topics as for example boys are more interested in the police or in technology (e.g., Lehmann, 1994). Thus, terrorism which is related to the work of the police and possibly also the technical aspect of surveillance might be more appealing for boys.

Figure 3: ICC of Task 1, Item 3 (examination "Press") – powers of politicians and the police



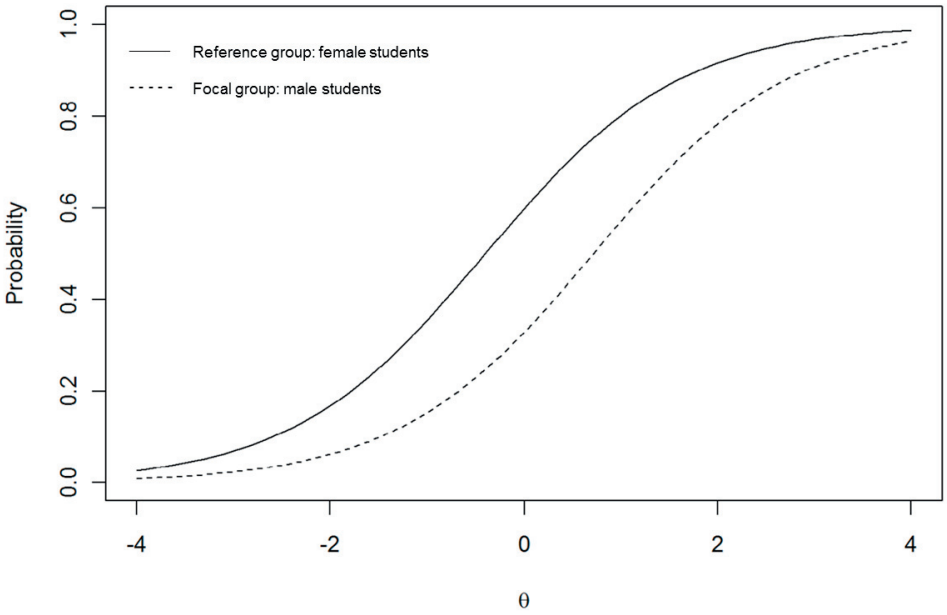
The next DIF item favors boys, too (Figure 3). Their probability of solving the item correctly is about 50 % whereas the girls' chances are about 35 % at an item difficulty of zero logits.

The content of the item can be described as follows: The students have to refer to the increasing powers of politicians and the police. In their summary of the author's argumentation they have to include this aspect which merely is a comprehension task. Possible interpretations of DIF again meet the assumptions on gender specific reading interests which makes the topic of police and possibly also the power of politicians more appealing and interesting for boys. Thus boys might write a more detailed summary or are more likely to highlight some aspects which lead to a better evaluation of their answers.

For Item 9 the relationship is the opposite: the item favors girls (Figure 4). The difference in probability of solving the item correctly is higher than it was for the first two items. The girls' chances are about 60 %, the boys' chances amount to about 35 % at an item difficulty of zero logits.

The students had to refer to the author's attempts to convince the readers of his opinion by analyzing the grammar of the text. As national and international research has already pointed out, this is a girls' domain (see Section 3.2). Thus it seems that also in the *Abitur* examinations girls show higher abilities in analyzing the text and gain more points for their answers.

Figure 4: ICC of Task 2, Item 9 (examination "Press") – analysis of the author's grammar



All in all, the conducted analyses point out that the two considered *Abitur* examinations of English as a foreign language contain a relatively small number of DIF items in comparison to international large scale assessments. While these assessments include up to 30 % DIF items (Mullis et al., 2000; Walther et al., 2008), one examination shows no gender-specific DIF item and the second contains 23 % (3 out of 13 items). Although the girls obtain higher total scores, there are two items favoring the boys and only one favoring the girls. In other words, without the two items favoring the boys, the difference between genders would be even higher. Taking into consideration that there is one item favoring girls and two items favoring boys, it cannot be proclaimed that one gender is clearly favored by the items in terms of solving a large number of items significantly better.

8. Conclusion and discussion

The results of this research show important indications with regard to the fairness of centrally executed *Abitur* examinations and their intended functions. On the one hand it can be appraised to what extent centralized examinations make a contribution to fairness in the educational system. On the other hand the awareness regarding this topic and its significance for different student subpopulations is raised. However, it must be taken into consideration that the data are based on the examinations of only one German federal state and because of the different examination procedures the results may neither be generalized for the whole of Germany

nor for other subjects. Besides, there could be a bias in the sample. It could not be controlled whether all student characteristics (for example age distribution) are represented appropriately.

With regards to the first research question concerning the performance of boys and girls in the *Abitur* examinations, only slight differences could be revealed. These have to be validated by further research comprising more examination tasks. The significant differences found between boys and girls speak in favor of a better performance by the girls in three cases and by the boys in one case.

As to the second research question investigating the fairness of *Abitur* examinations for boys and girls, a positive conclusion can be drawn. The first examination contains no DIF item and thus can be seen as fair to both genders. The second examination has to be estimated: There are only a few items where DIF can be detected and the effect favors not only one group: One item is easier for girls while two evaluation criteria favor the boys. This means that without the differential item functioning in these items, the difference between genders would be even higher. Nevertheless, the percentage of DIF items is lower than it is in international large scale assessments which were constructed according to test theory and tested in advance. Thus, it cannot be concluded that unfair items regarding boys and girls are included in the *Abitur* examinations in English as a foreign language.

The results can be interpreted in light of the test construction and the process of evaluating the answers. The examinations are constructed in a multilevel process involving experts from different subjects and professional disciplines to revise the examination tasks. Finally, they are piloted by a group of teachers taking the examinations from the student perspective. With regards to the results of the study at hand, the assumption may be made that this complex and profound process of test construction leads to fair examinations. Another point of interest is the evaluation of the students' answers by providing evaluation criteria for each task. These criteria contain a slight scope for the teachers by only specifying a maximum amount of points. Nevertheless, the presented results give no evidence for an unfair evaluation considering boys and girls.

Concerning gender-specific domains in the *Abitur* examinations, the presented study confirms findings of previous research. According to the findings concerning reading habits (e.g., Lehmann, 1994) a gender-specific role behavior can be detected within the students' selection of their examination. Firstly, the selection of the examination is congruent with stereotypical choices: More girls chose the examination which is based on a poem and more boys chose the one based on an article from the international press. Secondly, the DIF effect of the items can be linked to gender-specific domains. Girls on the one hand seem to be more sensitive to linguistic aspects in the text and solve the task of analyzing the grammar significantly better than boys. Boys on the other hand perform significantly better than girls when they have to summarize a given text which covers a typically male domain. These first tentative conclusions must be reconsidered from a didactical perspective to generate constructive suggestions for teaching.

In terms of Educational Governance the centralized *Abitur* examination seems to be implemented at school level as it is supposed to be from the administrative point of view – at least considering the gender-specific fairness of this high-stakes test. The student subgroups of boys and girls have approximately equal conditions in the examinations and the transition to tertiary education is not affected by the test. Thus it could be shown (for one subject) that the centralization of the *Abitur* has a tendency to provide more equal opportunities to pass the *Abitur* compared to the non-centralized examination format.

Furthermore, the test is fair for boys and girls in different schools and districts in North Rhine-Westphalia which is an important result in regard to Educational Governance. Nevertheless, the examinations are designed centrally within only one state. Thus the aspect of equity and equal opportunities cannot be ensured across states. On top of this, the weight of this centralized examination in the final mark varies across states from four up to 27 % which can be seen as a rather low or medium level of standardization in comparison to other countries.

With respect to implications for administration and test construction, the study reveals that the adaptation of test theory – as a standard procedure for large scale assessment tests – to the investigated examination format "centralized *Abitur* examinations" has the advantage that the tasks can also be described in scientific terms and thus the process of test construction can be revised and improved.

Methodological implications direct the attention to a triangulation with additional qualitative methods for the investigation of DIF items. In order to test the hypothesis of gender-specific topics within the DIF items, and to reveal potential additional aspects for an explanation of DIF, some qualitative research interviewing students or teachers should be attached.

Further research could involve the evaluation process of the *Abitur* examinations. A reason for the DIF items may be the teachers' evaluation of the students' answers. As the evaluation criteria provide a range of points to mark the students' answers, the teachers' subjective opinion about the students might influence their judgment. This hypothesis could be proved by a blind re-evaluation, with external teachers who do not know the students' gender.

Another aspect for subsequent research could be additional student characteristics and background information as explanatory factors for differences in content and linguistic performance. In addition to this, further school subjects and examinations from additional years should be analyzed to provide a thoroughly researched answer to the question of fair *Abitur* examinations and an appropriate implementation of centralized *Abitur* examinations in terms of Educational Governance. Only this extension would allow for a well-founded interpretation regarding the fairness of the examination format as a whole. As the explorative study at hand focuses on intensive courses there should also be a broader perspective on basic courses (basic courses are characterized by a lower number of lessons per week than intensive courses and, depending on the federal state, two or three basic courses are chosen by the students as *Abitur* examinations).

References

Aktionsrat Bildung – Vereinigung der Bayerischen Wirtschaft e.V. (Ed.). (2011). *Gemeinsames Kernabitur: Zur Sicherung von nationalen Bildungsstandards und fairem Hochschulzugang*. Retrieved from http://www.aktionsrat-bildung.de/fileadmin/Dokumente/Gutachten_Gemeinsames_Kernabitur.pdf

Altrichter, H., Brüsemeister, T., & Wissinger, J. (Eds.). (2007). *Educational Governance: Handlungskoordination und Steuerung im Bildungssystem*. Wiesbaden, Germany: VS.

Altrichter, H., & Maag Merki, K. (Eds.). (2010). *Handbuch Neue Steuerung im Schulsystem*. Wiesbaden, Germany: VS.

Amrein, A., & Berliner, D. (2002). High-Stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1–74. Retrieved from http://epaa.asu.edu/ojs/article/view/297/423

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum.

Bache, I., & Flinders, M. (Eds.). (2004). *Multi-Level governance*. Oxford, United Kingdom: Oxford University Press.

Bellenberg, G. (2008). Zur Nutzung von zentralen Abschlussprüfungen als Bausteine eines umfassenden Qualitätssicherungs- und -entwicklungskonzepts – ein Baustellenbericht. In W. Böttcher, W. Bos, H. Döbert, & H. G. Holtappels (Eds.), *Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive* (pp. 223–233). Münster, Germany: Waxmann.

Bevir, M. (2009). *Key concepts in governance*. Los Angeles, CA: SAGE.

Bishop, J. H. (1999). Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review, 6*(2), 349–398.

Bornstein, M. H., Hahn C.-S., & Haynes O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language, 24*(3), 267–304.

Bos, W., Bonsen, M., & Gröhlich, C. (Eds.). (2009). *KESS 7 – Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7*. Münster, Germany: Waxmann.

Camilli, G., & Shepard, L. (1994). *Methods for identifying test bias*. Thousand Oaks, CA: Sage.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing. Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum.

Cole, N. S. (1997). *The ETS Gender Study: How females and males perform in educational settings*. Princeton, NJ: Educational Testing Service.

Creemers, B. P. M., Kyriakides, L., & Sammons, P. (2010). The state of the art of Ecuational Effectiveness Research: Challenges for research methodology. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 3–74). London, United Kingdom: Routledge.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Drechsel, B., & Artelt, C. (2007). Lesekompetenz. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, & R. Pekrun (Eds.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (pp. 225–248). Münster, Germany: Waxmann.

Eickelmann, B., Kahnert, J., & Lorenz, R. (2013). Untersuchung zur Frage von geschlechtsspezifischer Fairness im Zentralabitur im Fach Mathematik. In K.

Schwippert, M. Bonsen, & N. Berkemeyer (Eds.), *Empirische Bildungsforschung* (pp. 147–166). Münster, Germany: Waxmann.
Eickelmann, B., Kahnert, J., Lorenz, R., & Bos, W. (2011). Das Zentralabitur in Nordrhein-Westfalen aus der Lehrerperspektive. Veränderungen für den Unterricht. *SchulVerwaltung NRW, 12*, 31–32.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
Enderlein, H., Walti, S., & Zurn, M. (Eds.). (2010). *Handbook on multi-level governance*. Cheltenham, United Kingdom: Edward Elgar.
Featherston, M. (2011). *High-Stakes testing policy in Texas: Describing the attitudes of young college graduates* (Master's thesis). Retrieved from https://digital.library.txstate.edu/bitstream/handle/ 10877/3484/fulltext.pdf?sequence=1
Fend, H. (2008). *Neue Theorie der Schule: Einführung in das Verstehen von Bildungssystemen* (2nd ed.). Wiesbaden, Germany: VS.
Fend, H. (2011). Die Wirksamkeit der neuen Steuerung – theoretische und methodische Probleme ihrer Evaluation. *Zeitschrift für Bildungsforschung, 1*(1), 5–24.
Halbheer, U., & Reusser, K. (2008). Outputsteuerung, Accountability, Educational Governance – Einführung in Geschichte, Begrifflichkeiten und Funktionen von Bildungsstandards. *Beiträge zur Lehrerbildung, 26*(3), 253–266.
Hamilton, L. S., Stecher, B. M., Russell, J. L., Marsh, J. A., & Miles, J. (2008). Accountability and teaching practices: School-Level actions and teacher responses. In B. Fuller, M. K. Henne, & E. Hannum (Eds.), *Strong stakes, weak schools: The benefits and dilemmas of centralized accountability* (pp. 31–66). Bingley, United Kingdom: Emerald.
Hargreaves, A., & Fullan, M. (2012). *Professional capital. Transforming teaching in every school*. New York, NY: Teachers College Press.
Hartig, J., & Jude, N. (2008). Sprachkompetenzen von Mädchen und Jungen. In E. Klieme (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 202–207). Weinheim, Germany: Beltz. Retrieved from http://www.pedocs.de/volltexte/2010/3149/pdf/978_3_407_25491_7_1A_D_A.pdf
Herman, J. L. (2004). The effects of testing on instruction. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning Accountability Systems for Education* (pp. 141–166). New York, NY: Teachers College Press.
Heubert, J. P., & Hauser, R. M. (1999). *High-Stakes: Testing for tracking, promotion and graduation*. Washington, DC: National Academy Press.
Hirnstein, M., & Hausmann, M. (2010). Kognitive Geschlechtsunterschiede. In G. Steins (Ed.), *Handbuch Psychologie und Geschlechterforschung* (pp. 69–86). Wiesbaden, Germany: VS.
Holland, P., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Hornberg, S., Valtin, R., Potthoff, B., Schwippert, K., & Schulz-Zander, R. (2007). Lesekompetenzen von Jungen und Mädchen im internationalen Vergleich. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert, & R. Valtin (Eds.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 195–223). Münster, Germany: Waxmann.
Ingenkamp, K. (1989). *Diagnostik in der Schule. Beiträge zu Schlüsselfragen der Schülerbeurteilung*. Weinheim, Germany: Belz.
Jürgens, E. (2005). *Leistung und Beurteilung in der Schule. Eine Einführung in Leistungs- und Bewertungsfragen aus pädagogischer Sicht* (6th ed.). Sankt Augustin, Germany: Academia.
Kahnert, J. (2014). *Das Zentralabitur im Fach Mathematik. Eine empirische Analyse von Abitur- und TIMSS-Daten im Vergleich*. Münster, Germany: Waxmann.

Kampshoff, M. (2007). *Geschlechterdifferenz und Schulleistung: Deutsche und englische Studien im Vergleich*. Wiesbaden, Germany: VS.

Klein, E. D., Kühn, S. M., van Ackeren, I., & Block, R. (2009). Wie zentral sind zentrale Prüfungen? Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich. *Zeitschrift für Pädagogik, 55*(4), 596–621.

Klieme, E. (2003). *Zur Entwicklung nationaler Bildungsstandards: Eine Expertise*. Bonn, Germany: BMBF, Referat Öffentlichkeitsarbeit.

Köller, O., & Klieme, E. (2000). Geschlechtsdifferenzen in den mathematisch-naturwissenschaftlichen Leistungen. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *Dritte Internationalen Mathematik- und Naturwissenschaftsstudie zur mathematischen und naturwissenschaftlichen Bildung am Ende der Schullaufbahn* (Vol. 2, pp. 373–404). Opladen, Germany: Leske + Budrich.

Koretz, D., & McCaffrey, D. (2005). *Using IRT DIF methods to evaluate the validity of score gains*. Retrieved from http://ipea.hmdc.harvard.edu/files/ipea/Koretz_McCaffrey__Methods_to_Evaluate_-_Report_660-1.pdf

Koretz, D., McCaffrey, D., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions*. *CSE Technical Report 551*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Kühn, S. M. (2010). *Steuerung und Innovation durch Abschlussprüfungen?* Wiesbaden, Germany: VS.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, United Kingdom: CUP

Lehmann, R. H. (1994). Lernen Mädchen wirklich besser? Ergebnisse aus der internationalen IEA-Lesestudie. In S. Richter & H. Bruegelmann (Eds.), *Mädchen lernen anders lernen Jungen. Geschlechtsspezifische Unterschiede beim Schriftspracherwerb*. Lengwil, Switzerland: Libelle.

Lehmann, R. H., Gänsfuß, R., & Peek, R. (1999). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern an Hamburger Schulen. Klassenstufe 7. Bericht über die Untersuchung im September 1998*. Unpublished research report.

Lenzen, D., & Blossfeld, H.-P. (2009). *Geschlechterdifferenzen im Bildungssystem: Jahresgutachten 2009*. Wiesbaden, Germany: VS. Retrieved from http://dx.doi.org/10.1007/978-3-531-91835-8

Li, Y., Cohen, A., & Ibarra, R. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115–136.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lorenz, R. (2013). *Das Zentralabitur im Kontext der Bildungsgerechtigkeit – Schwierigkeit und Fairness der Abituraufgaben im Fach Englisch in NRW*. Münster, Germany: Waxmann.

Lorenz, R., Eickelmann, B., & Dohe, C. (2013). Fairness von zentralen Abituraufgaben. Geschlechtsspezifische Unterschiede im Fach Englisch in NRW. In N. McElvany, M. Gebauer, W. Bos, & H. G. Holtappels (Eds.), *Jahrbuch Schulentwicklung 17* (pp. 236–263). Weinheim, Germany: Juventa.

Lorenz, R., Kahnert, J., Eickelmann, B., & Bos, W. (2011). Mehr Gerechtigkeit durch Zentralabitur? Analysen einer Lehrerbefragung in NRW. *Zeitschrift Schul-Management, 42*(6), 24–27.

Maag Merki, K. (Ed.). (2012). *Zentralabitur: Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland*. Wiesbaden, Germany: VS.

Maag Merki, K., Langer, R., & Altrichter, H. (Eds.). (2014). *Educational Governance als Forschungsperspektive. Strategien. Methoden. Ansätze*. Wiesbaden, Germany: VS.

Madaus, G., & Russell, M. (2009). *The paradoxes of high stakes testing. How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age.

Moosbrugger, H., & Kelava, A. (Eds.). (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg, Germany: Springer Medizin.

Mullis, I. V. S., Martin, M. O., Fierros, E. G., Goldberg, A. L., & Stemler, S. (2000). *Gender differences in achievement: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College.

Nichols, S., Glass, G., & Berliner, D. (2005). *High-Stakes testing and student achievement: Problems for the No Child Left Behind Act*. Retrieved from http://nepc.colorado.edu/files/EPSL-0509-105-EPRU.pdf

Nikolova, R., & Ivanov, S. (2010). Englischleistungen. In W. Bos & C. Gröhlich (Eds.), *KESS 8. Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8* (pp. 49–66). Münster, Germany: Waxmann.

OECD – Organisation for Economic Co-operation and development. (2010). *PISA 2009 Ergebnisse. Was Schülerinnen und Schüler wissen und können: Schülerleistungen in Lesekompetenz, Mathematik und Naturwissenschaften* (Vol. 1). Bielefeld, Germany: Bertelsmann. Retrieved from http://deposit.d-nb.de/cgi-bin/dokserv?id=3548688undprov= Munddok_var=1unddok_ext=htm

OECD – Organisation for Economic Co-operation and development. (2013). *PISA 2012 results: What students know and can do. Student performance in mathematics, reading and science* (Vol. 1). Paris, France: OECD Publishing. Retrieved from http://www.keepeek.com/Digital-AssetManagement/oecd/edu cation/pisa-2012-results-excellence-through-equity-volume-ii_9789264201132-en

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26. Psychometrics* (pp. 125–167). Amsterdam, Netherlands: Elsevier.

Schräpler, P., & Schmidtke, K. (2011). Wer besteht das Abitur? Erfolg und Nichterfolg bei Abiturprüfungen in Nordrhein-Westfalen. *Statistik Kompakt, 6/2011*. Retrieved from http://www.it.nrw.de/statistik/querschnittsveroeffentlichungen/Statistik_kompakt/ausgabe6_2011.html

Stanat, P., & Kunter, M. (2003). Kompetenzerwerb, Bildungsbeteiligung und Schullaufbahn von Mädchen und Jungen im Ländervergleich. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 212–242). Opladen, Germany: Leske + Budrich.

Strüber, D. (2008). Geschlechtsunterschiede im Verhalten und ihre hirnbiologischen Grundlagen. In M. Matzner & W. Tischner (Eds.), *Handbuch Jungenpädagogik* (pp. 34–48). Weinheim, Germany: Beltz.

Venables, W. N., Smith, D. M., & R. Development Core Team (2011). *An introduction to R. Notes on R: A programming environment for data analysis and graphics*. Retrieved from http://www.r-project.org

Walther, G., Schwippert, K., Lankes, E.-M., & Stubbe, T. C. (2008). Können Mädchen doch rechnen? Vertiefende Analysen zu Geschlechtsdifferenzen im Bereich Mathematik auf Basis der Internationalen Grundschul-Lese-Untersuchung IGLU. *Zeitschrift für Erziehungswissenschaft, 11*(1), 30–46.