Large-scale Assessments in Education

RESEARCH Open Access

Comparability of teachers' educational background items in TIMSS: a case from Turkey

Elif Oz*

*Correspondence:
eoz2@nd.edu
Institute of Educational Initiatives, University of Notre Dame, 307 Carole Sandner Hall, Notre Dame, IN 46556, USA

Abstract

Large-scale international assessment studies such as Trends in International Mathematics and Science Study (TIMSS) or Programme for International Student Assessment (PISA) provide researchers and policy makers the opportunity to conduct secondary analyses to answer questions related to educational outcomes and compare the impact of certain inputs on student outcomes across countries. These comparisons are made under the assumption that the questionnaire items translated to different languages are understood in the same way by its participants. Presenting a case from Turkey, this paper shows that equivalency of questionnaire items is not always achieved. The case explores demographic information related to teacher preparation and the sample is drawn from eighth grade science and mathematics teachers participated in TIMSS 2007, 2011, and 2015 in Turkey. Descriptive analysis of data collected from these teachers and comparisons across subjects and years show that teachers may have misunderstood a question regarding their major, thus limiting potential claims related to teacher preparation in Turkey. Researchers and policy analyst who use secondary data collected by international assessment studies should be aware of such comparability issues in adapted items prior to conducting any secondary analyses.

Keywords: International large-scale assessments, Item adaptation, Translation, Contextual background questionnaires, TIMSS

Introduction

Large-scale international assessment studies such as Trends in International Mathematics and Science Study (TIMSS) or Programme for International Student Assessment (PISA) are critical evidence for charting educational progress as well as in shaping educational policies of countries (Ababneh et al. 2016; Klemencic 2010; Lockheed and Wagemaker, 2013; Paine and Zeichner 2012; Sjoberg 2015; Tobin et al. 2015). Germany's adaptation of new science standards at the national level in 2004 (Naumann 2005; Steffen and Hößle 2014), Jordan's development of new teacher training guidelines, and New Zealand's efforts to improve teachers' instruction in science after poor TIMSS and PISA results are only a couple of examples of policy changes based on international assessments (Ababneh et al. 2016; Tobin et al. 2015).

In addition to measuring students' performance in reading, mathematics, and science, large-scale international studies also collect a variety of background information from participating students and their parents, teachers, and school administrators. The rich collection of information enables policy-makers and researchers to conduct secondary analyses to understand the ways and extent to which various factors are related to important student outcomes such as scientific literacy or self-efficacy in mathematics and how these associations change across countries and over time (Anderson et al. 2007; Ferrini-Mundy and Schmidt 2005; Fertig 2003). Moreover, TIMSS and PISA produce publicly available, easily-accessible, and nationally representative educational data (Anderson et al. 2007). TIMSS and PISA are also well known by the general public since the results of these studies are used by the mass media (Dominguez et al. 2012; Sjoberg 2015). These features make the data collected by these large-scale international studies widely-used (Bonett 2002; Grek 2009; Klemencic 2010) and thus, they are critically important for their accuracy.

Because of their critical importance and widespread use, the demographic items developed by the international assessment studies, like all other items, valid representations of the context of participating countries is imperative. To collect information that can be comparable across countries, background surveys are developed in one or two languages and then the original items are adapted to different cultures and languages (Ebbs and Korsnakova 2015). The international studies put great effort into ensuring equivalent adaptations of background items, however there are cases where major differences in adaptations occur which result in variations in responses across countries or cultures (Bray et al. 2020; Sjoberg 2015). On the other hand, researchers may have the tendency to taking the comparability of the background items as granted (Avvisati et al. 2019; Bray et al. 2020). This paper presents one case from TIMSS in which survey question choices and responses highlight items that are "lost in translation", thus raising at least some questions about the broader validity for making high-stakes decisions.

Although there are multiple papers that discuss these adaptation problems, very few of them used actual data from large scale international studies to detect possible interpretation differences across different versions of tests or surveys (e.g. Arffman 2010, 2012; Bray and Kobakhidze 2014; Bray et al. 2020; Ercikan 1998). Moreover, most of these studies focus only on translation of items developed to measure a psychometric construct (Arffman 2010, 2012; Ercikan 1998). There are very few studies that investigate comparability of background questionnaire items and open the validity of the information collected through these background items to discussion (see Bray and Kobakhidze 2014; Bray et al. 2020).

Drawing from Bray et al. (2020) recent study that showed major adaptation differences in different versions of items related to students' outside-of-school educational activities in PISA that resulted in variations in interpretations of these items across countries, this paper presents a case from TIMSS where differences in meaning between the Turkish and English versions of a teacher questionnaire item in TIMSS resulted in inconsistent responses from the participating science and mathematics teachers in Turkey. Complementing the findings of Bray et al. (2020), this paper contributes to the discussion related to item adaptations of background questionnaire items by providing an example where translation issues may result in differences in interpretations of a background questionnaire item in TIMSS. This paper also contributes to the discussions related to "data literacy" of secondary data users by providing evidence for the existence of translation issues even for the items that are designed to measure "objective traits" of participants (Avvisati et al. 2019; Bray et al. 2020).

The case explores demographic information related to teacher preparation collected in TIMSS 2007 and 2011 that was originally intended to be used to investigate the relationship between science teachers' professional preparation and their students' achievement in science in Turkey. The sample for the case under discussion is drawn from eighth grade science and mathematics teachers whose students participated in TIMSS 2007, 2011, and 2015 in Turkey. As a proof of concept, descriptive analysis of data collected from these teachers and comparisons across subjects and years along with responses that do not reflect the teacher preparation policy context in Turkey show that it is very likely that teachers misunderstood a question regarding their post-secondary major, thus limiting potential claims related to teachers' background knowledge and student outcomes in Turkey. This paper also informs different stakeholders regarding inconsistencies in teachers' responses to the educational background item used in TIMSS 2007, 2011, and 2015 Teacher Contextual Background Questionnaires in Turkey which may be possibly due to translation issues in the Turkish version of the questionnaire.

In the following sections, I discuss the literature on validity of adapted items and bias related to the poor adaptations of items developed by international assessments and summarize the background questionnaire item adaptation process followed by TIMSS. Then, I provide contextual information about Turkey's teacher preparation and certification policies that are helpful to understand the roots of the problem with the construction and translation of the teacher preparation items to Turkish. I present and discuss the discrepancies and anomalies that I observed in teachers' responses while conducting a study that was designed to investigate the relationship between science teachers' professional preparation and their students' science achievement. Based on the comparisons of the original and adapted items in TIMSS 2007 and 2011 and teacher preparation context at the time of data collection, I explain these discrepancies in data with adaptation problems related to the Turkish versions of the item. I conclude by discussing the importance of taking the context of the country into account in adaptattions of contextual background items that are developed through large-scale, international studies and by providing recommendations for international survey developers and researchers who intend to use contextual data collected through large-scale international surveys.

Validity of adapted items in large-scale international assessment studies

Before delving deeper into the literature on validity of adapted items used in international studies, I would like to provide definitions for translation and adaptation of items. In the international assessment literature, the terms translation and adaptation refer to different concepts although they can also be used interchangeably. In their guidelines for translating and adapting tests, The International Test Commission (ITC 2017) uses the term test adaptation to refer to a set of "activities" to "mov[e] a test from one language and culture to another" (p. 3). These activities include not only the translation process but also ensuring the equivalency of items and validity of the constructs or adapting the format of the test to the target language or culture. Translation refers to a much "simplistic approach to transporting a test from one language to another with no regard for educational or psychological equivalence" (ITC 2017, p. 3) In this study, the threat to validity of poorly adapted items will be discussed which will also includes poor translation of tests.

Validity of a test or an item is an essential concept in test construction and psychometric measurement. Although it has different types in test construction, in general terms, validity is defined as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of conclusions drawn from some form of assessment" (American Psychological Association (APA), 2021). In international assessment studies where assessments are constructed in different languages and cultures, the adequacy and appropriateness of conclusions drawn should not depend on the language or the culture the assessment is constructed and administered. This can be achieved by making sure that the tests developed in different languages and cultures measure the same constructs which is called item equivalency. When item equivalency between different versions of an item is not achieved, it is called a "translation error" but it is an overlooked issue in the field of large-scale international assessments (Solano-Flores et al. 2009, p. 80).

Developing items that are equivalent in meaning in different languages and different cultures is one of the main concerns to ensure validity of items developed for an international study. Translation is necessary but not sufficient to create equivalent items in different languages or cultures. According to Solano-Flores et al. (2009), translation errors are inevitable and capturing the same meaning in two versions of an item is theoretically impossible. Therefore, there is some room for translation error that can be acceptable. On the other hand, these errors should not cause any item bias which may be detrimental to the validity of the test administered in different cultural settings (Solano-Flores et al. 2006). Moreover, in addition to accurate translations, item adaptations are needed to ensure that the item implies the same meaning regardless of the language the item was administered (Solano-Flores 2019).

In general, bias in cross-cultural surveys "occurs if score differences on the indicators of a particular construct [...] do not correspond to differences in the underlying trait or ability [...] (van de Vijver and Tanzer 2004, p. 120). Therefore, it is the opposite of test equivalence. For a measurement to be valid, the test should be equivalent across different cultures and therefore, any biases that may prevent equivalency of items should be investigated (van de Vijver and Tanzer 2004). One type of such biases is called "item bias" which refers to an item's inability to measure the same attribute across different groups (Mellenbergh 1989). Although it is not the only reason, poorly translated or adapted items is one source of item bias (van de Vijver and Poortinga 1997). However, poor translation can also be a byproduct of vagueness in meaning in the original item itself which may also cause item bias (van de Vijver and Tanzer 2004).

Constructing and validating an item and then translating it to another language may increase the possibility of differential item functioning (van de Vijver and Poortinga 1997). In the context of large-scale assessments that are administered across multitude of countries and therefore in multitude of languages, it is not possible to simultaneously construct the items in different languages. Therefore, items designed and adapted by large-scale international assessment studies are more prone to item bias due to errors in translation and adaptation of items and careful attention should be paid to the adaptation process. When it comes to errors in translation of items, Solano-Flores et al. (2009)

group the sources of translation errors in large-scale international assessments into 10 categories: "(1) style, (2) format, (3) conventions, (4) grammar and syntax, (5) semantics, (6) register, (7) information, (8) construct, (9) curriculum, and (10) origin" (p. 82). Covering what these dimensions refer to is outside of the scope of this paper as not all of the dimensions are applicable to adaptations of contextual items which are not designed to measure an aptitude. However, grammatical errors can be problematic in adaptations of contextual background questionnaires where incorrect use of grammar or unusual use of words may make the question susceptible to misunderstanding. Another source of error applicable in adaptations of background items are semantic errors which refers to misalignment in meaning between the original and adapted item. Using words that do not reflect the context (i.e. register) is another source of translation error that is applicable to adaptations of contextual background questionnaires.

Equivalency of items adapted to different cultures by international large-scale surveys

In the field of test development, the term *item adaptation* is used in lieu of item translation because simply translating an item does not guarantee whether an item produces similar responses for individuals taking the test in different languages (Geisinger 1994). For the responses to be comparable, individuals that have similar cognitive ability should respond to the question in similar ways regardless of characteristics of test takers such as gender, ethnicity, or the language the test was taken in. This is called *item equivalency*. Although this is a term generally used for psychometric test construction, it is also important for adaptations of survey items that are designed to collect accurate demographic information from participants (Bray et al. 2020; van de Vijver and He 2016).

Adapting items to various languages as well as different cultures or contexts so that the item could be interpreted by all participants as intended is a daunting task. The challenges in measuring constructs or collecting information that can be comparable across different countries or populations are discussed in the literature (e.g. Arffman 2012; Avvisati et al. 2019; Bray et al. 2020; Rutkowski and Rutkowski 2019). Studies show that equivalency of items is not always achieved by large scale international surveys and translation issues exacerbate the problem (Arffman 2012).

The majority of the studies that discuss adaptation problems of international studies focus only on participants' responses to questions that are designed to measure a certain construct such as reading comprehension or scientific literacy (Arffman 2012; Ercikan 1998; Geisinger 1994). These studies investigate differences in responses of students from different countries and cultures to the items that are designed to measure a psychometric construct. To do so, researchers use statistical procedures to detect possible translation issues. A considerable amount of bias detected by these studies are found to be related to differences in languages which also includes errors in translation of items (Arffman 2012).

Ercikan (1998) used differential item functioning (DIF) to test the equivalence of the achievement items developed in English and then adapted to French for students who participated in the Science Study of 1984, conducted by the Institute of Educational Assessment (IEA) in Canada. Differential Item Functioning (DIF) is predominantly used as a tool to test to what extent an item provides different results for certain groups of test takers (Clauser and Kathleen 1998). It may signal potential translation problems if there are different patterns in students' responses who take the test in different languages but who are expected to perform similarly.

Using DIF, Ercikan (1998) reported that 18 items, 26% of the whole instrument, showed some red flags for potential translation problems. Judgmental reviews of these items revealed that translation problems are the possible reason for the observed DIF for eight of these 18 flagged items. In a similar study, this time using TIMSS 1995, Ercikan (2002) reported adaptation problems with one fifth of the items in Mathematics and Science Tests.

A similar analysis using the International Adult Literacy Survey responses in English and French versions showed that the differences in item difficulty between the individuals who take the different versions of the survey can be attributable to the translation issues (Blum et al. 2001). Investigations of the items that are problematic in item difficulty revealed translation errors that changed individuals' understanding of the question (Blum et al. 2001).

Comparing English and Finnish versions of three texts used in PISA 2000 Reading tests, Arffman (2010) revealed that none of the text pairs were equivalent to each other. She identified six common translation issues that might resulted in variations in difficulty of the texts. Later on, she compared English and Finnish versions of two additional texts from PISA 2000 and report differences in language as well (Arffman 2012).

Although there are studies and papers that investigate and discuss the translation and adaptation issues with the international large-scale assessment studies, very few of them specifically focus on the items developed to collect contextual background information. However, the comparability of contextual background information collected are also crucial for test developers, policy-makers, and researchers (Kirsch and Braun 2020; Rutkowski and Rutkowski 2019; van de Vijver and He 2016).

For example, in TIMSS and PISA, students' performance is predicted by administering only a portion of the test items to students. Students' responses to the rest of the test items are estimated using their responses to the questions presented to them, in addition to the background information collected from them that are known to be associated with student achievement (IEA 2011; OECD 2009). Any misinterpretations of the background questionnaire items may result in inaccurate predictions of student achievement (Rutkowski and Rutkowski 2010). Moreover, background questionnaires also serve the purpose of "contextualizing education" as it makes investigating the relationship between contextual differences and student outcomes possible (Rutkowski and Rutkowski 2010, p.425). Therefore, possible misinterpretations of the items designed to collect contextual information from the students, parents, teachers, and administrators are worth being investigated (van de Vijver and He 2016).

The studies that focus on contextual background items developed by international large-scale assessments are scant. One background information where item adaptations are considered problematic and investigated in depth is shadow education, supplemental education that students receive outside of school (Bray and Kobakhidze 2014; Wolf 2002). Varying shadow education practices across countries makes adaptations of items related to shadow education particularly challenging. For instance, in Japan, the term "juku" refers to the institutions that provide supplementary instruction outside of school

(Wolf 2002). Therefore, when the word “juku” is used in Japanese version of an item related to shadow education, students understand what is meant more easily than their counterparts in other countries who are not familiar with such institutions. Therefore, adapted versions of such items could be interpreted differently by students in countries such as Latvia or Slovak Republic where shadow education practices are not common (Bray and Kobakhidze 2014). For instance, Wolf (2002) had to exclude Colombia in his study on shadow education due to problems in translation of the item that asks about students’ participation in shadow education. However, he did not provide any further information about the details of the translation problem or why they thought there was a translation problem. High rates of shadow education participation in TIMSS 1995 reported by Colombian students that exceeded even that of Japan and Korea could likely be the reason (Bray and Kobakhidze 2014).

In addition to the translation issues due to the variations in shadow education practices across countries, Bray et al. (2020) show that item adaptation issues across countries may contribute to differences in students’ interpretation of items in different languages. Back-translating the adapted PISA items related to outside-school-time educational activities in nine different languages shows that there may be differences in intended and “received” meanings across different versions of the same item (Bray et al. 2020, p.93). The authors conclude that these differences in interpretation make the data collected using these items incomparable across countries.

No other studies that used international large-scale assessment data reported any problems with the adaptations of the contextual background questionnaires, to my knowledge. This may be partly related to the lack of statistical tests that can be used to detect problems in background questionnaires. Detecting adaptation problems in the background questionnaire items could be possible if the responses of the participants do not make sense in the context of the country which the item is adapted to. To illustrate this, I present a case from TIMSS where I detect a discrepancy in teachers’ responses to the item about their post-secondary education in Turkey, which is possibly due to weak adaptations of the item to Turkish. Before presentation of the case, I summarize common adaptation practices used in TIMSS.

Adaptation process for contextual questionnaires in TIMSS

According to the Methods and Procedures document prepared by the TIMSS team, adaptation of contextual questionnaires go through the same diligent process as the assessment instrument (Ebbs and Korsnakova 2015). Each participating country has their own National Research Coordinators (NRCs) who are responsible for the primary adaptations of items to the language used in instruction in the country.

The NRCs are required to follow the adaptation guidelines disseminated by TIMSS including certain qualifications that the translators should have. Having extensive knowledge in both English and the target language and being well-versed with the culture of the target country are examples of such qualifications. Based on the guidelines, some terms could be adapted to the country’s context if needed. For this purpose, “[t] he guidelines for translation and adaptation provide countries with detailed descriptions of the intent of each required adaptation to clarify the meaning of the terms used and to enable the translators to select the appropriate national term or expression to convey the intended meaning" (Ebbs and Korsnakova 2015, p. 7.11). For instance, for the term "professional development" in English, the countries are required to change the term to something that would refer to "the supplemental training provided to teachers during their professional careers" (Ebbs and Korsnakova 2015, p. 7.11). The NRCs are also responsible for documenting their adaptation process by developing national adaptation forms specifically for the contextual background questionnaires. However, these national adaptation forms are not publicly available.

After the NRCs complete their adaptation processes the adapted items go through a verification process conducted by TIMSS and an external agency to assess the equivalency of the items (Ebbs and Korsnakova 2015). As in the case of translators selected by the participating countries, the verifiers are required to meet certain qualifications such as having "university-level education" or "mother tongue proficiency in the target language" (Ebbs and Korsnakova 2015, p. 7.13). However, residing in the participating country of interest is not one of the required qualifications. In cases where the verifiers do not reside in the target country, being in "close contact with the country and its culture" would be sufficient (Ebbs and Korsnakova 2015, p. 7.13).

Despite the guidelines provided and diligence in working on the adaptation of items administered in international large-scale assessment studies, there may remain issues with inappropriate item adaptations to different cultures (Hambleton 2002). This may be especially problematic in international large-scale studies where the data are used for comparisons across countries or for high-stakes decision-making. Next, I present a case for item adaptation issues that may arise in background questionnaires of international assessment studies.

A case from TIMSS: teacher preparation and student achievement in Turkey

In Turkey, there are two different routes to become a science teacher at the middle school level. One is through graduating from a four-year-long science teaching in middle grades program offered by colleges of education in Turkey. The other one is through a-year-long pedagogical training program which can be completed during or after getting a bachelor's degree in science (e.g. biology, chemistry, physics, Talim ve Terbiye Kurulu Baskanligi (TTKB) 2014).

There are major differences between the two types of teacher preparation routes in terms of the number and intensity of pedagogical and subject matter courses, which I explain in more detail in the following section. Due to these differences, there is some discussion around the quality of the programs and their graduates in Turkey (Azar 2011; Gurol et al. 2018). Most of the studies designed to investigate this difference in teacher quality are confined to the opinions of the program participants and faculty about the quality of these two programs (Gurol et al. 2018). There are very few studies that investigate the relationship between the teacher preparation paths and student achievement in Turkey using empirical evidence (Abazaoglu and Tasar 2016; Atar 2014; Yildirim 2013).

The issue of translation arose when I conducted a study to investigate any differences in students' science achievement related to the teacher preparation programs that their science teachers attended in Turkey. To answer my research question related to teachers' majors, I conducted a regression analysis where I included students' science achievement as the outcome variable and their science teachers' post-secondary degree major as the key explanatory variable, controlling for other variables that are known to be associated with student achievement. TIMSS data were the only representative data in Turkey that were available to conduct this analysis. I chose to use the TIMSS 2011 data set since it was the most recent one that was available at the time.

Before I present the problem with the adaptation of the teacher preparation item undur discussion to Turkish, I provide some additional background information about teacher preparation and certification policies in Turkey which might be useful to understand the discrepancies in teachers' responses.

Teacher preparation and certification policies in Turkey

As previously mentioned, there are two different pathways to become a middle grades science teacher in Turkey. One path is to graduate from a middle grades science teaching program offered by a school of education. This is a four-year-long program where teachers take pedagogical and subject-matter courses that are designed to help teachers to develop the necessary skills to teach science to middle grade students. The program is developed by the Higher Education Committee and the departments that offer this program require the prospective teachers to take the same courses that the Higher Education Committee determined in 1985 (Kavak et al. 2007; Sozer 1992; Turkmen 2017). Therefore, in terms of required courses and total credits to be completed, middle grades science teaching programs look very similar across universities in Turkey. Moreover, although there have been some changes in teacher preparation programs once in every decade since 1985, these changes were minor. As such, there is uniformity in the middle school science teaching programs offered by universities in Turkey.

Another pathway to become a middle grades science teacher in Turkey is through completing a pedagogical formation program that can be completed during or after completion of a bachelor's program in natural sciences or engineering (i.e. biology, molecular biology and genetics, chemistry, chemistry engineering, physics, and physics engineering, TTKB 2014). The main difference between the formation programs and the regular teaching training programs is in the course requirements and the total number of credits to complete to be certified to be a science teacher. For instance, to graduate from a university-based middle school science teaching program, a prospective teacher completes at least 72 credits of pedagogical course work in addition to subject matter content and other course requirements (Higher Education Council 2018). On the other hand, to successfully complete the pedagogical formation program, a prospective teacher completes only 25 credits of course work (Higher Education Council 2014).

In contrast, the number of subject matter courses prospective teachers are required to complete is higher for those who follow the pedagogical formation route. For instance, while prospective teachers in the science teaching program are required to complete between 13 and 20 science, technology, engineering, and mathematics (STEM) courses (approximately 40–75 credits), prospective teachers in the pedagogical formation program should have finished between 26 and 30 STEM courses (approximately 80–95 credits) to earn their bachelor's degree. In addition, the level of science courses in a natural science bachelor's program is much higher than the level of the STEM courses required for the students in the science teaching program. Further complicating matters, a teacher who graduated from a physics program could graduate without taking any biology or earth science courses throughout their undergraduate education. However, middle school science teachers are generalists—they are required to teach topics in biology, chemistry, and physics. While a teacher's extensive knowledge in physics may be an advantage for her or his students to understand physics more, it could be a disadvantage in other science subjects.

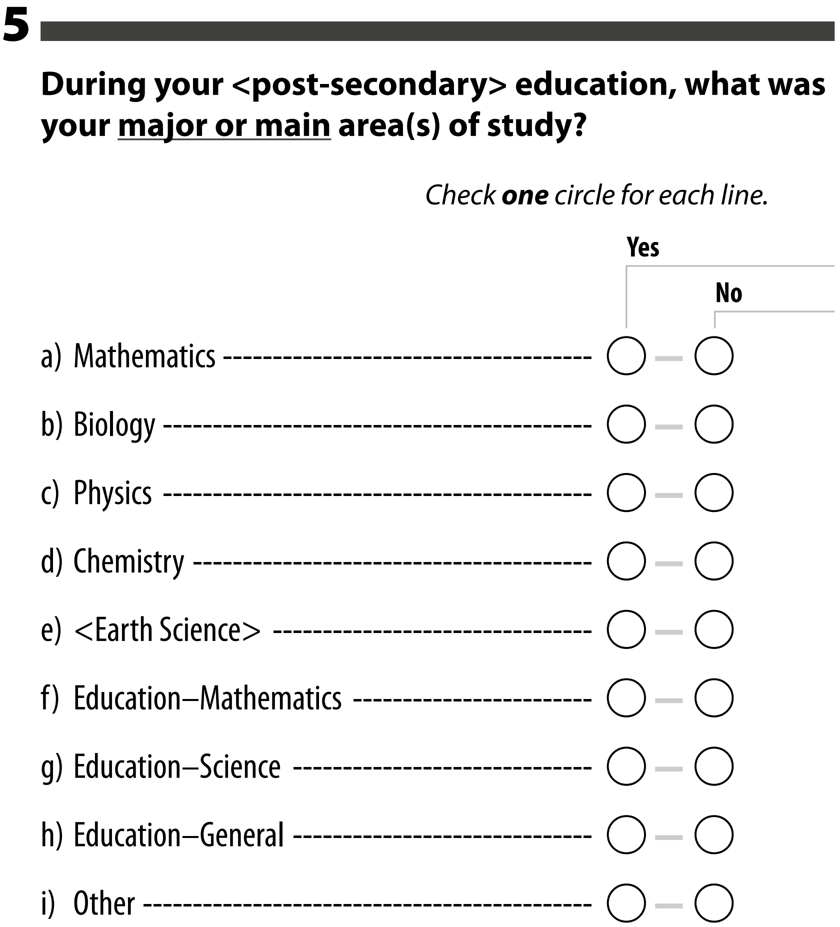
Due to these differences in teacher preparation programs in Turkey, there are discussions regarding the quality of teachers that each program produces. Therefore, the purpose of my study was to investigate the relationship between student achievement and the type of teacher preparation program their teachers graduated from in Turkey.

In addition to the policies related to science teacher preparation programs in Turkey, more contextual information about the policies that limits having multiple majors and minors at the post-secondary level is useful to understand the translation issue discussed in this paper. An undergraduate student in Turkey can graduate from a maximum of two majors at the same time if they meet certain criteria for a double major (Resmi Gazete 2010). However, students from an education program are not allowed to choose a non-education major as their second major. The opposite is also true for students in non-education majors: a student from a non-education major cannot choose a major in education as part of their double major program. However, students in a non-education major can be certified to teach in their field if they complete a pedagogical formation training during or after their undergraduate education. Similar to the double major program rules, most of the universities in Turkey do not allow an undergraduate student to have more than one minor (Bogazici University, n.d.; Hacettepe University 2013; Istanbul University 2014). However, a student can be in a double major and a minor program at the same time. Therefore, a student can have a maximum of two majors and a minor at the same time. Understanding these constraints are central to the potential translation issues found within the TIMSS data.

Teacher preparation items in TIMSS

There are two items on each of the eighth grade science and mathematics teacher background questionnaires in TIMSS 2007, 2011 and 2015 about participating eighth grade teachers' education: (1) "What is the highest level of formal education you have completed?" and (2) "During your < post-secondary > education, what was your major or main area(s) of study?" The second item is designed to determine teachers' major(s) in the field that they received their degrees. Except for the 2007 mathematics teacher background questionnaire,[1] the choices presented to teachers are (1) biology, (2) physics, (3) chemistry, (4) earth science, (5) mathematics, (6) education-mathematics, (7) education-science, (8) education-general, and (9) other. Teachers could select as many of these choices as applicable (IEA 2011). Figure 1 shows the exact question that was asked in TIMSS 2011. The original version of the question is in English and it has not been changed since 2007.

[1] In 2007, the options provided to the mathematics teachers are: (1) mathematics, (2) education-mathematics, (3) science, (4) education-science, (5) education-general, and (6) other.



5

During your <post-secondary> education, what was your major or main area(s) of study?

*Check **one** circle for each line.*

	Yes	No
a) Mathematics	○	○
b) Biology	○	○
c) Physics	○	○
d) Chemistry	○	○
e) <Earth Science>	○	○
f) Education–Mathematics	○	○
g) Education–Science	○	○
h) Education–General	○	○
i) Other	○	○

Fig. 1 The TIMSS 2011 item for teachers' major and minor (IEA 2011)

Detecting the adaptation problem

Throughout the data preparation and cleaning process for my regression analysis, it became apparent that science teachers in Turkey interpreted the question related to their major in different ways: an unignorable number of science teachers selected many of the choices provided to define their major. For instance, 30% of the teachers selected choices b (biology), c (physics), d (chemistry), g (education-science), and h (education-general) altogether as their major or minor. As explained before, in Turkey, individuals can have a maximum of two majors and one minor at the same time when they graduate from college. Therefore, in the most extreme case (considering the possibility that they included their minors in their responses), the maximum number of choices selected at the same time is three if the question was interpreted by the teachers as intended.

To investigate further, I checked responses of mathematics teachers from the same year and science and mathematics teachers' responses to the same item in 2007 and 2015 questionnaires as well. Figure 2 shows the distribution of the number of options selected by the science teachers in TIMSS 2007, 2011, and 2015. The numbers in each rectangle box in the figure show the number of choices that teachers selected together and the percentages of the teachers who selected that many choices.

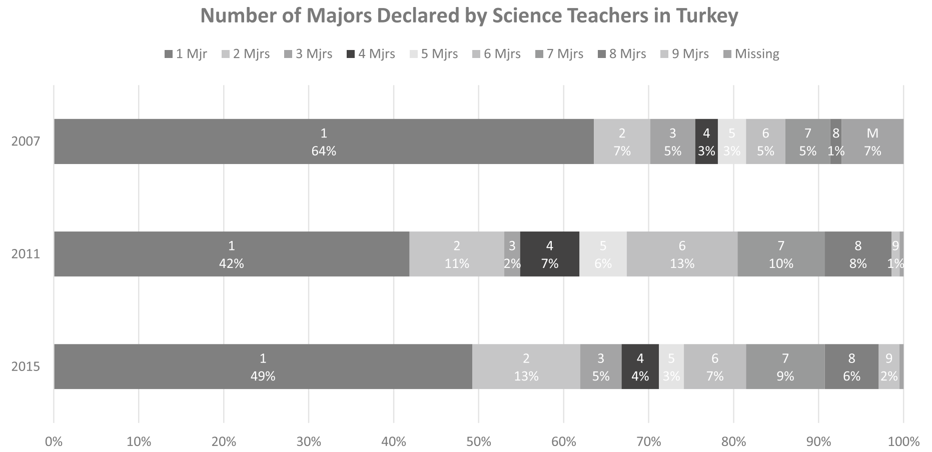


Fig. 2 Number of majors declared by the science teachers in Turkey

Table 1 Original items in English, their adapted versions in Turkish, and back-translations of the adapted items to English

	Original item	Item adapted to Turkish	Item back-translated to English
2007	During your < post-secondary > education, what was your major or main area(s) of study?	Yüksek öğreniminizde hangi alanda eğitim aldınız?	During your post-secondary education, which field did you receive education in?
2011	During your < post-secondary > education, what was your major or main area(s) of study?	Yüksek öğreniminizde hangi anadal ya da alanlarda eğitim aldınız?	During your post-secondary education, which major or fields did you receive education in?

According to Fig. 2, 17%, 45%, and 33% of the science teachers declared more than three major areas of study in 2007, 2011, and 2015, respectively. Moreover, based on the restrictions, individuals in an education-related program cannot apply to a double-major program in a non-education field or vice versa. This also limits the possible combination of choices that teachers can report if they interpret the question as intended. However, in the TIMSS data set, 11% (2007), 51% (2011), and 40% (2015) of the science teachers declared majors or minors in education and non-education fields which is not possible based on the policy context of Turkey.

These anomalies in science teachers' responses created a need for checking the adapted version of the item to Turkish.[2] Table 1 presents the original items in English and Turkish and my back translations of the Turkish version of the items to English for 2007 and 2011 questionnaires. As depicted in Table 1, when back-translated from Turkish to English, the Turkish version of the item in the 2011 questionnaire read, "During your post-secondary education, which major or fields did you receive education in?".[3]

[2] Finding the adapted versions of the TIMSS questionnaires was not easy. They are not available online and there is no contact information provided anywhere that researchers can use to reach out for such information. I searched for articles that is written in Turkish and that used the same data set and reached out to their authors to obtain the adapted versions of the survey.

[3] As a native speaker of Turkish, I translated these items back to English.

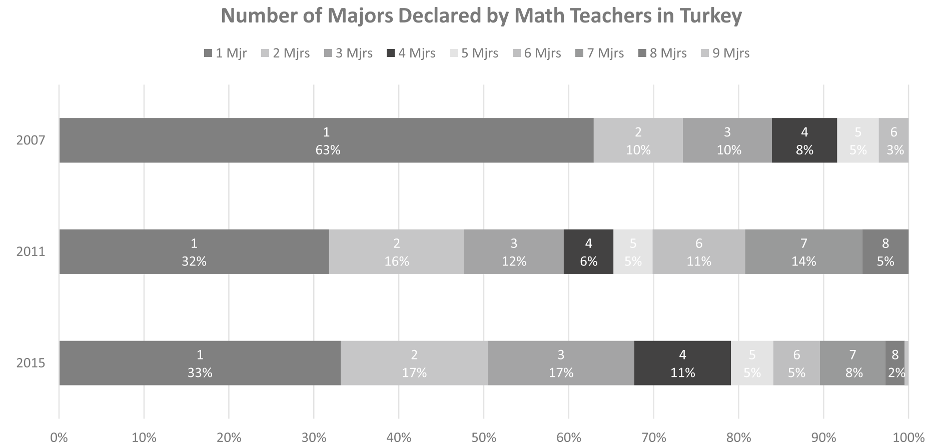


Fig. 3 Number of majors declared by the mathematics teachers in Turkey

Based on the patterns in teachers' answers and the verb selection (i.e. "receive") in the Turkish version, it is very likely that some of the teachers misinterpreted the question as the courses they had taken throughout their post-secondary education.

To investigate further, I analyzed eighth grade mathematics teachers' responses in TIMSS 2007, 2011, and 2015, hypothesizing that they would provide a similar pattern. Figure 3 presents the distribution of the number of major areas of study the mathematics teachers reported in Turkey. Indeed, mathematics teachers' responses have a similar pattern. In this case, 16%, 41%, and 32% of the mathematics teachers selected more than three major areas of study in 2007, 2011, and 2015, respectively. Moreover, 32% (2007), 63% (2011), and 60% (2015) of the mathematics teachers' selections include both education and non-education fields which is not possible as explained earlier.

Another interesting pattern that emerged in both science and math teachers' responses is the difference in the percentages of teachers that declared more than three major areas of study in 2007 compared to 2011 and 2015 data. Between 2007 and 2011, the percentage of teachers who selected more than three major areas of study increased from 17% to 45% for science teachers and from 16% to 41% for mathematics teachers, a great difference in teachers' responses collected four years apart.

Since the pattern in teachers' responses in 2007 seemed different than the patterns in 2011 and 2015, I also checked the Turkish version of the adapted item included in the 2007 teacher questionnaires and found that the item was adapted slightly differently to Turkish in 2007 than in 2011. As depicted in Table 1, when the item from 2007 was translated back from Turkish to English, this time it read "in your post-secondary education, which field did you receive education in?" Although the question was still asking the field that the teachers receive their education in, which is likely problematic, using the singular form (i.e. "field") instead of the plural (i.e. "fields") may have increased the possibility that the question was interpreted by the teachers as the major that they graduated from instead of the courses they took.

In addition to the problems related to the translation of the question root, the options provided to science and mathematics teachers may aggravate the situation. Each teaching program in Turkey has a unique name that does not change across universities.

Middle grades science teaching programs offered has the same name (i.e., Fen Bilgisi Ogretmenligi) across all universities. Likewise, middle grades mathematics teaching programs have a unique name (i.e., Ilkogretim Matematik Ogretmenligi). When science and mathematics teachers could not see the exact names for the program that they graduated from among the options, their uncertainty might have increased. Moreover, providing mathematics teachers the same options that science teachers have for their major probably confused participating mathematics teachers even more than the science teachers (see Fig. 1) because people who graduated from any science fields are not allowed to teach mathematics in middle schools in Turkey. On the other hand, they might take science courses as an elective or even as a requirement during their post-secondary education. When mathematics teachers see biology, chemistry, or physics as their options they might think that they are asked about the courses they took in college. This may also lead to a misinterpretation of the intended meaning of the question asked.

Discussion

At the beginning of their meeting report on comparability of measures designed by international surveys across cultures, Avvisati et al. (2019) stated, "[W]hen surveys go beyond measuring objective attributes (e.g. age or household size) and behaviors (e.g. unemployment or job-seeking behaviors) and aim to assess subjective attitudes, […] or psychological traits […] new challenges for the validity and comparability of survey results emerge (p. 1). Presenting a case from students' out-of-school activities questionnaire items in PISA, Bray et al (2020) show that it may even be true for measuring objective attributes. Presenting a case from TIMSS, this current paper contributes to Bray et al. (2020) discussion pertaining to the comparability of translated questionnaire items intended to collect information about objective attributes. Linguistic differences between the original and translated items as well as an in-depth analysis of participants' responses that do not reflect the policy context in the country provide supporting evidence that even for "objective measures" there may arise problems with collecting information that can be comparable across cultures or countries.

In comparative studies, assessing the equivalency of items adapted to different cultures and languages can be frequently ignored in practice (Avvisati et al. 2019). Yet many studies and reports use teachers' major information collected by TIMSS as integral to their comparative analyses (e.g. Abazaoglu and Tasar 2016; Akiba et al. 2007; Akyuz 2014; Blomeke et al. 2016; Burroughs and Chudgar 2017; Gustafsson and Nilsen 2016; Mullis et al. 2016). Excepting Akiba et al. (2007), these comparative studies include TIMSS data from Turkey (see for instance, Abazaoglu and Tasar 2016; Blomeke et al. 2016; Gustafsson and Nilsen 2016). There are also unpublished master's or doctorate theses that use teachers' major area of study data from Turkey collected in TIMSS. TIMSS data are also used for policy analysis purposes. It means that the results of the studies and the inferences or decisions made based on the results may be questionable since the information collected does not accurately reflect the educational profiles of teachers in Turkey (Bray and Kobakhidze 2014). For instance, policies focused on the role of course taking by potential teachers, or the number of majors of pre-service teachers and their correlation with student outcomes would be inaccurate based on demographic question distributions that violate national policies and general practices.

In addition to cross-country analyses, the TIMSS data collected from Turkey are also used for analyses within Turkey (Ceylan 2014). Using data collected through ill-adapted items could be more detrimental to these studies since there are very few representative data sets that are available to educational researchers interested in Turkey. Making comparisons of the results across different data sets to check for consistency is not always possible especially for analyses that requires teacher data to be matched to student data. Conclusions reached about teacher preparation in Turkey using this data may poorly inform policy and research on teacher preparation. For instance, implementation of the formation program could be encouraged and supported by the Ministry of Education claiming that there are no differences between teachers with a formation and teachers with a degree in education in terms of their students' achievement even it may not be the actual case. In the following paragraphs, I share my recommendations that may help improve the quality of research studies that use data from international assessment studies.

The current paper supports Bray et al. (2020) conclusion that researchers should not take for granted the item equivalency of background questions in international assessment studies. For researchers, policy analysts, or other users of secondary data who intend to use international large-scale assessment data in their analyses, it is imperative to be aware of the adaptation issues in assessment items, as well as in background items (Avvisati et al. 2019; Bray et al. 2020).

Detecting the poorly-adapted items in advance would save time and resources and reduce the possibility of disseminating inaccurate information. Getting to know the item adaptation process and reaching out more documentation about item adaptations may help the secondary data users detect possible adaptation issues. Also, reading countries' responses to the curriculum questionnaires may give a general idea about the policy context of the country. As illustrated in this paper, examining the responses of the participants in detail to see whether they make sense or not in the cultural and policy context of the countries prior to data analysis may help identify possible issues with item adaptations.

For organizations that conduct international large-scale assessment studies, increasing the transparency in how the items are adapted could help the secondary data users to detect any potential translation problems early on. Currently, the adapted versions of the TIMSS contextual questionnaires are not publicly available. Including the background surveys administered in the documentation package that can be downloaded with country data would help researchers detect possible misinterpretations of the items that they are interested in early on. In Turkey, the adapted questionnaires are not available online and the e-mail address provided on the website which is the only point of contact to request anything related to TIMSS is not responsive.

Furthermore, the item adaptation guidelines shared with the NRCs should be publicly available. Making these guidelines available to researchers and policy analysts could be beneficial for predicting possible sources of adaptation problems. In addition to the guidelines, making the National Adaptation Forms that NRCs are required to submit publicly available would also help to detect the sources of poor adaptations. Sharing these guidelines, documentation, and policies and procedures with the public may provide the opportunity for the researchers to make their own assessments of the item adaptation quality. Moreover, this would not only bring more transparency to the item adaptation process but also make the NRCs more accountable to the public.

As shown above, at times, changes in adapted items occur over time while the original items stay the same. This is especially problematic for longitudinal analyses where it is assumed that changes in participants' responses can be attributed to changes throughout time. However, as in the Turkish case, it is possible that the observed changes in responses could be due to slight changes in translation of a question. Bray et al. (2020) warn researchers about such changes when conducting longitudinal analyses. Checking the adapted version of the item for each year included would give an idea about not only possible misinterpretations but also changes in responses throughout time that can alternatively be explained by changes in translation. Although the adapted versions of the questionnaires may not be available online, contacting the institution (for TIMSS, NRCs) that is nationally responsible for administering the surveys may help the researchers obtain the information they need. Finally, the research field could also help support increasing validity of the surveys by sharing any discrepancies in responses that may be due to translation issues with the test developers as well as with other researchers.

This study focused on only one country —Turkey—and one demographic question as a case for the need to improve the comparability of information collected by large-scale assessment studies and to communicate the issues related to item adaptations with secondary data users so as to better advance the field of educational research and policy decisions. Admittedly, the focus on a single country and question cannot in itself be the foundation for major changes to test creation protocol. On the other hand, by focusing on one case in depth would help secondary data analysts be cognizant of and detect possible adaptation issues.

International assessment studies are invaluable opportunities for countries to obtain information about where their students' knowledge and skills are at compared to students in different countries. As mentioned, for some countries, including Turkey, data collected from international assessments could be one of the few representative data sets that the researchers can have access to. Therefore, the data collected from international assessment studies could contribute to the current knowledge in education. Moreover, tremendous amounts of resources and effort are put in to conduct these studies, therefore, improving the accuracy of the data to be collected by the international studies would help these resources be used as effectively as possible. Improving the quality of data collected by these studies are of great importance since it not only informs educational policies in participating countries but also educational research in general.

Acknowledgements

Thank you to Courtney Preston, Matthew Kloser, and Carolyn Herrington for their feedback on the manuscript.

Authors' contributions

As the sole author, I conducted the analysis and wrote the whole manuscript. The author read and approved the final manuscript.

Funding

The author did not receive any funding to conduct this research.

Availability of data and materials

The TIMSS data that support the findings can be found in the repository of IEA (url: https://www.iea.nl/data-tools/repository).

Declarations

Competing interests

The author does not have any financial or non-financial competing interest.

Received: 8 September 2020 Accepted: 25 February 2021
Published online: 12 March 2021

References

Ababneh, E., Al-Tweissi, A., & Abulibdeh, K. (2016). TIMSS and PISA impact – The case of Jordan. *Research Papers in Education, 31*(5), 542–555.

Abazaoglu, I., & Tasar, M. F. (2016). Relations of characteristics of science teachers and students with the student achievement in science: A case analysis according to TIMSS 2011 data. *Elementary Education Online, 15*(3), 922–945.

Akiba, M., LeTendre, G. K., & Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational Researcher, 36*(7), 369–387.

Akyuz, G. (2014). The effects of student and school factors on mathematics achievement in TIMSS 2011. *Education and Science, 39*(172), 150–162.

American Psychological Association (n.d.). Validity. In APA dictionary of psychology. Retrieved January 30, 2021, from https://dictionary.apa.org/validity

Anderson, J. O., Lin, H., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education, 5*(4), 591–614.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research, 54*(1), 37–59.

Arffman, I. (2012). International education studies: Increasing their linguistic comparability by developing judgmental reviews. *International Scholarly Research Network, 2012,* 1–11.

Atar, H. Y. (2014). Multilevel effects of teacher characteristics on TIMSS 2011 science achievement. *Education and Science, 39*(172), 121–137.

Avvisati, F., Le Donne, N., & Paccagnella, M. (2019). A meeting report: Cross-cultural comparability of questionnaire measures in large-scale international surveys. *Measurement Instruments for the Social Sciences, 1*(8), 1–10.

Azar, A. (2011). Quality or quantity: A statement for teacher training in Turkey.". *Journal of Higher Education and Science, 1*(1), 36–38.

Blomeke, S., Olsen, R. V., & Suhl, U. (2016). Relation of student achievement to the quality of their teachers and instructional quality. In T. Nilsen & J. Gustafsson (Eds.), *Teacher quality, instructional quality, and student outcomes*, (pp. 21–50). Springer.

Blum, A., Goldstein, H., & Guérin-Pace, F. (2001). International Adult Literacy Survey (IALS): An analysis of international comparisons of adult literacy. *Assessment in Education: Principles, Policy & Practice, 8*(2), 225–246.

Bogazici University. (n.d.). Bogazici Universitesi yan dal programlari yonergesi. Bogazici University. http://boun.edu.tr/tr-TR/Content/Ogrenciler/Ogrenci_Isleri/Yonetmelik_ve_Ic_Tuzukler/Yandal_Yonergesi

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice, 9*(3), 387–399.

Bray, M., & Kobakhidze, M. N. (2014). Measurement issues in research on shadow education: Challenges and pitfalls encountered in TIMSS and PISA. *Comparative Education Review, 58*(4), 590–620.

Bray, M., Kobakhidze, M. N., & Suter, L. E. (2020). The Challenges of Measuring Outside-School-Time Educational Activities: Experiences and Lessons from the Programme for International Student Assessment (PISA). *Comparative Education Review, 64*(1), 87–106.

Burroughs, N. & Chudgar, A. (2017). *The role of teacher quality in fourth-grade mathematics instruction: Evidence from TIMSS 2015* [Policy brief]. International Association for the Evaluation of Educational Assessment.

Ceylan, E. (2014). Examining item difficulties with respect to science teachers' backgrounds and their views on science instruction. *Education and Science, 39*(172), 138–149.

Clauser, B. E., & Kathleen, M. (1998). An NCME instructional module on using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–43.

Domínguez, M., Vieira, M. J., & Vidal, J. (2012). The impact of the Programme for International Student Assessment on academic journals. *Assessment in Education: Principles, Policy & Practice, 19*(4), 393–409.

Ebbs, D., & Korsnakova, P. (2015). Translation and translation verification. In M. O. Martin, I. V.S. Mullis, & M. Hooper (Eds.). *Methods and Procedures in TIMSS 2015.* TIMSS & PIRLS International Study Center.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(1998), 543–553.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3–4), 199–215.

Ferrini-Mundy, J., & Schmidt, W. H. (2005). International comparative studies in mathematics education: Opportunities for collaboration and challenges for researchers. *Journal for Research in Mathematics Education, 36*(3), 164–175.

Fertig, M. (2003). Who's to blame? The determinants of German students' achievement in PISA 2000 study. IZA Discussion Papers No. 739. Institute for the Study of Labor.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*(4), 304–312.

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24*(1), 23–37.

Gurol, M., Turkan, A., & Som, İ. (2018). Pedagojik formasyon sertifika programinin degerlendirilmesi. *Elektronik Sosyal Bilimler Dergisi, 17*(65), 103–122.

Gustafsson, J. E., & Nilsen, T. (2016). The Impact of school climate and teacher quality on mathematics achievement: A difference-in-differences Approach. In T. Nilsen & J. Gustafsson (Eds.). *Teacher quality, instructional quality and student outcomes,* Springer.

Hacettepe University. (2013). Hacettepe Universtesi yandal programi yonergesi. Hacettepe University. https://www.hacettepe.edu.tr/duyuru/yonergeler/1723,98.pdf

Higher Education Council. (2014). *Pedagojik formasyon egitimi sertifika programina iliskin usul ve esaslar.* Higher Education Council. http://www.yok.gov.tr/web/guest/icerik/-/journal_content/56_INSTANCE_rEHF8BIsfYRx/10279/7052802

Higher Education Council. (2018). *Fen bilgisi ogretmenligi lisans programi.* Higher Education Council. http://www.yok.gov.tr/documents/10279/41805112/Fen_Bilgisi_Ogretmenligi_Lisans_Programi.pdf

IEA. (2011). TIMSS 2011 *Teacher Questionnaire Science <Grade 8>*. TIMSS & PIRLS International Study Center.

International Test Commission (2017). ITC Guidelines for translating and adapting tests (second edition). *International Journal of Testing,* DOI: https://doi.org/10.1080/15305058.2017.1398166Istanbul University. 2014. Istanbul Universitesi yandal programi yonergesi. Istanbul University. http://cdn.istanbul.edu.tr/FileHandler2.ashx?f=yandal-yo%CC%88nergesi-10.04.2014.pdf

Kavak, Y., Aydin, A., & Akbaba-Altun, S. (2007). *Ogretmen yetistirme ve egitim fakulteleri (1982–2007).* Higher Education Council.

Kirsch, I., & Braun, H. (2020). Changing times, changing needs: Enhancing the utility of international large-scale assessments. *Large-scale Assessment in Education, 8,* 10.

Klemencic, E. (2010). The impact of international achievement studies on national education policymaking: The case of Slovenia–How many watches do we need?. In A. W. Wiseman. (Ed.). *The Impact of international achievement studies on national education policymaking* (Vol. 13, pp. 239–266). Emerald Group Publishing Limited.

Lockheed, M. E., & Wagemaker, H. (2013). International large-scale assessments: Thermometers, whips or useful policy tools? *Research in Comparative and International Education, 8*(3), 296–306.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127–143.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics.* IEA.

Naumann, J. (2005). TIMSS, PISA, PIRLS and low educational achievement in World society. *Prospects, 35*(2), 229–248.

OECD. 2009. *PISA data analysis manual: SPSS* (2nd ed.). OECD.

Paine, L., & Zeichner, K. (2012). The local and the global in reforming teaching and teacher education. *Comparative Education Review, 56*(4), 569–583.

Resmi Gazete. (2010, April 24). Yuksekogretim kurumlarinda onlisans ve lisans duzeyindeki programlar arasinda gecis, cift anadal, yandal ile kurumlar arasi kredi transferi yapilmasi esaslarina iliskin yonetmelik. *Resmi Gazete*.

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies, 42*(3), 411–430.

Rutkowski, L., & Rutkowski, D. (2019). Methodological challenges to measuring heterogenous populations internationally. In L.E. Suter, E. Smith, & B. Denman (Eds.). *SAGE Handbook of Comparative Studies in Education.* SAGE.

Sjoberg, S. (2015). PISA and global educational governance – A critique of the project, its uses and implications. *Euroasia Journal of Mathematics, Science & Technology Education, 11*(1), 111–127.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education, 4*(43).

Solano-Flores, G., Contreras-Nino, L. A., & Backhoff-Escudero, E. (2006). Translation and adaptation of tests: Lessons learned and recommendations for countries participating in TIMSS, PISA and other international comparisons. *Revista Electronica de Investigacion Educative, 8,* 2.

Solano-Flores, G., Backhoff, E., & Contreras-Nino, L. A. (2009). Theory of test translation error. *International Journal of Testing, 9*(2), 78–91.

Sozer, E. (1992). Universitelerde 1982 oncesi ve sonrasında ogretmen egitimi ile ilgili program uygulamalari. *Kurgu Dergisi, 10,* 259–278.

Steffen, B., & Hößle, C. (2014). Decision-making competence in biology education: implementation into German curricula in relation to international approaches. *Euroasia Journal of Mathematics, Science & Technology Education., 10*(4), 343–355.

Talim ve Terbiye Kurulu Baskanligi. (2014). Ogretmenlik alanlari, atama ve ders okutma esaslari. *Milli Egitim Bakanligi Tebligler Dergisi, 77*(2678), 256–299.

Tobin, M., Lietz, P. Nugroho, D. Vivekanandan, R., & Nyamkhuu, T. (2015). Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific Region. Australian Council for Educational Research.

Turkmen, H. (2017). Science Teacher Preparation in Turkey. In J. E. Pedersen, T. Isozaki, & T. H. Charlotte (Eds.). *Model science teacher preparation programs: An international comparison of what works*. Information Age Publishing.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(1), 29–37.

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Europeenne de Psychologie Appliquee, 54,* 119–135.

van de Vijver, F. J. R., & He, J. (2016). Bias assessment and prevention in noncognitive outcome measures in context assessments. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning, methodology of educational measurement and assessment*. New York: Springer.

Wolf, R. M. (2002). Extra-school Instruction in Mathematics and Science. In D. F. Robitaille & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data*. New York: Kluwer Academic Publishers.

Yildirim, A. (2013). Teacher education research in Turkey: Trends, issues and priority areas. *Education and Science, 38*(169), 175–191.

Publisher's Note