MATHEMATICAL COMPETENCIES IN HIGH SCHOOL EXIT EXAMINATIONS AND IN TIMSS/III – VALIDATION OF HIGH STAKES-TESTING THROUGH AN INTERNATIONAL LARGE SCALE ASSESSMENT INSTRUMENT

Julia Kahnert, Birgit Eickelmann, Wilfried Bos & Manuela Endberg
Institute for School Development Research, TU Dortmund University, Germany

Abstract

In Germany students gain access to universities by passing the high school exit examination (Abitur). By using probabilistic scaling methods, this study examines whether central high school exit examinations as a form of high-stakes testing allow for valid information in measuring competencies. For this approach a cohort of students simultaneously took their high school exit examinations in mathematics and the TIMSS/III-test (Mullis et al., 1998). Both instruments entail information about students’ mathematical competencies. The presented study establishes in how far both results are comparable and which consequences for high-stakes testing as well as for TIMSS/III can be deduced.

Introduction, Background, and Objectives of Study

The presented research project is conducted as a sub-project for the project “Quality management of central examinations” organized by the Institute of School Development Research at Dortmund University. The project is funded by the school ministry in North Rhine-Westphalia. Data was collected in the survey period during February and March 2011.

This paper deals with the topic of high school exit examinations as a certain form of high-stakes testing in Germany. Exemplarily for the subject mathematics, students’ performance data was collected and analysed. One of the most important findings in line with our proposed research questions is the result of the comparison between the mathematics A-level examination\(^1\) and the pre-university TIMSS-test. As the title of the paper suggests, we sought validation of high-stakes testing (in form of the central exit examination) through an international large scale assessment instrument (the pre-university TIMSS-test mathematical); The main objective was to find out whether both testing procedures measure the same mathematical competencies and therefore being able to state whether performance results for both tests appear to be comparable.

To start off, we will provide some crucial background information concerning the current situation of the German school system.

Based on the unexpectedly poor results of German students in international assessment studies such as PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study), numerous reforms for the German school system have been triggered. During the last couple of years, national and international educational controlling systems predominantly shifted from input control to output control (Böttcher 2004; Maag Merki, 2010). While education policy formerly focused on centrally controlling educational systems through detailed curricula, economic resources and professionalizing teachers, output control now attaches more importance to setting objectives and to the evaluation of results (Bellmann & Weiß, 2009).

The overall goal is to put a stronger focus on educational performance by introducing various forms of external evaluations for the schools. One major aspect of this is the implementation of central high

---
\(^1\) A-Level examination and high school exit examination: Both terms will be used interchangeably throughout this paper.
school exit examinations (Maag Merki, 2008). These central A-level examinations can be seen as a core feature of New Governance. The shift from decentralized school exit examinations to centralized exit examinations has lately become a special matter of interest for educational researchers.

One main reason for the introduction of central high school exit examinations was the lack of comparability of graduations as well as variances of student achievement both in-between schools and in-between federal states. So, in order to raise comparability and to reduce the achievement variance, a huge wave of implementing central A-level examinations occurred in Germany since the year 2005. Some federal states (e.g. Bavaria, Baden-Wuerttemberg and Saarland) actually have a long tradition of central examinations, whereas others only quite recently shifted from decentralized examination procedures to centralized ones. Today, in 15 out of 16 federal states in Germany central high school exit examinations have been implemented; the only exception being Rhineland-Palatinate.

In Germany, students gain access to universities and colleges by passing the high school exit examination (Abitur). Therefore, graduation from an academic high school is the highest graduation achievable in the German school system. However, this ‘high-stake’ does only apply for the students. Teachers and schools do not have to fear any negative consequences in case of low success rates or low average student test performance as opposed to other states using high-stakes testing, e.g. the USA.

This paper focuses on the aforementioned developments by casting light on the shift towards centralized A-level examinations in one federal state of Germany, namely North Rhine-Westphalia, for the subject of mathematics.

Our main research question is to examine whether the high school exit examinations as a form of high-stakes testing allow for valid information in measuring—in our case mathematical—competencies. For this approach a cohort of students simultaneously took their high school exit examinations in mathematics and the TIMSS/III-test (Mullis et al., 1998). Both instruments entail information about students’ mathematical competencies. So, in order to frame our findings on centralized A-level examinations we used the standardized pre-university TIMSS-test as a means of comparison by applying probabilistic scaling methods. The presented study establishes in how far both results are comparable and which consequences for high-stakes testing as well as for TIMSS/III can be deduced. This study’s research approach combines national and international perspectives in order to link high-stakes testing performances to large scale assessment tests results. The international TIMS-Study is therefore used to verify the validity of the national high-stakes testing method.

**Theoretical or Conceptual Framework**

In many countries high-stakes testing is both well-established and critically discussed. We will thus give a bit of descriptive information as well as empirical findings on high-stakes testing procedures. The following not fully exploiting overview is mostly based on papers and findings from the USA.

**Definition of high-stakes testing**

“High-stakes tests are tests from which results are used to make significant educational decisions about schools, teachers, administrators and students.” (Amrein & Berliner, 2002a, 5). High-stakes testing comprises a specific reform approach which links results from comparative assessment tests to rewards and sanctions for schools, teachers and students. The predominant use of high-stakes tests in the US-American school system is based on the “No Child Left Behind Act” (NCLB) introduced by former president George W. Bush in 2002. According to this act, each child is to be supported in developing their individual talents and potentials in order to reach the set default competency levels needed to pass these kinds of centralized assessment tests. High-stakes testing is distinctive to raise school accountability by linking comparison tests results to rewards and sanctions for students (e.g. to not graduate school) and additionally providing comparable performance data (Bellmann & Weiss, 2009).
As already mentioned before, high-stakes tests hold serious consequences (high-stakes) for different agents of the school system. These consequences may turn out to be positive in the sense of rewards in case of good testing results or respectively negative in the form of sanctions for bad testing results. The following paragraph comprises a short list of possible positive and negative consequences for schools, teachers and students:

- **On school level** financial rewards may be given out to successful or improved schools, whereas in case of low results “[…] state government has the power to close, reconstitute, or take over low performing schools” (Amrein & Berliner, 2002b, 2).
- **Teachers** may be rewarded with financial bonuses in case of high average class scores or improvement of average class scores; in case of low average class scores teachers can be displaced or removed by the state government.
- For **students** high testing scores may lead to special diplomas or scholarships. Some US states use low testing scores to prevent students from graduating high school, i.e. they do not receive their high school diploma. As Amrein and Berliner state: “even if they meet all other requirements for graduation but fail the high school graduation exam, they are denied a high school diploma” (Amrein & Berliner, 2002b, 2)

In order to find out the effects of high-stakes testing procedures, several international studies were conducted (e.g. Amrein & Berliner, 2002a, 2002b, 2002c; Hamilton & Koretz; Nichols, Glass & Berliner, 2005). Bellmann & Weiβ, 2005) differentiate the main results of these studies in intended and unintended effects of high-stakes testing. According to Maag Merki (2010) the intended effects can be found on various dimensions in different manifestations.

**Intended effects of high-stakes testing**

Maag Merki distinguishes intended effects of high-stakes testing on the teacher level from the teaching level and the school level.

- **Teacher level**: High-stakes tests are deemed to have positive effects on teachers’ behavior, especially as an increase in motivation and working more effectively.
- **Teaching level**: On the teaching level high-stakes tests might lead to increased use of more complex teaching methods, a stronger focus on student support, stronger focus on students’ learning results and integration of central contents into the individual teacher curriculum.
- **School level**: On school level high-stakes tests are supposed to result in a higher standard of teacher professionalism as well as an increase of teacher cooperation. Furthermore, high-stakes tests are helpful to raise the significance of school quality.

Of great interest for empirical educational research is the assumed intended effect of increased student achievement through high-stakes testing. A great number of studies have been conducted to examine this thesis (e.g. Amrein & Berliner, 2002a, 2002b; Bishop, 1998; Bishop & Wößmann, 2004; Jürges, Schneider & Büchel, 2005; Jürges, Schneider, Senkbeil & Carstensen, 2009; Nichols, Glass & Berliner, 2005; Wößmann, Ludger, 2003a, 2003b). However, up until now only inadequate evidence to support the proposition that high-stakes tests raise student achievement could be found: “[…] there is no consistent evidence that high-stakes testing works to increase achievement” (Nichols, Glass & Berliner, 2005).

Moreover, regarding the intended effect of increased student achievement through the implementation of high-stakes tests and high school exit examinations, Amrein and Berliner (2002) report that “the data […] suggest that after the implementation of high-stakes tests, nothing much happens” and that “the data […] also suggest […] that after the implementation of high school graduation exams, academic achievement apparently decreases (Amrein & Berliner, 2002a, 57).”
Unintended effects of high stakes testing

- **Cheating teachers:** For example: Increasing time limits on timed portions of tests, changing students’ answers to test questions, helping students to give the correct answers to test questions, providing hints, rephrasing or clarifying test questions (Amrein & Berliner, 2002b, 46)

- **Cheating students:** Test results may not only be influenced by teachers, but also by the students. Worth mentioning is “student cheating”, meaning that during the testing students cheat in order to score higher results (Nichols & Berliner, 2005, 53).

- **Teaching to the test:** Teachers may systematically try to prepare students for the test situation by adapting their teaching contents as well as the overall curriculum according to the testing procedures and question types that are asked in the test as well as teaching them test-taking strategies. Amrein and Berliner (2002) describe the problem of the teaching to the test-effect as follows: “When teachers teach to the test, students become experts at answering test questions without entirely understanding the concepts behind their answers.” (Amrein & Berliner, 2002b, 41)

German education policy renders central A-level examinations as appropriate evaluative procedures to control schooling and learning processes in order to achieve increased school quality. The following paragraph lists several reasons for the implementation of central high school exit examinations as well as their ascribed functions.

Reasons for implementing central high school exit examinations in Germany

One starting point for the increased implementation of central high school exit examinations in Germany was the unexpectedly negative results of German students in international school assessment studies such as PISA and TIMSS. These revealed remarkable differences in students’ competencies, especially in relation to the factors gender and social background (cf. e.g. Baumert 2000a, 2000b; Klieme et al., 2010). Furthermore, these findings hinted at a great variance in students’ achievement in and in-between federal states, entailing a very questionable comparison of graduations with regard to the procedures of A-Level examination in different federal states and between different school types (e.g. academic high schools – comprehensive schools) (Neumann, Nagy, Trautwein & Lüdtke, 2009; Trautwein, Köller, Lehmann & Lüdtke, 2007). These observations induced enhanced output control through school-external and interscholastic control elements, e.g. centralized A-level examinations. Therefore, in order to align the value of high school graduations throughout the majority of Germany by means of standardized testing contents and central examination processes, central high school exit examinations have been implemented in 15 out of 16 federal states in Germany. Only the state of Rhineland-Palatinate is still conducting A-level examinations in a decentralized manner, meaning that both the development and the evaluation of examination questions are done by subject teachers. As opposed to this, centralized high school exit examination questions are developed by the school ministry in collaboration with a committee of subject experts, while the assessment of the written examinations is done by subject teachers using a standardized list of how to evaluate the different parts of the written exams.

In Germany, central A-Level examinations can be characterized as high-stakes tests for students, because the exam results hold serious consequences for students’ future career decisions. By passing the high school exit examination, German students gain access to universities and colleges in Germany. The “Abitur” as graduation from an academic high school is also a prerequisite for taking up certain qualified jobs.

There are no negative consequences for German students, who do not pass their A-level examination other than not gaining university and college access. They still receive graduation (although of a lower
qualification level). In this context it is also important to mention that in Germany there are several alternative ways to achieve the qualification of “Abitur” (e.g. during the course of certain kinds of vocational trainings, in night school, etc.)

There are no consequences, neither positive nor negative, for teachers or schools in case of low passing and graduation rates.

In summary, central A-level examinations were implemented in Germany to fulfil certain functions.

Functions of high school exit examinations in Germany

Certain functions are ascribed to central exit examinations. Kühn (2010) lists functions referring to quality, selection and allocation, legitimation, comparability, feedback, innovation, controlling, professionalization, relief and motivation (Kühn, 2010a). These functions can be distinguished in external and internal functions with external functions aiming at schools’ publicity and internal functions concerning all participants of school life [cf. Table 1] (Wacker, 2008).

Table 1  
Functions of central exit examinations

<table>
<thead>
<tr>
<th>Society level</th>
<th>School life level</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Quality function</td>
<td>- Feedback function</td>
</tr>
<tr>
<td>- Selection and allocation function</td>
<td>- Innovation function</td>
</tr>
<tr>
<td>- Legitimation function</td>
<td>- Control function</td>
</tr>
<tr>
<td>- Comparability function</td>
<td>- Professionalization function</td>
</tr>
<tr>
<td></td>
<td>- Relief function</td>
</tr>
<tr>
<td></td>
<td>- Motivation function</td>
</tr>
</tbody>
</table>

Quality function entails for example the guarantee of students gaining knowledge and competencies of at least default standard level needed for leading an autonomous life and taking up an occupation. With the aid of uniform quality standards, central exit examinations therefore assess students’ cognitive output, which is seen as important input for universities or firms. Moreover, based on nation-wide uniform quality standards the comparability function ascertains comparability of graduations in and in-between federal states, between different schools and different types of schools.

On school life level central exit examinations hold a feedback function for students as well as teachers. Students receive feedback for their academic performance, teachers gain information for future teaching processes.

Also, due to central exit examinations professionalization of teachers and students increases. For teachers, this happens through optimization of teaching processes related to central exams. Students become more professional by increase of autonomous learning within the scope of exam preparations.

The core of the central A-level examinations is made up by the exam questions which are to assess the learning progress of students at the end of academic high school. Due to the high significance of these questions, we will now give a short overview of the current state of research for high school exit examination questions.

State of research for high school exit examination questions

In Germany there are two approaches towards the evaluation of A-level examination questions. On the
one hand, students and teachers evaluate actual exam questions through questionnaires (cf. e.g. Landesinstitut für Schule und Medien Berlin-Brandenburg, 2009; Maag Merki, 2008; Lorenz et al., 2011; Kahnert et al., 2012). These researches revealed that students and teachers see the aspiration level, content, comprehensibility and complexity of exam questions as appropriate. On the other hand, content-related analyses of these questions are executed (cf. e.g. Bolle-Bovier, 1994; Brockhage & Weghöft, 2004; Kühn 2010; Paul, 2002). The findings show that predominantly traditional tasks and questions are used instead of innovative ones. Material is mostly provided in form of texts and images; experimental tasks (e.g. useful for chemistry) are seldom if ever employed in the construction of A-level examinations. Typically, more professional knowledge is retrieved whereas learned competencies are rarely focused. Also, reproduction and reorganization outweigh problem-oriented tasks and questions in most cases.

Until today, a research gap exists for analysis of what A-level examination questions actually measure. This is where our research interest lies. The presented study is exemplarily conducted for the subject mathematics. The main objective is to find out whether high school examination questions measure a mathematical competency. International standardized school assessment studies already provided definitions of mathematical competencies. That is why we chose the pre-university TIMSS-test for the purpose of comparison. We pursue the research question whether our national high-stakes test measures the same mathematical competencies as the international pre-university TIMSS-test. The following paragraph states reasons why we chose the subject mathematics as basis for the upcoming analyses.

Why mathematics?
Mathematics is of particular research interest for several reasons. First, in Germany mathematics is from the very beginning of primary school until the end of academic high school (as well as throughout the schooling process in other school forms) characterized as a compulsory subject for all students. Furthermore, mathematics, or more precisely mathematical competency, is defined as one of the (eight) key competencies of the 21st century for life-long learning (Kommission der Europäischen Gemeinschaften, 2008, 6).

The basis of mathematical competency is mathematical literacy, which is defined by the OECD, PISA Deutschland (1999) as follows:

„Mathematical literacy is an individual’s ability, in dealing with the world, to identify, to understand, to engage in, and to make well-founded judgments about the role that mathematics plays, as needed for that individual’s current and future life as a constructive, concerned, and reflective citizen.“ (OECD, PISA Deutschland, 1999)

The classification of mathematics as a key competence can therefore be seen as one reason for the significant role mathematics plays in international school assessment studies, the most well-known being TIMSS and PISA. Moreover, acquisition of mathematical competency is considered one of the main objectives of the German school system, because it constitutes an important precondition for social participation as well as for a free, autonomous and independent conduct of life (Neumann, 2010).

Now, we will introduce our research project entailing test instruments, objectives, research questions, data basis and data collection as well as applied methods of analysis.

Research Project

Test instruments
For the description of test instruments we would like to emphasize in how far both instruments differ from each other.

Central high school exit examinations in mathematics in North-Rhine Westphalia
In North Rhine-Westphalia all exit examinations for graduating upper secondary education are
conducted as central written examinations. Contentwise and thematically all examination questions are formulated according to the obligatory standards of the current curricula for upper secondary education (Ministry for Schools and Further Education of North Rhine-Westphalia, 2011).

The central high school exit examinations in mathematics for the year 2010/2011 in North Rhine-Westphalia had to cover the following three thematic priorities: analysis, linear algebra/geometry and stochastic theory. In compliance with certain formal regulations, the subject teachers were to choose a set of examination questions from a large question pool.

TIMSS (Trends in International Mathematics and Science Study)
TIMSS is an international school assessment study which is often referred to in line with other international large school assessment studies, especially PISA and PIRLS. TIMSS is conducted by the IEA (International Association for the Evaluation of Educational Achievement). The main purpose of the study is the international comparison of students’ competencies and achievement in school mathematics and science. The study is conducted every four years since 1995 for grade 4 and grade 8 students.

The pre-university TIMSS-test, which was used in the process of our data collection, is a special case. As part of TIMSS/III (Third International Mathematics and Science Study) it has the objective to assess the mathematical and science competencies of students at the end of academic high school (in Germany Grade 12/13), right before the entering time into university/college.

Central High School Exit Examinations and pre-university TIMSS-test
Central A-Level examinations have been implemented in Germany in order to raise the comparability of academic high school degrees in-between schools as well as in-between federal states. The central high school exit examinations are to be seen as a half-standardized method of measuring student achievement. Students take their central A-level examinations which are based on consistent curricular standards at the same time and under the same testing conditions (e.g. time frames, opportunities to choose examination questions from a pool of questions, etc.). Still, central A-level examinations have to be classified as a half-standardized test method due to the fact that subject teachers assess their students’ test results. Assessment is done according to a standardized list of assessment criteria that also states how many points are maximally to be assigned for one question or question part. However, since there are no centralized standards for right or wrong answers (the teachers are merely provided with sample solutions), teachers are given some kind of leeway for the assessment of their students’ answers. Opposed to this the pre-university TIMSS-test is a fully standardized test, e.g. each question can be answered either correctly or incorrectly with a default number of points assigned to each question.

Objective of study
The aim of this paper is to reveal whether the half-standardized method of student assessment in mathematical competency is as valid as and therefore comparable to a standardized method of measuring students’ mathematical achievement. Previous studies (TOSCA (Transformation of Secondary School Systems and Academic Careers) and TIMSS/III) have shown that there is a correlation between students’ grades in mathematics and the TIMSS test results. Therefore, in an attempt to answer our main research question we chose the pre-university TIMSS-test in mathematics, which was part of TIMSS/III, for comparison with central high school exit examinations in mathematics. The pre-university TIMSS-test is characterized by a high level of curricular validity and matches the requirements of the central A-level examinations.

Research questions
The main objective of this research paper is to examine how valid the central high school exit examinations in mathematics as a half-standardized method of testing student’s mathematical achievement are compared to a fully standardized test method of measuring students’ mathematical competency, in this case the pre-university TIMSS-test. We will therefore pursue the following research questions in this paper:
In how far is it possible for central A-level examination questions to be scaled along with a standardized test format and to be described by using the Partial Credit Model?

In how far is it possible to describe central A-level examination questions along with a standardized test format using a 2-dimensional model which covers differences in both test formats?

Do students tend to perform better in their central A-level examinations or in the standardized test format?

Do students who perform well in their central A-level examinations also score high in the standardized test format?

High-stakes testing is distinctive to raise school accountability by linking comparison tests results to rewards and sanctions for students (e.g. to not graduate school) and additionally providing comparable performance data (Bellmann & Weiß, 2009). Of particular research interest on national and international scale are studies about intended (e.g. performance growth) and unintended effects (teaching-to-the-test) of high-stakes testing (cf. e.g. Eickelmann et al., 2011; Maag Merki, 2011; Wößmann, 2003; Amrein & Berliner, 2002; Bishop, 1994). With regard to these findings this study examines the intended effect of high-stakes testing to assess students’ competencies in a specific subject (mathematics).

**Research Methods and Modes of Inquiry**

In spring 2011 advanced level mathematics academic high school students worked on the tasks of the pre-university TIMSS-test in order to collect their mathematical competencies. On another testing day these students took their official central high school exit examinations. It needs to be pointed out here that these students had to take their high school exit examination in mathematics on this particular day due to the fact that the examination is centrally organized and taken by all advanced level mathematics high school students of North Rhine-Westphalia on the same day. The central A-level examination can be characterized as a high-stakes test for the students since passing results in achieving general access to all German universities and colleges. It also opens the door to quite a large number of qualified jobs. The pre-university TIMSS-test was scheduled prior to the central A-level examination; hence, the pre-university TIMSS-test has to be considered a low-stakes test because its results would not have any negative or positive consequences for the students.

The central mathematics examination results for each student as well as a list of points gained for each partial task per student were collected by the subject teacher in the form of an assessment sheet filled in by the teachers themselves. For each student the results for both the high school exit examination (as a half-standardized test) and the TIMSS-test (as a fully standardized test) have been analyzed by Rasch scaling methods and checked for correlations and their rank correlations.

**Sample and Data Sources**

In order to answer our research questions, several preliminary considerations were needed. First, we needed an A-level cohort in one federal state. Our research is based on the A-level cohort for 2011 in North Rhine-Westphalia. For the next step, we needed an A-level examination subject. We decided on mathematics. The reasons for choosing mathematics were given earlier in this paper. For comparing the central A-level examination as a form of half-standardized testing with a standardized competency test, we chose the pre-university TIMSS-test mathematics. We settled on the TIMSS-test for two reasons. First, because it matches the learning content and learning objectives of the upper secondary classes and second, because the arrangement of the TIMSS-test questions is basically equivalent to the requirements to be met in the written A-level examinations.

Furthermore, the TIMSS-test holds an overall curricular validity, which we even had confirmed by experts before the actual comparison took place.

Our sample consists of 318 academic high school graduates who passed their final examination in spring 2011. The data is compiled for 14 schools and a total of 17 advanced level mathematics classes. Additionally the students’ mathematical competencies were assessed through the pre-university TIMSS-test which is part of the international TIMSS-Study.

We have very detailed information on each student’s A-level performance in the form of assessment
sheets which were filled in by their mathematics teachers during the evaluation of the individual student’s examination results. For each examination question and its sub-questions the assessment sheet in the end holds information on how many points can be maximally achieved and how many points each student actually obtained. Exactly the same range of students who took their central A-levels in mathematics also took the TIMSS-test, so that we also have access to each student’s competency test performance.

Methods of Analysis
By applying the Rasch model, performance assessment tests like the A-level examination questions as well as the pre-university TIMSS-test questions can be analyzed with due respect to test theory. Whether a question was answered correctly or incorrectly will be recorded for each person. In the course of this item and person parameters are calculated. The item parameter is defined as the item difficulty, whereas the person parameter entails the person ability which therefore marks the position or value of the person on the latent variable (Bühner, 2006). This position on the latent variable is estimated by use of the person parameter. To accomplish that, the accumulated values are used, in our case the number of solved items (Bond & Fox, 2007; Bühner, 2006). The estimated person and item parameters have the same unit of measurement and thus can be mapped onto the same scale, the logit/metric scale. This scale allows for values ranging from plus to minus infinity; most of the time, however, values between plus three and minus three occur. (Bos et al., 2003; Bühner, 2006). Negative values signify easy items and persons with lower ability, while positive values indicate difficult items and persons with higher ability (Bühner, 2006).

In order to scale tests which entail more than two solution categories, as it is the case for the A-level examination questions and the TIMMS-questions, the Partial Credit Model can be used.

The Rasch Model in the form of a Partial Credit Model (Masters, 1982; Whrigt & Masters, 1982) is also applicable for manifest variables with more than two solution categories.

For evaluation of the model quality, statistics can be used to show in how far the empirical data matches the assumptions of the theoretical model. For the calculation of the overall model, ConQuest provides the final deviance. This value is important for the model comparison to see for example whether a one-dimensional model matches the data better than a multidimensional one. There are several other fit values to indicate which model fits the data best. Worth mentioning are Akaike Information Criterion (AIC) (Akaike, 1973), Consistent AIC (CAIC) (Bozdogan, 1987) and Bayes Information Criterion (BIC) (Schwarz, 1978). For all these values the rule applies that lower values imply a better fit of the theoretical model to the empirical data. Information about the model quality can also be gained from the fit-statistics on item level. These statistics show whether single test items correspond with the model assumptions. Conquest provides infit and outfit statistics, which count among the residual values and are based on the difference between the empirically calculated frequencies of solution and the probabilities of solution according to the model (Bond & Fox, 2007; Rost, 2004; Whrigt & Masters, 1982; Wu, Adams, Wilsons, & Haldane, 2007). For the evaluation of the model conformity of a test item the infit-value weighted mean square (MNSQ) proved to be successful (Voss, 2006). MNSQ has an expected value of 1 and can only take positive values. Values greater or smaller than 1 indicate problems of model conformity of the item. Infit-values greater than 1.20 and smaller 0.80 are labeled problematic (Adams, 2002).

To test whether A-level examination questions measure the same mathematical competencies as the TIMMS-questions, a multidimensional Rasch model is also calculated.

By means of rank correlation it is further queried whether students who perform well in the high school exit examinations also achieve high scores in the pre-university TIMSS-test. A rank correlation coefficient measures how well a random monotone function is able to describe the correlation between two variables. (Klieme, Artelt, Hartig, Jude, & Köller, 2010). We chose Spearmans Rho, since it assumes equidistant ranks (Klieme et al., 2010). In our case the rank correlation between student performance in the TIMSS-test (variable X) and in the A-level examination mathematics (variable Y) is to be calculated. Based on the maximal achievable scores of both tests, the rank orders are formed in
such a way that for each value of both variables X and X the number of places is allocated as the rank gained when all values are arranged according to their extent. Thus, the highest value is ranked as 1. The closer the rank correlation coefficient is to 1, the higher is the correlation between both variables.

Findings
To answer the first research question we will now check whether the exit examination data and the TIMSS-data can be scaled together in a one-dimensional model. For the following analysis we choose the one-parametric one-dimensional Rasch model. We use the partial-credit-model (Masters 1982; Wrigh, 1982), because each cut-off point, each value of an item, for both the mathematic exam questions and the TIMSS-questions are considered. The maximal points to be obtained for each TIMSS-question are three, whereas in the mathematics exam questions nine points can be maximally gained. Using this model, the data needs not to be dichotomized plus there is no loss of valuable information. We have 90 examination items and 20 TIMSS-items to be scaled jointly. In both tests the three mathematical contents ‘analysis’, ‘linear algebra/geometry’ and ‘stochastic theory’ are represented. So, considering the contents both tests are comparable. All 318 sampled students took the pre-university TIMSS-test. The exit examination questions are arranged in a multi-matrix-design so that not all students worked on all of the eight tests available. Instead, the teacher was to choose three question sets whereas one of them needed to represent ‘analysis’. The other two sets could be chosen randomly; however, at least two content areas needed to be covered. Nevertheless, the presented results are to be interpreted regardless which question sets the students worked on (local stochastic independence), for we are assuming that each question set measures the same mathematical competency.

The following figure shows the results of the joint scaling of the exit examination questions and the TIMSS-questions.

![Figure 1. Scaling results for A-level-questions and TIMSS-questions (one-dimensional)](image-url)
To the left of the metric the latent distributions of the person ability parameter are indicated by ‘x’. Each ‘x’ represents 2,9 cases. An ‘x’ within the lower range of the scale (x < 0) refers to low person ability, whereas an ‘x’ within the higher range of the scale (x > 1) depicts high person ability. In parallel, to the right of the metric the item difficulties ranging from 1-110 are indicated. For each value point of each item the difficulties are denoted. The first figure signifies the item number (e.g. 5). The figure after the point represents the cut-off points (e.g. 5.1). For those items that entail 0 or only 1 point to be obtained, no cut-off points are indicated in the figure. For these items person ability cannot be reckoned since it is formed by the number of items solved (partly) correctly. If 0 points are reached for an item, however, this implies that this item was not solved. Congruent with the person ability, items in the higher range of the scale are harder to solve; respectively, lower ranked items are easier to solve. An item is labelled ‘easy’ if it was solved by a great number of high-school graduates. Items 1-90 are exit examination questions and items 91-110 are pre-university TIMSS-test questions.

To answer the question whether both test-data can be jointly scaled within the partial-credit-model, we need to look at the item fit-statistics. For evaluation of the model quality we use the in-fit-value MNSQ. Our evaluation is oriented towards PISA defaults, stating in-fit-values greater than 1.20 and smaller than 0.80 as indicators for problematic questions. While reviewing the values we found 43 cut-off points not to be compliant with the model. This means, there are items the cut-off points of which are not selective, e.g. not demanding a higher person ability of students in order to reach two instead of one point. Since we checked each cut-off point, each point achievable for each item, and counted those as individual items, 43 out of 447 are not compliant with the model. So, there is only a small share of items being not compliant with the model. Further, these items all stem from the pool of exit examination questions; all TIMSS-items are compliant with the model (as was to be expected, since these items were already used and tested in a sample in 1995). The exit examination questions, however, are analysed this way for the very first time. Due to the fact that only a small share of items is non-compliant, our research question can still be answered: it is possible to jointly scale both tests in a partial-credit-model, taking into account the restriction of non-compliant items.

Using one-dimensional scaling does not yet show whether both tests measure the same mathematical competency. Thus, we will further check whether both tests can be scaled better using a two-dimensional model, with the first dimension being the exit examination question sets and the second dimension entailing all TIMSS-questions. Through the quality criteria final deviance, BIC, CAIC and AIC we can then decide if the one- or the two-dimensional model fits the data better, implying that either only one or two different kinds of mathematical competencies are measured. Moreover, the two-dimensional model can be used to confirm in which of the tests the students show better performance.

Figure 2 (cf. Page 12) shows the results for the two-dimensional scaling.

Evaluation of the in-fit statistics reveals that 40 cut-off points of the items are non-compliant with the model. As before, these involve only exit examination items; all TIMSS-questions are compliant with the model. Model compliance is therefore similar to the one-dimensional model. Comparison of the quality criteria shows that for the one-dimensional model all chosen values are smaller than for the two-dimensional model: the one-dimensional model fits the data better than the two-dimensional one. (cf. Table 2).
This can be seen as a first hint towards both tests measuring the same mathematical competency; a second clue being the correlation of 0.723 of both dimensions. So, there is a medium linear correlation between both dimensions.

In calculating the average person ability, the two-dimensional model also makes for the analysis whether students perform better in the high school exit examinations or in the TIMSS-test. The average person ability for the exit examination questions is 0.356 and for the TIMSS-test 0.292. Thus, students on average show better performance in the exit examination than in the TIMSS-test. This was to be expected, because students were much more familiarized with the exit examination format than with the TIMSS-test. Also, students worked on the TIMSS-test prior to the exit examinations, so that there might have been a learning effect, as well. Finally, the high school exit examination as a high-stakes test is of much greater importance to the students.

Thirdly, we want to pursue the question whether students who perform well in the exit examination in
mathematics also achieve high scores in the TIMSS-test. In order to answer this we used a rank correlation. From the rank correlation of 0.505 we can deduce that there is a concordant monotone correlation between examination and TIMSS-test performance data, meaning that high performance in the exit examination goes hand in hand with high scores in the TIMSS-test. However, a correlation of 0.505 is merely defined as medium correlation. We can therefore conclude that high performance in the A-level examination does not necessarily go along with high performance in the TIMSS-test, and vice versa. This result remains valid even if we consider the attenuation correction which estimates how high two variables would correlate if there were no errors in measurement, meaning both variables would have perfect reliability. Including the attenuation correction, correlation between both tests would increase to 0.666 which is still to be interpreted as medium correlation.

Conclusions, Scholarly or Scientific Significance, and Implications

Our findings induce further research on national and international level, especially aiming at linking national tests to international approaches. Focusing on the quality discussion of national testing instruments and their impact on school accountability, a more in-depth discussion about structural prediction features of classroom teaching ensues. The approach to use international large-scale-assessment tests to validate national high-stakes tests can be of further international use. Future research may show whether these findings could be confirmed for bigger samples in other subjects and other high-stakes testing countries.

Additionally, the data allows for even further analyses, given the fact that during data collection also a student questionnaire covering students’ background variables was used. Therefore, in using structural equation modelling, it would be possible to check in how far mathematical performance can be ascribed to teaching and classroom variables or mathematical interest. With this we might be able to find out the importance of teaching quality or even the significance of teaching to the test-effects on students’ mathematical performance.

In conclusion, referring back to educational controlling systems using high-stakes testing, these findings indicate a potential benefit in using competence-orientated test tasks such as those from international comparative studies in national high-stakes tests.

Finally, we would like to introduce the idea of incorporating central examination questions of previous years into the national options of international school assessment studies, so that both test formats, although having different basic functions, might become better co-ordinated or even linked together in their intention to measure mathematical competencies.

References


