Are the tests scores of the Programme for International Student Assessment (PISA) and the National Educational Panel Study (NEPS) science tests comparable? An assessment of test equivalence in German Schools

Helene Wagner*, Inga Hahn, Katrin Schöps, Jan Marten Ihme, Olaf Köller

IPN - Leibniz Institute for Science and Mathematics Education at Kiel University, Olshausenstraße 62, D-24118 Kiel, Germany

ARTICLE INFO

ABSTRACT

The aim of this study is to link the science scale of the German *National Educational Panel Study* (NEPS) with the science scale of the *Programme for International Student Assessment* (PISA). One requirement for a strong linking of test scores from different studies is a sufficient similarity of the tests regarding their constructs. The present study aims to assess the similarity of the operationalized constructs of the NEPS and PISA scientific literacy tests with the aim to link the scales of the two tests. A linking study was carried out for this purpose in which 1079 students worked on the tasks of both studies. The results of the comparison between NEPS and PISA indicated a high overlap regarding their constructs. However, both studies deal with missing responses differently. The linking via equipercentile equating showed a high classification consistency which was highest when missing responses were ignored in both studies.

1. Introduction

In 1997, the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany decided on Germany's regular participation in international large-scale-assessments. Germany participates in the *Trends in International Mathematics and Science Study* (TIMSS) every four years at the end of primary school as well as in the *Programm for International Student Assessment* (PISA) every three years at the end of lower secondary education.

However, these studies only allow cross-sectional analyses and only address specific age groups. Until recently, no large-scale study measuring the development of competencies over the lifespan had been carried out in Germany. The *National Educational Panel Study* (NEPS; Blossfeld, 2008) which started in 2009 is the first German attempt to close this gap by assessing the development of skills and competencies over the lifespan (Hahn et al., 2013). NEPS strives to connect with national and international large-scale assessment studies to achieve a common interpretation of scores (Blossfeld, 2008). However, comparing test results from different studies is a challenge because they are based on different frameworks and their results are not reported on the same scale. Therefore, the test instruments have to be linked to a common scale.

This study examines the comparability of the Grade 9 NEPS science test with the PISA science test. Connecting both tests could extend the interpretation of their test scores. Until now, no proficiency levels have been defined in NEPS. Hence, the NEPS results cannot be interpreted and reported in a criterion-based manner. The link between the NEPS and PISA tests can allow for classification of the NEPS test scores in the criterion-based international reference framework of PISA which is well established in the public educational debate in Germany. The longitudinal design in NEPS could help to identify the determinants of competence acquisition which can predict the performance in the PISA test. Furthermore, the link between NEPS and PISA could be used to investigate in NEPS samples to what extent the performance on the international PISA scale can predict success in upper secondary education and the further professional career.

Due to the fact that the NEPS and PISA studies deal differently with missing responses, we also investigated how the different treatments of missing values affect the comparability of the test scores and the quality of the linking. To link the NEPS and PISA tests to a common scale, 1079 9th grade students took both tests in a linking study.

According to Kolen and Brennan (2004) the linking of test scores from different studies requires sufficient similarities of the tests with regard to:

- Inferences: To what extent are the scores for the two tests used to draw similar types of inferences? In other words, to what extent do the two tests share common measurement goals?
- Populations: To what extent are the two tests designed for testing similar populations?
- Characteristics and conditions of the measurement: To what extent do the two tests share common measurement conditions, for example, with regard to test format, administration conditions, test length, etc.?
- Constructs: To what extent do the two tests measure the same construct?

To closely examine these aspects, the next section will look into the similarities and differences of the NEPS and PISA frameworks.

2. Comparing the scientific literacy tests of NEPS and PISA

2.1. Inferences

NEPS and PISA have different objectives. The aim of PISA is to monitor educational systems at the end of lower secondary school in terms of student performance (OECD, 2013). This goal is realized every three years by a cross-sectional overview of the educational level of 15-year-old students. The aim of NEPS is to provide longitudinal data of the competence development from early childhood to late adulthood in Germany. In order to achieve this goal the data collection in NEPS is embedded in a multicohort sequence design (von Maurice, Sixt, & Blossfeld, 2011) which makes it possible to compare the educational level of 9th grade students from different cohorts. In other words, despite the different objectives of NEPS and PISA the measurements of these studies allow to assess the educational level of students at the end of lower secondary school.

2.2. Target populations

The target population of the NEPS test are 9th grade students (von Maurice et al., 2011). PISA examines the competence of 15-year-old students (15 years and 3 months to 16 years and 2 months of age). In Germany the target population of 15-year-old students for PISA 2012 was defined as the persons born in 1996. The analysis of the composition of the PISA sample in Germany in 2012 showed that 48% of the selected students attended Grade 9, 33% of them attended Grade 10 and 11, and 19% of them attended Grade 7 and 8 (Sälzer & Prenzel, 2013). Hence, the target populations in NEPS and PISA are not identical, but the overlap of both selected samples is high.

2.3. Characteristics and conditions of the measurement

PISA is a cross-sectional study which in 2012 assessed mathematics, reading, science and financial literacy of 15-year-old students (OECD, 2014b). The 53 items of the science test were split into three clusters and presented to students with seven mathematics clusters and three reading clusters in thirteen test booklets. Each booklet consisted of four clusters with each cluster representing 30 min of test time. Each student worked on one to two science clusters so that each item was processed by a sufficient number of students.

NEPS provides longitudinal data on educational processes and competence development in information and communication technologies, mathematics, reading and science (von Maurice et al., 2011). The 28 items of the NEPS science test were presented in 2010 in 28 min and each person got the same items in a fixed sequence (Schöps & Saß, 2013).

In Germany data collection and processing for PISA 2012 and NEPS 2010 were coordinated by the IEA Hamburg. Both tests examined in this linking study were administered as a paper pencil test. The majority of the items in NEPS 2010 and PISA 2012 had a closed-constructed

response format (OECD, 2014a; Schöps & Saß, 2013). However, PISA 2012 also used an open-constructed response format (32% of the total number of items).

NEPS and PISA deal differently with missing responses. PISA 2012 used a two-stage procedure for handling missing responses (OECD, 2014b): in the first step, not–reached and not valid items were ignored and omitted items were scored as incorrect when estimating the item parameters. In the next step, the estimated item parameters were used for the estimation of person parameters where missing responses were scored as incorrect. In contrast, NEPS 2010 ignored all missing responses for the estimation of item and person parameters (Pohl & Carstensen, 2012).

A number of studies (De Ayala, Plake, & Impara, 2001; Pohl, Gräfe, & Rose, 2014; Rose, von Davier, & Xu, 2010) showed that scoring missing responses as incorrect leads to a bias in the estimation of parameters and to the overestimation of the reliability. Based on these results we assume that ignoring the missing responses in NEPS and PISA will increase the comparability of their test scores (hypothesis two) and their scales (hypothesis four), and hence the quality of linking (hypothesis six).

2.4. Operationalized constructs: comparing the contents of the science tests of NEPS and PISA

The definition of scientific literacy used by NEPS includes aspects of the *concept of competence* as defined by Weinert (2001), and of the *concepts of scientific literacy* developed by the American Association for the Advancement of Science (American Association for the Advancement of Science, 2009) and by PISA (OECD, 2006). Therefore, the NEPS scientific literacy framework has a substantial overlap with the scientific literacy framework from PISA 2012 (Fig. 1).

Fig. 1 shows that the frameworks of both studies differ in the number of components used for assessing scientific literacy: The framework of NEPS only considers the *content-related components* which are related to the *knowledge of science* (KOS) in PISA, and *process-related components* which are related to the *knowledge about science* (KAS) in PISA. The PISA framework differentiates further and also distinguishes between the competencies *identifying scientific issues*, *explaining phenomena scientifically* and *using scientific evidence*. At this point, it can be concluded that the frameworks of the two studies differ in their conceptual scope. But how different are they on the task level?

This question can be examined using the theory of bias and equivalence of van de Vijver (1998), which was modified for the purposes of equivalence research in the linking studies by Pietsch, Böhme, Robitzsch, and Stubbe (2009). Pietsch et al. (2009) suggested assessing the similarity of the operationalized constructs of two tests by regarding their conception, their dimensional structures and their scales.

According to Pietsch et al. (2009) two tests are equivalent regarding their conception when their constructs have equivalent frameworks. In order to analyze the conceptual equivalence of the scientific literacy in NEPS and PISA (Wagner, Schöps, Hahn, Pietsch, & Köller, 2014) seven experts in the field of science didactics familiar with large-scale assessments classified the NEPS items according to the categories of KOS and KAS and to the competencies in PISA. The results showed that 79 percent of the NEPS items could be assigned to the contents of the PISA framework. However, according to five of the seven raters some of the KOS components in PISA (earth and space systems and technology systems) were not covered by NEPS items.

3. Linking methods and linking studies

Different methods of linking can be applied depending on the level of equivalence of the two tests. Mislevy (1992) and Linn (1993) differentiate between five types of linking: *equating, vertical scaling, concordance, projection* and *moderation*. Fig. 2 illustrates the different linking methods which, in terms of their applicability depend on the