Studies in Educational Evaluation 47 (2015) 58–67

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Comparing apples with oranges? An approach to link TIMSS and the National Educational Panel Study in Germany via equipercentile and IRT methods

Annika Nissen [a,*], Timo Ehmke [a], Olaf Köller [b,c], Christoph Duchhardt [d]

[a] Leuphana University of Lüneburg, Scharnhorststraße 1, 21335 Lüneburg, Germany
[b] IPN – Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany
[c] Centre for International Student Assessment, Affiliated Institute of the Technical University Munich, Marsstraße 20-22, 80333 Munich, Germany
[d] University of Bremen, Bibliothekstr. 1, 28359 Bremen

ARTICLE INFO

Article history:
Received 20 January 2015
Received in revised form 10 July 2015
Accepted 14 July 2015
Available online 26 August 2015

Keywords:
Student assessment
Equating methods
Linking study
Mathematics competencies
Large scale assessment
TIMSS

ABSTRACT

The aim of this linking study is to classify the National Educational Panel Study (NEPS) according to the criterion based reference frame of the Trends in International Mathematics and Science Study (TIMSS). One requirement for a strong linking is a high conceptual overlap of the frameworks. The results of the comparison between NEPS and TIMSS indicate a high overlap. To link the studies the equipercentile equating and an IRT linking approach are compared. The results show that both methods showed similar descriptive statistics and satisfying classification consistency. Nevertheless, the equipercentile equating has small advantages and is therefore more suitable. Thus, inferences are possible on a population level, but the results should not be reported or interpreted on an individual level.

1. Introduction

There are many national and international large-scale assessment studies which measure mathematics competencies of primary and secondary school students as, for example, the *Trends in International Mathematics and Science Study* (TIMSS; Mullis, Martin, Ruddock, O'Sullivan & Preuschoff, 2009) and the *National Educational Panel Study* (NEPS; Blossfeld, von Maurice & Schneider, 2011) in Germany. Increasingly, national and international assessments have been linked with the aim to explain differences in the results of the corresponding studies (e.g., Grønmo & Olsen, 2007; Neidorf, Binkley, Gattis & Nohara, 2006; Nohara, 2001; Wu, 2010) or to use the benchmarks in other national studies (e.g., Cartwright, Lalancette, Mussio & Xing, 2003; National Center for Education Statistics, 2013; Phillips, 2007).

In Germany, the Standing Conference of the Ministers of Education and Cultural Affairs of the Laender in the Federal Republic of Germany (KMK) calls for a link between national and international assessment studies (KMK, 2006). However, it is not feasible to simply compare the assessments and to relate the results accordingly for several reasons. Amongst others, the assessments are based on different frameworks and the results are not reported on the same scales. To enable a valid comparison of the findings, instruments from both studies have to be linked to a common scale.

This study examines a linkage between the mathematics tests from TIMSS 2011 and NEPS 2010 that were developed for the same age cohort (students at the end of grade 4/beginning of grade 5). NEPS is a longitudinal study in Germany which analyzes educational processes over the lifespan. In NEPS, no proficiency levels are defined. So the results of NEPS cannot be interpreted in a criterion-based manner. TIMSS 2011 is an international study which assesses mathematics and science achievement of students with the overarching goal of improving teaching and learning. In TIMSS, achievement is reported at four proficiency levels (international benchmarks). These benchmarks describe what students typically know and can do in mathematics (Mullis, Martin, Foy & Arora, 2012). Linking NEPS with TIMSS has the advantage that the results of NEPS could be classified in the international and criterion-based reference frame of TIMSS. By this

* Corresponding author.
E-mail addresses: anissen@leuphana.de (A. Nissen), tehmke@leuphana.de (T. Ehmke), koeller@ipn.uni-kiel.de (O. Köller), christoph.duchhardt@uni-bremen.de (C. Duchhardt).

http://dx.doi.org/10.1016/j.stueduc.2015.07.003
0191-491X/

means, linking could extend the interpretations of the NEPS scores. For example, students who do not exceed the TIMSS low international benchmark, or who reach the advanced international benchmark, could be examined longitudinally within NEPS.

Accordingly, 733 fourth graders took both test forms, the TIMSS mathematics test for the end of grade 4 and the NEPS mathematics test for the beginning of grade 5. Due to the fact that "different methods and different groups do not produce identical 'equatings'" (Lord & Wingersky, 1984, p. 455), we compared two different linking approaches. Furthermore, different linking methods are based on different assumptions. Hence, using two methods and comparing the results is a form of quality control (Kolen & Brennan, 2010). In this study, the equipercentile equating and item response theory (IRT) scale transformation were applied to find out which method fits the data better concerning (1) descriptive statistics such as means, standard deviations, skewness, and kurtosis and (2) classification accuracy according to the TIMSS international benchmarks. One requirement for a strong linking is a high conceptual overlap of the studies, while the utility and reasonableness of a linking can be influenced by the degrees of similarity (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Kolen & Brennan, 2010; Linn, 1993; Mislevy, 1992). Therefore, the frameworks and test specifications of the two studies will be compared beforehand.

In Section 2, after a short description of the TIMSS and NEPS frameworks and test specifications, an introduction to linking methods will be given. Afterwards, we will focus on the current state of research, giving an overview of some studies which concentrate on different aspects of linking. In Section 3, the research questions will be formulated. In Section 4, the methods of the linking-study will be presented. Section 5 is devoted to the results. In a first step, we will look at common features and differences in the frameworks and test specification of the two studies. After this, the results of the linkings will be presented, first for the equipercentile linking and then for the IRT linking. Afterwards, we will compare the outcomes, going on to demonstrate the results of the conversion. In Section 6, the results will be interpreted and discussed.

2. Theoretical background

2.1. About TIMSS

In 1995, the Third International Mathematics and Science Study (TIMSS; current: Trends in International Mathematics and Science Study; Mullis et al., 2009) was conducted by the International Association for the Evaluation of Educational Achievement (IEA) for the first time. TIMSS 2011 is a cross-sectional study which assesses mathematics and science achievement of students in grades 4 and 8. As it was repeatedly conducted after 4 years, it created the opportunity to measure trends across age cohorts in the educational systems. The overall aim of the study is to compare the countries' outputs of their educational system and to identify the influencing factors. Thus, it means to explain the various countries' differing results, and to identify possibilities which could support the improvement of the respective educational systems. In 2011, more than 60 countries participated in TIMSS.

In addition to the assessment instruments in mathematics and science, students' social background and other characteristics are collected via questionnaires. The basis for the frameworks of the tests is formed by the curriculum model which consists of three aspects: (1) what the students in the countries are expected to learn (intended curriculum); (2) what the students actually learn in the classrooms (implemented curriculum); and (3) the final outcome (attained curriculum).

2.1.1. TIMSS 2011 – mathematics test for fourth graders

The TIMSS 2011 assessment for fourth graders took 72 min and contained 177 mathematics items and 175 science items. The participation for the selected sample was obligatory. The items were included in 14 booklets, each one with four blocks, i.e., two mathematics and two science blocks. IRT was used to scale the data and develop the reporting scales. For the constructed responses (46.9%) a two-parameter model was used, for the multiple-choice items (53.1%) a three-parameter model was applied. Additionally, a partial credit model was employed for nine items. In order to estimate the students' personal proficiency, the test instruments of TIMSS apply the plausible value approach (Foy, Brossman & Galia, 2012).

The mathematical framework differentiates two domains, the content and the cognitive domain. The content domain includes the subdomains number (50%), geometric shapes and measures (35%), as well as data display (15%). The cognitive domain differentiates between knowing (40%), applying (41%), and reasoning (19%) (Foy et al., 2012).

2.2. About NEPS

The National Educational Panel Study (NEPS) is a longitudinal study in Germany. The aim is to measure competencies over the lifespan (from early childhood to late adulthood) and describe the education processes and the developmental trajectories (Blossfeld et al., 2011). The main interest is to measure the development of competencies and ascertain possible aspects of impact. Therefore, NEPS uses a multi-cohort sequence design. There are six cohorts, each starting at different stages of transition in the educational system: newborns, four-year-olds, fifth graders, ninth graders, university freshmen to higher education students, and adults aged between 25 and 65 years. Four starting cohorts began in 2010, one in 2009, and one in 2012. The participants will be followed across their lifespan.

NEPS was initiated and funded by the German Federal Ministry of Education and Research (Blossfeld, von Maurice & Schneider, 2009). For the long-term organization of the study, the *Leibniz Institute for Educational Trajectories* (LIfBi) was established in 2014. NEPS assesses four competence areas (Weinert et al., 2011): (1) domain-general cognitive abilities and capacities; (2) domain-specific cognitive competencies (German language, mathematics, and science competencies); (3) meta-competencies and social competencies; and (4) stage-specific attainments, skills, and outcome measures.

2.2.1. NEPS 2010, grade 5 – mathematics test

In our equivalence-study, we used the NEPS mathematics test for fifth graders implemented at the beginning of their school year. The participation was optional. The NEPS mathematics test is based on the concept of mathematical literacy, known, for example, from PISA (OECD, 2010). The framework differentiates between two domains: the content areas as well as mathematical and cognitive processes. The content areas include the subdomains quantity (33%), change and relationship (25%), space and shape (21%), and data and chance (21%). The mathematical and cognitive processes differentiate between argumentation, communication, modeling, problem solving, representing, and applying technical skills (Weinert et al., 2011).

For fifth graders, the test takes about 30 min including 25 items. The NEPS test is scaled based on the IRT (Duchhardt & Gerdes, 2012). A one-parameter Rasch model is used. In NEPS, students' Weighted Likelihood Estimates (WLE) scores are calculated as estimates for the students' achievement scores. The mathematics items are multiple choice (13 items), complex multiple choice