Schwippert, Knut; Wendt, Heike

It's all about validity: preparing TIMSS and PIRLS background questionnaires for the 21st century

Tertium comparationis 23 (2017) 1, S. 28-46

Tertium Comparationis
Journal für International und Interkulturell
Vergleichende Erziehungswissenschaft
Vol. 23, No.1, pp. 28–46, 2017

It's all about validity: Preparing TIMSS and PIRLS background questionnaires for the 21st century

Knut Schwippert
University of Hamburg

Heike Wendt
TU Dortmund University

Abstract

International large-scale assessments are facing several challenges: it is expected that they deliver reliable and comparable information on different levels of educational systems as well as documenting the situation of educational systems every few years as trend information over time. This means that the instruments used to gather the information need to be adapted to actual situations in each cycle of the study as well as focusing on information that is sufficiently important for trend observations over the different cycles. In order to gather actual information as well as assessing reliable and valid information over time, this involves walking on a tightrope. The challenge here is to predict what is currently important and what will also be relevant in the future.

1. Introduction

Learning processes are not easy to understand – input, process, and output factors are multiple correlated –; this is what can be observed by analyzing available data from large-scale assessments. However, correlations are no proof of causality. We need to understand how these factors interact, which of them are independent, which are dependent, and which of both can be changed or influenced by developments or reforms in the educational sector. To access this information within large-scale assessments background questionnaires are addressed to relevant groups of persons. Many of this background information can be reduced to their empirical values: thereby it seems to be easy to compare their values between participating

countries. To understand what these variables explain in a specific country it is necessary to understand their relation to social, cultural, economic and historical background factors. This kind of interpretation makes the difference between just numerous values and content related interpretations of differences in the observations.

The aim of this article is to open the view for information that is gathered in the context of large-scale assessments. It is an attempt to sensitize for challenges and to initialize discussions for further development. Only with agreed objectives of future large-scale assessments specific recommendations could be given. The article will promote the understanding of the logic behind large-scale assessments, their opportunities but also their limits.

This article will describe the purpose of international large-scale assessments (LSA) and will shed light on the theoretical and practical background of the process of developing tests and background questionnaires in its context from a country-specific perspective, namely Germany. To understand differences between countries or changes within one country over time means to take the specific circumstances during the assessments into account. Comparing values is a mathematical operation, but to understand the content of the values means to understand them in a valid manner. Without reflecting the interpretation of the results its validity is questionable. Last but not least structural issues of LSAs will be addressed and conceptual strategies for the adaptation, development and preparation of international large-scale assessments will be provided. It should be noted that no specific recommendations on item or scale level will be provided on a general level for large-scale assessments at the end of this article since – as it will be shown – this would be misleading in the understanding of LSA as cooperative projects. Nevertheless from a country specific perspective we point on issues that need to be addresses by looking on actual challenges in the German educational system.

2. Large-scale assessments: Figures vs. content

Large-scale assessments (LSA) are well known not only by researchers or specialists but also by an interested public informed by general media. In many countries the basic results of large-scale assessments find their way into the public. Interested persons find more and detailed information on well prepared websites. The scientific community encompasses the whole world.

Most of the ‘consumers’ of the results provided by public press or internet understand the results just as league tables. Unfortunately, these league tables lead to a popular perception of LSAs representing only a (negligible) part of the surveys. Only those who dig deeper into the aim of the studies for which they are designed

and those who understand the limits of the surveys will receive a more helpful picture of the (different levels of the) educational systems that lie in the focus.

Results need to be reflected – what do the results of a single country or the comparison of many countries tell us. The benefits of large-scale assessments only evolve by informed and reflected interpretation of the results. The numerous results are interesting by themselves but for an understanding content-related (valid) interpretations are necessary.

In the following, brief descriptions of the main aims of large-scale assessments are given. Following two well established trend surveys – TIMSS (*Trends in International Mathematics and Science Study*) and PIRLS (*Progress in International Reading Literacy Study*) – and their frameworks for the development of achievement tests and background questionnaires are introduced. As a main focus the interpretation – in the sense of valid information – of background variables is discussed. Only in the context of an argument-based approach to validation it can be understood that LSAs are not just league tables but offer valuable information to understand, to interpret and to develop educational systems, by taking the appropriate (valid) interpretation of the results into account.

2.1 Large-scale assessments as system monitoring studies

The first challenge – delivering reliable information on different levels of educational systems – is addressed by the initiator and the executing organization of such a study. Depending on its research interest, the focus of the study may vary. Since large-scale assessments require resources such as money, knowledge and the time of the participants, such studies have to focus on gathering necessary information to make informed decisions at a political, administration, district, school, classroom or private level. Since many large-scale assessments are mainly designed to offer information about educational systems, these surveys serve as system monitoring studies. For instance, within these studies, information is gathered about students' family backgrounds as covariates. One international leading organization that initializes large-scale assessments is the *International Association for the Evaluation of Educational Achievement* (IEA). The IEA is an organization founded by international-honored researchers who explored the possibility of comparing educational systems of different countries by understanding this as a natural experiment (Husén, 1967). This view was founded upon the observation that – empirically speaking – independent variables that can explain changes or developments do not change within a country-specific context: in other words, they do not have variance within one system. For example, this can include information about the structure of the educational system (tracked or not tracked) as well as concerning the question whether national assessments exist. International large-scale assessments offer the

possibility to evaluate these variables by comparing them between different educational systems. However, by creating an international comparison, the core question is to explore the important variables in an educational system that can prompt enhancements for students once their values differ. Nonetheless, the question remains: What is the global goal of educational systems to enhance students in different subjects, motivation or social values? What are international common goals or what goals are important only for specific regions of the world? Finally, what variables need to be varied to achieve expected and desired changes?

The main field of expertise of the IEA is to assess the cognitive achievement of students, including reading, mathematics and science literacy, which is taught across the world. To assess these student abilities, first a framework needs to be designed that explains everyday life experiences that students need to learn or master to participate in social life in an informed, critical, and autonomous manner, with the perspective of being a gainful and responsible member of society (Tenorth, 1994). For younger students, this means that they need to learn the fundaments of cultural techniques (like reading, calculating and basic natural scientific knowledge) step by step in a systematical manner.

Next to the cognitive abilities of students, it is also important to assess the circumstances of learning because they open the possibility for changes that help or enable the students in their learning process. Since students' learning is affected by many circumstances, the infrastructure of their learning environments in both their families and their schools and classes – and finally in their peers' out of school and families – needs to be pictured in a survey that aims towards improving students' achievement and their future prospects in society.

The second challenge – to document the situation of educational systems every few years as trend information – encompasses the first challenge to further identify meaningful variables for assessing both achievement and background variables. Large-scale assessments designed as cross-sectional studies offer a specific quality of information on the educational systems that participate in these surveys. Nevertheless, cross-sectional studies offer a snapshot of the system, without the possibility to understand or extrapolate changes in both the average achievement level of students as well as causal inferences on those circumstances that may cause changes in student achievement. In order to investigate changes, longitudinal study designs are necessary. Since international large-scale assessments are mainly designed as system monitoring studies, the main focus lies on the observation of system-relevant information, whereby the students' ability is one out of many possible indicators for the success of the educational system.[1] Beginning in the mid-1990s, the frequency of international large-scale assessments became so high that the cycles of studies could be used as longitudinal information at the system level. Pre-

viously the 'steps' between studies focusing on the same or similar objectives were so wide that each of the studies was state of the art within its time but incomparable over time. To offer continuous information about educational systems the interval between the surveys became shorter and the assessment of achievement and background information was carried over to the next cycles, marking the birth of (system monitoring) trend studies. However, by designing these trend studies, a new challenge emerged: How can changes be measured in changing systems? Does the construct in one cycle of the study still represent the same in a later cycle? From the perspective of test theory – the *conditio sine qua non* – the fundamental axiom is "if you like to measure change, do not change the measure" (Tukey & Beaton cited in Mullis, Drucker, Preuschoff, Arora & Stanco, 2012, p. 3). However, if the reality changes how can we access the 'new' state of the art reality if we are constrained not to change the tools to assess it?

Transferred to large-scale assessments this means: measuring trend is a challenge! The IEA set up two worldwide studies which document the development of educational systems and manage the challenges of large-scale assessments as trend studies. These studies – namely TIMSS and PIRLS – are presented in the following section.

2.2 IEA trend studies: TIMSS and PIRLS

Gaining the possibility to live one's own life self-determined in today's world requires mastering the basic culture techniques that enable a person to communicate, obtain information independently, take part in daily life and have the opportunity to accomplish a productive job, namely being able to read, calculate and have some basic natural scientific knowledge. Students around the world start to learn this basic prerequisite from primary grade onwards. Owing to the importance of students' enhancement in schools, in 1960 the IEA decided to conduct the first international comparative pilot study focusing on mathematics, science, reading comprehension, geography and non-verbal ability (Kyriakides & Charalambous, 2014). Four years later (1964), the *First International Mathematics Study* (FIMS) was conducted and in 1970–71 the data collection of the *First International Science Study* (FISS) took place (IEA, n.d.). The *Second International Mathematics Study* (SIMS) followed in 1980–82 and the *Second International Science Study* (SISS) in 1983–84. Finally, in 1995 a milestone for assessing mathematics and natural sciences occurred as the *Third International Mathematics and Science Study* (TIMSS) was carried out. This TIMS study was the beginning of a four-year cycle by assessing the mathematics and natural science ability of students in primary (3–4 graders) and secondary (7–8 graders) school. After 1999 – the first repetition

of TIMSS – the name of the cycle was adapted to its real design: *Trends in International Mathematics and Science Study* (followed by the year of the assessment).

Although the assessment of reading started with the first pilot study (1960), it took until 1990–91 until the *International Reading Literacy Study* (IRLS) was conducted to investigate the reading ability of 9 and 14 year old students (IEA, n.d.). About ten years later, the second milestone in the context of international large-scale assessments was marked, with the beginning of the five-year cycle of the *Progress in International Reading Literacy Study* (PIRLS) as a trend study of fourth graders' reading ability.

Germany participated in the IEA *Six-Subject Study* in 1970, with (from its national perspective) disappointing results. Due to the dominant research paradigm in qualitative-oriented humanitarian sciences, Germany did not participate in the ongoing IEA international large-scale assessments for the next 20 years. Credit belongs to Rainer Lehmann, who ensured that Germany participated in 1991 with representative samples from East and West Germany (right after the fall of the Berlin Wall) and four years later in TIMSS 1995. Due to the disappointing results again, German politicians decided to participate mandatorily in international system monitoring studies addressing the core culture techniques of society (reading, mathematics and science ability). Hence, Germany has participated in PIRLS from the first cycle in 2001 until today (2016) and with the primary grade cohort in TIMSS since 2007.

Since the results of international comparative studies such as TIMSS and PIRLS are read very carefully by politicians, administration and research in Germany, the value of the studies strongly depends on their information content. Next to the appropriate achievement tests, the richness of the information content of the studies is based on background questionnaires. This offers the necessary independent variables that can be used to find and explain differences in the students' abilities. Once patterns of explanations for the differences are identified – based on either the national or international level – this information can be used to initialize and undertake reforms in the educational sector. Some may initialize effects very soon while others may take longer depending on the level in the educational system where these reforms take effect.

Although the political decision to participate in international system monitoring studies has been made, the value of the studies depended on how accurately actual changes in society and its consequences for the educational system are addressed in these studies. For this reason, from a German perspective it is important that a powerful research framework is available that covers the most actual challenges in the educational system, as well as ensuring the participation of a significant number of countries. Most interesting are those countries that do not have the same but

similar challenges in their system as well as acting in similar circumstances as Germany. By looking at the top-scoring countries, international large-scale assessments like TIMSS and PIRLS offer Germany important information about what is possible in educational systems. However, the core issue is to understand how neighbors or partners in unions handle the same or similar challenges given that they have comparable political, social and historical background under which the educational systems developed or are administered.

Within the IEA the development of LSA is based on rich experience in conducting international assessments around the world. Based on the documented previous knowledge a successful design of future studies is only promising if current theoretical models are taken into account as well. To give an idea of the orientation in the process of the development of test instruments and background questionnaires the following two sections give a brief insight.

2.2.1 Conceptual framework for achievement tests

The concepts of the IEA studies are inspired by the perspective to contribute to an educational system in which students are prepared for their lives ahead. This mission is clearly apparent once the achievement definitions of the actual TIMSS 2015 and PIRLS 2016 cycles are read:

> TIMSS 2015 Mathematics Framework: Mathematics is essential in daily life for such activities as counting, cooking, managing money, and building things. Beyond that, many career fields require a strong mathematical foundation, such as engineering, architecture, accounting, banking, business, medicine, ecology, and aerospace. Mathematics is vital to economics and finance, as well as to the computing technology and software development underlying our technologically advanced and information-based world (Grønmo, Lindquist, Arora & Mullis, 2013, p. 11).
>
> TIMSS 2015 Science Framework: The development of an understanding of science is important for students in today's world if they are to become citizens who can make informed decisions about themselves and the world in which they live. Every day they will be faced with a barrage of information, and sifting fact from fiction and understanding the scientific basis of important social, economic, and environmental issues is possible only if they have the tools to accomplish this (Jones, Wheeler & Centurino, 2013, p. 29).
>
> PIRLS 2016 Reading Framework: The PIRLS definition of reading literacy is grounded in IEA's 1991 study, in which reading literacy was defined as 'the ability to understand and use those written language forms required by society and/or valued by the individual' (Mullis, Martin & Sainsbury, 2015, p. 11).

Based on these definitions, the tests for assessing students' achievement are designed whereby they reflect tasks that are age-appropriate displayed and address the required cognitive abilities to solve these tasks. By developing tasks and the corresponding test items, the experience that the students usually have at that age is

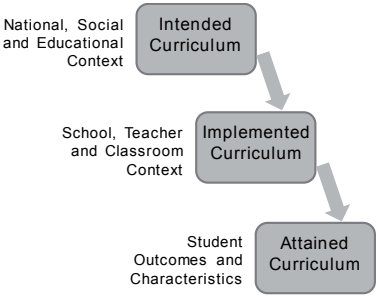
taken into account. Although the test items themselves are artificial – because they are displayed on paper or screen – they nevertheless address the interesting subject-related (cognitive) abilities of the students.

For the IEA studies, expert teams worked on the assessment frameworks and filled them with relevant subject areas and content domains of the interesting abilities. For example, the reading ability tests presented to the students comprise different kinds of texts (informational and narrative), for which items are created that address different areas of comprehension. In a similar way, mathematics (number, geometric shapes and measures, as well as data display) and science (life science, physical science and earth science) tests are developed addressing the cognitive domains of knowing, applying and reasoning (Hooper, Mullis & Martin, 2013).

To ensure that the designed assessment framework also covers the relevant areas in the national educational systems, a broad agreement exists concerning which abilities are taught in school around the world and how this can be assessed. For this reason, in IEA studies the assessment framework follows the common teaching goals described by the participating countries in their national curricula.

In IEA studies, the curriculum is understood "as the major organizing concept in considering how educational opportunities are provided to students and the factors that influence how students use these opportunities" (Mullis, 2013, p. 4). Within IEA's curriculum model (see figure 1), three (hierarchical) levels are distinguished. The top level of this model represents the legal level of national curriculum. In the official written documents, the intention of the teaching and learning process is described. Since this is the normative- and theoretical-driven level, it is described as the intended curriculum. Thinking further from this theoretical framework towards practical issues at the next level, the implemented curriculum is located. Within this (theoretically deducible) curriculum, the realized aspects of the intended curriculum are summarized. The implemented curriculum reflects the notion that – due to practical reasons or circumstances – not all intentions of the intended curriculum could be transferred into the practical teaching and learning process. For example, the textbooks used do not cover exactly the intended curriculum, while teachers cannot – e.g. due to social interactions with the students or different levels of students' abilities, motivation, or interest – realize all intended aspects of the documented (intended) curriculum. Finally, the realized (attained) curriculum reflects the notion that not everything that has been taught has been learned by the students, for whatever reason.

Figure 1: TIMSS curriculum model (source: Mullis, 2013, p. 5)



Since the national (intended) curricula contain the official goals for teaching and learning but not a description of the circumstances or background information of the educational system, the IEA decided to gather further information concerning the national curriculum with a curriculum questionnaire which is answered by the national research coordinator of the actual study. The international study center highlights:

> The questionnaire is designed to collect basic information about the organization of the mathematics and science curriculum in each country, and about the content of these subjects intended to be covered up to the fourth and eighth grades. It also includes questions on attrition and retention policies, the local or national examination system, as well as goals and standards for mathematics and science instruction (Martin, Mullis & Foy, 2013, p. 97 f.).

To understand the process of teaching and learning by considering the opportunities to learn, all three levels of the curriculum – intended, implemented and realized – must be taken into account. Only an understanding of the different layers of intentions and implementations in the teaching and learning process – by taking the level-specific circumstances gathered by background questionnaires into account – allows a purposeful development of educational systems.

2.2.2 Conceptual framework for background questionnaires

The stock-taking of the ability level of students by itself does not offer any indications about what needs to be changed in the educational system to enhance students' achievement. For this reason, in the IEA studies background questionnaires are addressed to different actors in the educational system.

> In today's technologically-centered society, understanding how to improve student learning in mathematics and science [and fostering reading achievement] is vital for educational policy makers, as well as principals, teachers, and parents. A strong foundation in [reading

> ability,] mathematics and science is crucial for student's academic and professional development, and fundamental to the prosperity and welfare of the global community (Hooper et al., 2013, p. 61 [supplements by the authors]).
>
> [The studies collect data] about how educational systems throughout the world deliver and promote learning …. These data on system structure, school organization, curricula, teacher education, and classroom practices reveal many pathways to teaching and learning. In particular, when compared across countries and in relation to student achievement, this information can provide insight into effective educational strategies for development and improvement (ibid.).

Accessing this information to gain a rich background for analyzing students' achievement helps to understand the current situation in educational systems and helps to compare it between different educational systems.

Specific challenges occur if the briefly presented frameworks are transferred into the context of trend studies. Besides the requirement to get appropriate and useful information to understand educational systems, the chronological perspective is a challenge for itself as will be shown in the following section.

2.2.3 Challenges by designing background questionnaires in trend studies

By establishing trend studies, the assessment of background data is a major challenge. The main focus of the IEA studies lies on the accurate assessment of students' achievement. Based on this condition, most of the assessment time is allocated to test administration. Relatively speaking, only a small portion of the time is used to obtain background information from the students. Moreover, the time that is planned to complete the home questionnaire by the parents is limited, since otherwise the response rate may drop or systematic cancellation for specific questions may occur. Given the fact that the background database is limited, it becomes even more crucial which questions are used and which are fruitful for trend analysis. Following the rule that the same information needs to be gathered in the same way – by either comparing different countries among each other or comparing countries' development over time – it is difficult to react to a dynamic educational system. To assess changing conditions in an appropriate way, it is necessary to adapt the background questionnaires accordingly.

Once scales in the background questionnaires are repeated over different cycles of a study, the usage of *item response theory* (IRT) helps to make comparable scale scores both over time and between countries (see Mullis, Martin & Hooper, 2016). By using appropriate IRT models, sources of bias in the context of the assessment can be detected. After identifying the bias, it is possible to encounter it in specific (scaling) models or enhance the scales for next cycles or stages of the study. Obtaining appropriate and reliable responses in each cycle means that if changes in the

scales need to be undertaken, comparisons over cycles of studies are not possible. Nevertheless, the alternative to maintaining scales with inappropriate characteristics would lead to misinterpretation regardless. As described by Mullis et al. (ibid.), the IRT approach is also used in the TIMSS and PIRLS studies. In general – as already mentioned – this is the state of the art to derive comparable scale scores. However, establishing this technique on a limited number of items (for each scale) causes other kind of challenges, as highlighted by Gustafsson and Rosén (2014). The core question (ibid.) that arises in the context of anchoring tests with only a few items relates to how reliable the concurrent modeling of the scales is, e.g. over time. As a conclusion, it should be noted that the items on which both the scales and the anchoring over time are based need to be selected very carefully.

In addition to these model-based issues, He and van de Vijver (2013b) highlight the necessity of carefully constructing questions and scales for cross-cultural studies like large-scale assessments. Following their research, the authors also recommend a careful investigation of the scales and items (e.g. by *differential item functioning* (DIF), *confirmative factor analysis* (CFA) or multilevel (HLM) approaches) as well as standardized test settings (defined as basic requirements in the TIMSS and PIRLS study design). Furthermore, by conscientiously considering possible sources of bias (e.g. construct, method and item bias), it needs to be distinguished in the process of analysis which kind of reasons could lead to variances in the answers between persons, groups or countries (ibid.). Some of them may be explained as cross-cultural effects given that they could likely be based on different response behavior or they simply guide to the fact that there are objective differences observed. Based on their research, He and van de Vijver (2013a) find empirical evidence for a cross-ethnic general factor by answering Likert-type scales, which could – under consideration of the measurement level of the scale (taking into account the measurement equivalence, such as construct-, measurement unit- and full score-equivalence) – be used for comparisons between countries. Nonetheless, by looking more closely at the interpretation of scales, it becomes obvious that depending on the group of answering persons, more latent information could lie behind the differences in the scales that need to be explored in depth.

At this point it becomes evident that the interpretation of scales is not only a technical driven approach that enables comparisons but it is also tightly associated with the interpretation of the scale – or in other words with its validity. The following section gives an overview of the changing interpretation of validity in social science.

2. The question of validity

With the development of tests during recent decades, the meaning of test criteria has also changed. The most affected criterion is the validation approach. The *Journal of Educational Measurement* dedicated a special issue to the topic of 'validity'. In the leading article, Kane (2013) describes the development of the understanding of validity to date:

> Criterion-Based and Content-Based Approach: By around 1915, the notion of criterion validity was in use. With a criterion measure that was assumed to approximate the 'real' value of the attribute of interest, validity could be evaluated in terms of the relationship between test scores and criterion scores (Thorndike, 1918). The early work on criterion-related validation mainly seems to have addressed applied problems in selection and placement with criteria specified in terms of desired outcomes (von Mayrhauser, 1992) (Kane, 2013, p. 4).
>
> The Construct Model: In their conceptualization of construct validity, Cronbach & Meehl (1955) shifted the focus from the development of a test for a given interpretation to the relationship between the test and a proposed interpretation. They developed their construct validation framework in terms of then-current views in the philosophy of science that theoretical constructs would be implicitly defined by their roles in a theory (ibid., p. 5)
>
> Argument-Based Approach to Validation: Cronbach (1982, 1988) and House (1980) proposed that the logic of evaluation argument could provide an effective framework for validation, and Cronbach (1988) suggested that a *validity argument* [emphasis in original] could provide an overall evaluation of the intended interpretations and uses of test scores by examining the evidence for and against the claims being made, including any evidence relevant to plausible alternate interpretations and uses. The analysis 'should make clear, and to the extent possible, persuasive, the construction of reality and the value weightings implicit in a test and its application' (Cronbach, 1988, p. 5) (ibid., p. 8).

Technically speaking, the mentioned 'tests' do not need to be achievement tests; rather, the whole arguments also remain valid for the assessment of background information. Furthermore, the answers to scales of background questionnaires are not interpreted as right or wrong – as is the case for achievement tests – but rather as a reflection of the agreement with other relevant constructs. These can be scales of interest and motivation as well as socio-economic or socio-cultural welfare, etc. The argument concerning how the constructs – based on either tests or scales within a questionnaire – have been interpreted as valid measures remains the same. Kane (2013, p. 3) highlights:

> Validity is not a property of the test. Rather, it is a property of the proposed interpretations and uses of the test scores. Interpretations and uses that make sense and are supported by appropriate evidence are considered to have high validity (or for short, to be valid), and interpretations or uses that are not adequately supported, or worse, are contradicted by the available evidence are taken to have low validity (or for short, to be invalid). The scores

generated by a given test can be given different interpretations, and some of these interpretations may be more plausible than others.

In other words, the argument-based approach induces a pragmatic perspective, whereby the validity is taken as the range of its interpretation and practical implications. It calls for "a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses" (Kane, 2006, p. 17).

Following this argumentation, a counting of positions in rank-order tables – like league tables – does not make sense: as soon as background variables are used to show differences between groups or are used as independent variables, it must be ensured that the content and interpretation related to the background variables (or scales) can be interpreted between countries or groups in the same way. If this cannot be ensured, it is very likely that the result will be like comparing apples and oranges.

In the context of LSAs not only the value of indicators or scales by itself has to be taken into account. Also the (country) specific circumstances are important to be considered. Social, cultural, economic, and historical background information need to be reflected to understand for example the role of the number of books at home. A possible valid interpretation of the number of books at home is not only the pure number of printed objects but its value gives also an (indirect) indication concerning the social, cultural, and economic status of the families. Looking on the correlation of achievement and the number of books at home gives a general (universal) tendency but to understand what the number of books at home 'really' measure one has to reflect its validity. By comparing effects of such variables (or scales) between countries in the context of LSA it is important to reflect if similarities and differences can be interpreted in the same manner. To avoid over-simplifications of effects over all participating countries it could be a constructive strategy to focus on a selection of countries (e.g. EU, OECD, Scandinavia, East-Asia) for which the proportion of between country variances is more likely smaller than in not selected countries. By taken e.g. cultural and historical background into account observations between the observed countries appear in a different light. In the context of the number of books at home and its influence on reading achievement in school it might be of interest to take into account whether a country has a written or an oral tradition for the transmission of stories and fairytales. This cast a differentiated light on books as a cultural symbol.

3. Strategies for the development of background questionnaires

Since nobody can predict future developments in educational systems, how large-scale assessments (and their instruments) should be designed to provide answers

to upcoming questions in the 21st century reflects one of the most important challenges.

Essentially two perspectives can be distinguished, namely a research-oriented approach and a challenge-oriented approach. Nevertheless, both approaches need to follow the standards of modern test theory taking into account the measurement equivalence between countries (see He & van de Vijver, 2013a) and over time, as well as considering the argument-based perspective on the validity of the indicators and scales used. As highlighted in the 'Education and Training Monitor' of the European Commission (European Commission, 2016), international comparisons should only be based on indicators that are internationally comparable. This sounds easy, but it is difficult to realize. The question of validity of the information remains due to the fact that cultural and historical differences in the interpretation of background variables influence its validity.

In this sense, for the development of background questionnaires in large-scale assessments it seems to be a fruitful strategy to consider empirical-based theoretical models that model the improvement of educational systems. In his meta-meta analysis, Hattie (2010) showed that those variables that are most effective in explaining differences in students' abilities address a fruitful cognitive process of learning. This means that distant variables like the structure of the educational system or the organization of everyday school life have (much) less predictive power than those showing how teachers support the learning process of students (e.g. cooperative learning, peer tutoring, direct instruction, feedback). Taking up this idea and combining it with observations of school and classroom research, Klieme and Rakoczy (2008) highlight that the question concerning what good teaching is should be addressed by teams of researchers from different fields of expertise, including empirical research, school effectiveness research and pedagogical psychological research. Research in the field of content knowledge, pedagogical content knowledge and pedagogical knowledge (Shulman, 1987) seems necessary to understand the complex process of teaching and learning in school. In Germany, the experts for pedagogical content knowledge do not have a long tradition in quantitative research. Based on these observations and combining the expertise of both fields seems promising to investigate the learning process in schools and classrooms. To develop scientific-based factual knowledge for teachers, systematic research regarding the input, process and output of the educational process is needed. To create a positive environment for teaching and learning processes, different levels of the educational system need to be evaluated (see i.e. Scheerens & Bosker, 1997). Studies like TIMSS or PIRLS offer the best conditions for this kind of research. Within these studies, the empirical educational research is combined with research about pedagogical content knowledge embedded into the framework of (interna-

tional) school effectiveness research (see Klieme & Rakoczy, 2008). In this context, one possible approach could be oriented towards the dynamic model of school effectiveness research of Creemers and Kyriakides (2006). This approach offers a research frame for understanding cross-sectional data as well as incorporating dynamic approaches that help to understand changes from the perspective of trend analysis. In this context, a third perspective carrying large-scale studies into the future could be to combine traditional background questionnaires and tests (either paper-and-pencil or electronic assessments) with assessment techniques that focus on the teaching and learning processes. This means that the study designs could be opened for approaches that enrich the cross-sectional data, e.g. with video recordings or systematic observations of the teaching and learning process; for instance, by using the *Classroom Assessment Scoring System* (CLASS) instruments (see Pianta & Hamre, 2009).

Returning to Hattie's (2010) findings to deliver promising suggestions for enhancing the quality of teaching and learning, Klieme and Rakoczy (2008) highlight that it seems to be a good strategy to explore characteristics of the acquisition of knowledge and motivational factors of students, such as: (a) structured, explicit, and disturbance preventing guidance of the lessons; (b) supportive, student-oriented classroom-climate (social climate); and (c) cognitive activation (of students). Based on these prerequisites, the sound learning process and the time on task is positively affected, leading to higher performance and conceptual understanding among the students. Moreover, the experience of autonomy, competence and social acceptance also guides to a higher motivation in the learning process of the students (see ibid.). Furthermore, for this process information captured by video recording or classroom observations could also be an interesting source of information.

4. Discussion

Challenges in educational systems are as diverse as the groups of protagonists involved. It has been known for some time that the expectations about the quality and quantity of information that teachers, parents, principals, administrational staff and politics possess strongly vary (Ross & Mählck, 1990). Within international large-scale assessment, it is difficult to address all of these demands in an appropriate way, e.g. due to different historical, social or economic circumstances, the same groups might have different perspectives and questions within the different countries. To enable the countries to address the national-focused interest, the IEA offers the possibility to create national adaptations and extensions of the background questionnaires.

Nevertheless, some questions can only be answered based on international comparisons. This does not necessarily mean that all participating countries need to ask the same questions, although either a union of countries that have the same information needs could be built up or regional contexts of countries could be taken into account for specific adaptations of the background questionnaire. For communities of interest, the Nordic countries (e.g. Iceland, Norway, Sweden, Denmark and Finland), the participating members of the European Union (EU) or the members of the OECD could serve as examples.

The necessity of the requirement of regional analysis is obvious since the population in Europe is becoming increasingly diverse. This emphasizes the role of education in supporting the integration of migrants and strengthening social cohesion (European Commission, 2016, p. 20).

The recommendations which can be drawn from the observations in this article are general statements that shall initiate a general discussion. The aim was not to provide specific variables or scales in general for future large-scale studies. General statements in this direction would be misleading. Without intensive research on specific research areas and without a systematic investigation what are the most important (research) questions a professional development of the next cycles of LSA is not possible.

Nevertheless, looking from a country specific angle on the possibilities that LSAs offer, specific interest can be named. In Germany definitely, the question of how to handle social heterogeneity and the impact of the individual background of the students seems to be one of the most urgent questions for our societies and therefore for all levels in the educational system. However, since the weightage of these challenges differ between the countries the best strategie in the context of LSAs is to negotiate on a common sense and shared objectives that should be addressed by these studies. These questions are truly driven by actual political and social challenges. The more specific the questions are focused the smaller might be the group of countries that agree on common (background) framework. The recommendation of this article is to follow actual developments of strong theories that model the impact of the educational system and to be flexible enough to address actual questions to all countries that are of importance and address questions to group of countries to cover their interest on specific information within their context and situation. Only due to meaningful planning of the design of tests and background questionnaires a solid data base will be available that allows (valid) interpretation of results from LSAs and the development of educational systems taken into account their individual circumstances.

The IEA offers the requisites to fulfil these recommendations therefore we are optimistic that LSAs will continue to offer important, reliable and valid information for the development of educational systems around the world.

Note

1. A common misunderstanding concerning large-scale assessments is that even though the students are tested intensely, the results of the tests cannot be used for individual diagnosis, due to the sampling plan and the specific method of administering the test items. Once single students would lie in the focus of a study, the testing time and herewith the number of items needs to be expanded to obtain reliable individual student measurements (independent of personal or group-related circumstances). The technique used in large-scale assessments is optimized to gather information from a representative sample of students with a sample of items that allow state of the art broad descriptions of the subject matter as well as being generalized to a population of students with known information about general learning circumstances in families, classes, schools and educational systems.

References

Creemers, B. & Kyriakides, L. (2006). A critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, *17*, 347–366.

European Commission. (Ed.). (2016). *Education and training monitor*. Retrieved February 3, 2017, from http://ec.europa.eu/education/sites/education/files/monitor2016_en.pdf

Grønmo, L.S., Lindquist, M., Arora, A. & Mullis, I.V.S. (2013). TIMSS 2015 mathematics framework. In I.V.S. Mullis & M.O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 11–27). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved February 3, 2017, from http://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf

Gustafsson, J.-E. & Rosén, M. (2014). Quality and credibility of international studies. In R. Strietholt, W. Bos, J.-E. Gustafsson & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 19–32). Münster: Waxmann.

Hattie, J.A.C. (2010). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

He, J. & van de Vijver, F.J.R. (2013a). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences*, *55*, 794–800.

He, J. & van de Vijver, F.J.R. (2013b). Methodological issues in cross-cultural studies in educational psychology. In G.A.D. Liem & A.B.I. Bernado (Eds.), *Advancing cross-cultural perspectives on educational psychology* (pp. 39–55). Charlotte, NC: Information Age Publishing Inc.

Hooper, M., Mullis, I.V.S. & Martin, M.O. (2013). TIMSS 2015 context questionnaire framework. In I.V.S. Mullis & M.O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 61–82). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved February 3, 2017, from http://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf

Husén, T. (Ed.). (1967). *International study of achievement in mathematics. A comparison of twelve countries / international project for the evaluation of educational achievement (IEA)*. Stockholm: Almqvist & Wiksell.

IEA. (n.d.). *Brief history of the IEA – More*. Retrieved February 3, 2017, from http://www.iea.nl/brief-history-iea-more

Jones, L.R., Wheeler, G. & Centurino, V.A.S. (2013). TIMSS 2015 science framework. In I.V.S. Mullis & M.O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 29–58). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved February 3, 2017, from http://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: Praeger.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *55*, 1–73. Retrieved February 3, 2017, from onlinelibrary.wiley.com/doi/10111/jedm.11200/epdf

Klieme, E. & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. *Zeitschrift für Pädagogik*, *54* (2), 222–237.

Kyriakides, L. & Charalambous, C.Y. (2014). Educational effectiveness research and international comparative studies: Looking back and looking forward. In R. Strietholt, W. Bos, J.-E. Gustafsson & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 33–49). Münster: Waxmann.

Martin, M.O., Mullis, I.V.S. & Foy, P. (2013). TIMSS 2015 assessment design. In I.V.S. Mullis & M.O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 85–98). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved February 3, 2017, from http://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf

Mullis, I.V.S. (2013). Introduction. In I.V.S. Mullis & M.O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 3–9). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved February 3, 2017, from http://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf

Mullis, I.V.S., Drucker, K.T., Preuschoff, C., Arora, A. & Stanco, G.M. (2012). Assessment framework and instrument development. In M.O. Martin & I.V.S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved February 3, 2017, from http://timssandpirls.bc.edu/methods/pdf/TP_Instrument_Devel.pdf

Mullis, I.V.S., Martin, M.O. & Hooper, M. (2016). Measuring changing educational contexts in a changing world: Evolution of the TIMSS and PIRLS questionnaires. In M. Rosén, K.Y. Hansen & U. Wolff (Eds.), *Cognitive abilities and educational outcomes* (pp. 207–222). Cham: Springer International Publishing.

Mullis, I.V.S., Martin, M.O. & Sainsbury, M. (2015). PIRLS 2016 reading framework. In I.V.S. Mullis & M.O. Martin (Eds.), *PIRLS 2016 assessment framework* (2nd ed.) (pp. 11–29). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved February 3, 2017, from http://timss.bc.edu/pirls2016/downloads/P16_FW_Chap1.pdf

Pianta, R.C. & Hamre, B.K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*,

38, 109–119. Retrieved February 3, 2017, from http://journals.sagepub.com/doi/pdf/10.3102/0013189X09332374

Ross, K.N. & Mählck, L. (1990). *Planning the quality of education. The collection and use of data for informed decision-making*. Oxford: Pergamon Press.

Scheerens, J. & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Shulman, L.S. (1987). Knowledge and teaching. Foundations of the new reform. *Harvard Educational Review*, *57*, 1–22.

Tenorth, H.-E. (1994). *‚Alle alles zu lehren'. Möglichkeiten und Perspektiven allgemeiner Bildung*. Darmstadt: Wissenschaftliche Buchgesellschaft.