SSOAR

Open Access Repository

www.ssoar.info

A meeting report: OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments (2018)

Behr, Dorothée; Zabal, Anouk

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

GESIS - Leibniz-Institut für Sozialwissenschaften

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

MEETING REPORT   **Open Access**

A meeting report: OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments (2018)

Dorothée Behr[*] and Anouk Zabal

In memoriam of Fons van de Vijver

Abstract

This report summarizes the main themes and conclusions from the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, which took place at the Organization for Economic Co-operation and Development (OECD), Paris, in June 2018. The five sessions covered the topics (1) etic (universal) vs. emic (culture-specific) measurement instruments, (2) language- and culture-sensitive development of measurement instruments, (3) international guidelines vs. implementation in countries and by translators, (4) tools and technological developments, and (5) quality control of translations. Key players in the field presented on best practice, lessons learned, and innovations and also made suggestions for moving the field forward.

Keywords: Cross-national, Cross-cultural, Translation, Adaptation, Comparability, Equivalence, Assessment, Test, Questionnaire, Instrument

Introduction

The OECD has recently launched a methodological seminar series to foster discussion among and cross-fertilization across the different stakeholders involved in designing, managing, and analyzing large-scale assessments. The seminars address both theoretical and practical developments (Organization for Economic Co-operation and Development, 2018; Thorn, 2018). With the Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC), to name but two major OECD studies, the OECD is one of the key players and drivers behind comparative assessment and, thus, very well placed to launch this important series. The topic chosen for the 2018 seminar was translation and adaptation of measurement instruments, given its central importance in achieving comparable data. William Thorn from the OECD, together with Dorothée Behr and Anouk Zabal from GESIS – Leibniz Institute for the Social

Sciences (Mannheim, Germany), were responsible for setting up the agenda and bringing together a unique group of speakers with wide-ranging international expertise. The talks by key players in the field, including both academics and practitioners, were followed by 113 international participants. The overarching questions "What is comparability?" and "How can translations be produced that meet the objectives for comparability?" were addressed across different stages of instrument development and production. The agenda was structured along the following topics (see Table 1):

The structure of the seminar reflected the fact that thinking about translation quality and comparability should essentially start at the development stage of the source instrument and not just at the translation stage. After all, if translatability or other comparability issues are only detected once the translation process has started, it is often too late to modify the source instrument to counteract these problems. The presenters in each session were encouraged to present and discuss current implementations and best practice, limitations, and future directions. The sessions

* Correspondence: dorothee.behr@gesis.org
GESIS – Leibniz Institute for the Social Sciences, P.O. Box 12 21 55, 68072
Mannheim, Germany

Table 1 Overview of topics covered at the seminar

Stage	Topics
Source instrument development	1. Etic (universal) vs. emic (culture-specific) measurement instruments
	2. Language- and culture-sensitive development of measurement instruments to ensure comparability, cultural relevance, and translatability
Translation	3. International guidelines vs. implementation by countries and by translators
	4. Tools and technological developments
Quality control	5. Quality control of translations

were organized with a view to triggering a constructive discussion among both presenters and the audience and towards fostering an exchange of ideas between the very heterogeneous players in the area of translation and adaptation of measurement instruments. This report is structured along the seminar topics, as outlined in Table 1.

Etic (universal) vs. emic (culture-specific) measurement instruments

The first session raised the fundamental question as to which kind of measurement instrument is best suited to achieve comparability in cross-national studies. The internationally widely acclaimed researcher Fons van de Vijver (2018) set the scene for the entire seminar with the first presentation. He made a convincing plea for the need to combine both etic and emic instruments. *Etic* instruments rely on the assumption of universally applicable constructs that can be "transported" into other cultures through translation. Advantages of such instruments include the ease of direct cross-cultural comparison and the use of tried-and-tested instruments. *Emic* instruments, on the other hand, rely on culture-specific operationalization of constructs; advantages of these instruments include increased ecological validity and construct coverage as well as the reduction of Western bias in the case on non-Western countries. Studies such as PISA or PIAAC predominantly follow an etic approach that calls for translation of source instruments and allows for only minor types of adaptations within a clearly defined framework. With the increase of countries and thus of cultural variation in such studies, three types of paradoxes come to the fore: (a) the "analysis paradox," according to which fewer conclusions can be drawn because scalar equivalence, the highest form of equivalence which allows for direct comparison of means, is increasingly difficult to achieve; (b) the "test design paradox," according to which the cultural coverage decreases since it is necessary to focus on content that has at least some relevance in all participating countries; and (c) the "test length paradox," according to which more items lead to

more design and analysis problems—longer instruments may be more informative for the different stakeholders, but they are also less likely to show a high level of invariance. Against this backdrop, van de Vijver proposed to combine etic and emic approaches in large-scale assessments as a way to maintain the advantages of both approaches. The strength of the combined approach was illustrated with findings from personality research in South Africa. According to this research, which took into account cultural specificities, social-relational aspects were identified as an important part of personality in South Africa. This would not have been uncovered had a standard personality instrument such as the Big Five been used (for more details, see Fetvadjiev, Meiring, Van de Vijver, Nel, & Hill, 2015).

While van de Vijver advocated for an integrative approach, Klaus Boehnke (2018) made a more radical plea for an emic, i.e., a culture-specific approach to instrument development (see also Boehnke et al., 2014). He presented a case study on an exclusively emic development of an instrument measuring "paternal warmth." As part of this study, students from five different cultures and languages independently developed items measuring paternal warmth for their specific culture of upbringing. Statistical analyses of these items were subsequently carried out. However, even this approach required an additional etic reference variable as an external validation variable for the emic scale. Furthermore, Boehnke compared the emic results to results from the internationally established etic Nurturant Father Scale. He concluded that the emic scale achieved better construct validity than the etic scale. According to Boehnke, "latent variable equivalence of emic scales across cultures" (p. 27) could be shown using his approach. Even though Boehnke's approach did not (yet) establish metric or scalar equivalence, it recommended, as did van de Vijver's approach, that the universal, one-size-fits-all approach to instruments needs to be questioned, possibly even revised to do justice to the heterogeneous needs and particularities of the diverse countries and cultures in a study.

Language- and culture-sensitive development of measurement instruments to ensure comparability, cultural relevance, and translatability

The second session focused on best-practice procedures used to develop source instruments that pave the way for comparability, cultural relevance, and translatability. The first speaker, Brita Dorer (2018), presented the so-called *advance translation approach* from the European Social Survey (ESS), the methodological flagship in the social sciences. First mentioned by Harkness and Schoua-Glusberg in the 1990s (1998), advance translations are translations of a pre-final source questionnaire. They are undertaken with the goal to identify

translatability and cultural problems early on so that these can be mitigated in an improved subsequent source version. Advance translations are based upon two principles, namely that the production of optimal instrument translations essentially starts at the design stage (Smith, 2004) and that translation problems are often only detected once an actual translation is attempted (as compared to "just looking" at the source instrument). The feedback from advance translations is typically used to revise or annotate the source instrument to make it suitable for cross-national implementation. Dorer described the design of the ESS advance translations and referred to similar methods such as translatability assessment.[1] She illustrated the usefulness of the approach with some examples: For instance, an item referring to "dependence on energy imports" was identified as problematic given that not all countries in the ESS rely on energy imports. In another instance, a construct-relevant distinction between "justice and fairness" was seen as problematic given that not all languages can linguistically maintain the appropriate nuances.

The second presenter in this session, Olivieri (2018), gave an overview of the *Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations* by the International Test Commission (2018). The guidelines address "development and adaptation issues across all aspects of tests that may impact fairness and validity when assessing linguistically and culturally diverse populations" (Oliveri, 2018, p. 5). Oliveri's presentation focused on the first section of the guidelines, namely linguistic aspects to take into account during test development and adaptation. For instance, she recommended including different linguistic groups in the design of tests to identify translation hurdles—and to avoid regional vocabulary, ambiguous words, and the use of construct-irrelevant product names, geography reference, and the like.

Overall, both talks in this session emphasized that consideration of translational, linguistic, and cultural aspects needs to be firmly integrated into source instrument development to prevent problems at later stages. What is now widely regarded as best practice in major large-scale studies should also be a role model for smaller studies.

International guidelines vs. implementation and perception of guidelines

The third session reflected on the role and perception of translation guidelines, which are produced and followed by many international studies. Behr (2018b) set the scene for this session by giving an overview of different types of guidelines. She started by differentiating between "overarching guidelines" on the cross-cultural research process in general, such as the *ITC Guidelines for Translating and Adapting Tests*[2] and the *Cross-Cultural Survey Guidelines*[3], and "project-specific guidelines." The overarching guidelines provide good practice and a framework for quality assurance and control in general and ideally inform project-specific guidelines. The project-specific guidelines were further subdivided into "general guidelines" and "detailed guideline." The former describe the translation approach in a specific study and thus ensure a uniform understanding of translation needs and procedures among all stakeholders in the study while the latter give concrete instructions at the item level, such as how to translate or adapt a particular term. After outlining the responsibilities of guideline developers (e.g., using clear language and defining feasible processes) and national teams (e.g., selecting appropriate staff and taking translation seriously), she concluded that guidelines were not meant to replace translation and decision-making competence—their goal would rather be to "empower" competent translation staff to take the right decisions. Finally, Behr called for certain guidelines (e.g., on different translation procedures such as double translation and reconciliation) to be backed empirically. This would allow good practice procedures that are common across many disciplines to gain a strong(er) foothold in the research community.

The second talk by Brita Upsing (2018) provided the first-of-its-kind assessment of how different translation players actually perceive translation guidelines and the frameworks set by international large-scale assessment studies, such as PIAAC and PISA (see also Upsing & Rittberger, 2018). For this, she analyzed translation guidelines from the first round of PIAAC and also qualitative data from 20 interviews with translators, verifiers, and project managers. She identified the value of translation training for project managers, who are typically not translators, in developing their understanding of what good translation is. She also emphasized the importance of training for translators, who are typically not assessment experts, by improving their understanding of the special requirements of cognitive assessments. Translators should be aware of item design characteristic that need to be maintained in the translation, for example, the role of distractors and the importance of literal or synonymous matches between stimulus and question. If

---

[1]Translatability assessment is described in some detail by Acquadro et al. (2018).

[2]https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf. Accessed 28 November 2019.

[3]https://ccsg.isr.umich.edu/. Accessed 28 November 2019.

translators understand the "big picture," they will be in a better position to take appropriate decisions. Furthermore, Upsing expressed a cautionary note: Detailed, that is, item-by-item guidelines might be misunderstood and prevent critical thinking. Hence, their number, length, and content should always be weighed against what is really needed. Ultimately, Upsing concluded that professional translators were key to the entire endeavor and should be involved throughout the various translation steps (e.g., also in the reconciliation step). The two talks encouraged the field of instrument translation to (further) tailor guidelines, training, and procedures to the different user groups and to actively seek feedback on how to improve guidelines and, overall, translation procedures.

Tools and technological developments

The translation industry is strongly shaped and influenced by translation tools and technological developments, such as developments in machine translation. The fourth session demonstrated that (some) large-scale surveys are already embracing these tools and developments to ensure comparable and high-quality translations.

In the first talk, Pettinicchi and Philip (2018) raised the question as to what extent machine translation could be used for an additional quality control step in SHARE, the Survey of Health, Ageing, and Retirement in Europe. SHARE is implemented as a computer-assisted interview. The study uses a large number of fills in the survey questionnaire, that is, dynamic text that is different depending on respondent characteristics (e.g., gender) or answers to previous questions. Dynamic text significantly increases the translation workload and the complexity of a questionnaire, and testing fills using human resources is a significant cost factor. Against this backdrop, Pettinicchi and Philip explored automatic tools to help identify avoidable mistakes such as spelling errors, missing translations, and flipped translations (e.g., "employed" followed by "unemployed" in answer categories instead of the other way round). Focusing on the language pair French and English and using an experimental design, they compared the off-the-shelf solutions Google, deepl, and Bing—all based on general corpora and neural networks machine translation—with Moses, a specifically trained in-house solution based on news and political corpora and a phrase-based approach to machine translation. The experiment involved a back translation from French into English using the various machine translation solutions, computing several indexes (e.g., on similarity between the original items in English and the English back translation from French), and flagging items to be followed up by human translators. In this experiment, the market solutions turned out

to be more effective. Regardless of this outcome, Pettinicchi and Philip see a need for a machine translation tool that is trained on domain-specific corpora, that is, on bilingual corpora from surveys, and that is thus sensitive towards the particularities of questionnaires (e.g., typical wording of response categories, typical form of addressing respondents, or recurring questionnaire-specific terms such as "showcards").

In the second talk, Danina Lupsa (2018) from cApStAn, the company that performs linguistic quality assurance and control in many large-scale assessments and other surveys, including PIAAC and PISA, introduced best-practice procedures and technology for project preparation and project execution in large-scale studies. Rather than re-inventing the wheel for survey research, existing technology, which complies with international standards, should be utilized. Technical project *preparation* should include segmenting the source instruments at the sentence level so that text can be translated and checked sentence-wise in translation tools. Furthermore, segmenting allows for a more effective use of translation memories, which identify and provide similar or identical previously translated text segments. At the level of project *execution*, translators should use computer-aided translation (CAT) tools. These tools are specifically designed for translators and provide a translation editor with a bilingual display of source and target text, translation memories, or glossaries with pre-determined terminology, to name but a few of the standard features of CAT tools. Linguistic quality control should make use of automated checks on completeness, consistency, spelling, formatting, and further pre-defined requirements, thus enabling linguists to focus on important meaning-related equivalence issues. Finally, Lupsa stressed that technical aspects of tools, files, and procedures should be considered early on during instrument design and be jointly worked on by developers, linguists, and tools experts.

Quality control of translations

The fifth and final session focused on quality control procedures. Stephen G. Sireci (2018) presented statistical and qualitative approaches for facilitating comparability in multilingual assessments. He noted that when dealing with validity issues the purpose of an assessment always needs to be considered. For example, evidence of comparability is not needed if interpretations of test scores are to be made within a language group, whereas evidence of comparability is urgently required if scores are to be compared across language groups. Producing comparable instruments calls for high-quality adaptation/translation procedures, statistical analyses of structural equivalence and differential item functioning, qualitative

analysis of item and method bias, and ultimately a sound validity argument for comparative inferences. Sireci looked more closely at different statistical methods for assessing equivalence, in particular confirmatory factor analysis (CFA) and multidimensional scaling (MDS).[4] He called for cautious interpretation of statistical results in general and viewed MDS as an under-used method for evaluating equivalence. Given the challenges inherent in multilingual assessment—despite all efforts put into quality assurance and control procedures—he proposed an "index of comparability" to inform data users about the level of equivalence of data. The exact specification of such an index is a matter for future research. The index should include findings from both qualitative and quantitative studies and flag instances of lack of measurement equivalence.

In the second talk, Steven Dept (2018) from cApStAn traced the evolution of translation verification over the past two decades. Simply put, translation verification consists of an additional check by a third person. Prerequisites for verification include a set of criteria, a common understanding of this set of criteria among all verifiers, and a reporting method. Moreover, verification needs to be embedded in a larger quality framework including detailed translation guidelines against which to verify and training of verifiers, to name but two features of such a larger framework. An even wider framework currently implemented in international assessments is *linguistic quality control* (LQC). In LQC, verification by a third person, including a thorough documentation of deviations (using pre-determined categories, and possibly severity or follow-up codes) is the first step. The subsequent steps include monitoring of corrective actions, analysis of field trial results, linking these results back to the translation output, and quantitative and qualitative reporting. Dept envisions a future where all stages of verification and subsequent feedback take place in a dedicated Translation Quality Management environment, where quality evaluation takes place in real time and produces metrics for a dashboard to be accessed by all relevant stakeholders. Furthermore, given the tension between budget and quality requirements in international studies, it might be feasible to have automated quality checks on text elements not crucial for measurement and human verification on key text elements.

However, this would require a shift towards more upstream preparation work before the translation actually starts.

The renowned expert in the field of cross-cultural assessment, Ron Hambleton (2018), concluded the seminar with his keynote on "Translating/Adapting Achievement Tests: PISA Guidelines, ITC Guidelines or a Mixture?" He presented and—at the same time—dispelled five common myths regarding tests adaptation across languages and cultures. These myths were, first, that everyone who knows two languages can translate well;[5] second, that a good literal translation ensures validity; third, that judgmental reviews are sufficient to identify problems; fourth, that a back translation design and the use of bilinguals to compile empirical data are sufficient for validation; and fifth, that all constructs are universally applicable and can be transferred into another culture using translation. To dispel these myths, he argued, for instance, that translators need to be knowledgeable about principles of test development and the two cultures involved. Furthermore, a simple back translation design neglects to look at the translation itself and can thus never assess the suitability for the target population. Hambleton introduced the Second Edition of the ITC Guidelines for Translating and Adapting Tests (2017), which further debunks these myths and also takes into account advancements in translation methodology and statistical methods since the First Edition from the 1990s. After comparing the ITC guidelines to the PISA translation guidelines (2018) with the goal to identify areas where these two guidelines could learn from each other, he concluded that the ITC guidelines could benefit from improved descriptions of the translation approach. For the PISA guidelines, he recommended to expand the review role of translators (e.g., by having translators make use of the Item Review Form by Hambleton & Zenisky, 2011) and to add at least some form of small-scale (cognitive) empirical study to the design (even 50 respondents would be sufficient for simple statistics).

The seminar ended with a summary of the key messages by the organizers of the event: (1) The presumed rivalry between emic and etic approaches in cross-cultural measurement may be resolved by focusing more on how they can complement each other and how an integration of both may yield a more appropriate approach. (2) Design and analysis should be driven by the clearly defined purpose of the study. (3) Collecting comparable data is complex and can only succeed if test developers, psychometricians, translation experts, and IT and platform experts collaborate early on. (4) Qualitative

---

[4]In brief, CFA is a method for testing whether the factor structures across the source and its translated/adapted versions are equivalent. The factorial structure is defined in advance and then tested. MDS is a way to graphically show the intercorrelations between items. A criterion for comparability is whether or not items are located in the same region. Dividing lines for the different regions are derived from theory rather than from a statistical procedure. For more information on CFA and MDS and further references, see International Test Commission (2017) or Braun and Johnson (2010).

[5]See also Behr (2018a) on the different competences required by translators.

and quantitative evidence needs to be triangulated so that researchers are in a strong position to evaluate equivalence and/or improve future studies based on lessons learned. (5) In times of rapid technological growth with continuously evolving statistical methods and IT and translation tools, it is important to be receptive to new opportunities and embrace innovations if these can be useful for international studies.

All in all, despite significant advancement in the field, comparability challenges and issues can never be ruled out completely. Against this backdrop, careful and transparent documentation of procedures, their strengths, and caveats remains a must to inform all future data users adequately. An "index of comparability," as suggested by Sireci, may be a further step to take one's responsibility towards the research and user community seriously.

Abbreviations
CFA: Confirmatory factor analysis; ESS: European Social Survey; ITC: International Test Commission; MDS: Multidimensional scaling; OECD: Organization for Economic Co-operation and Development; PIAAC: Programme for the International Assessment of Adult Competencies; PISA: Programme for International Student Assessment; SHARE: Survey of Health, Ageing, and Retirement in Europe

Authors' contributions
DB and AZ co-organized the seminar with the OECD. DB drafted the first version of this report, which was then reviewed and amended by AZ. Both authors read and approved the final manuscript.

Authors' information
Dorothée Behr is a senior researcher at GESIS-Leibniz Institute for the Social Sciences in Mannheim (Germany). Her services and research focus on questionnaire translation and adaptation as well as cross-cultural web probing. She has been responsible in PIAAC Cycle 1 and 2 for the translation guidelines for the background questionnaire. She also worked on the European Social Survey (ESS) and is now a member of the ESS Translation Expert Panel. She gained her PhD from the University of Mainz (Germany) on the topic of questionnaire translation.
Anouk Zabal is a senior researcher at GESIS-Leibniz Institute for the Social Sciences in Mannheim (Germany). Her work focusses on different survey methodological topics, including survey operations, item development, and questionnaire construction, as well as the translation and adaptation of survey instruments. As a member of various international consortiums (PIAAC Cycles 1 and 2, ESS, expert for ALL) and through her work in the National Project Management for PIAAC in Germany, she has extensive experience with international large-scale surveys and specifically the challenges of translating and adapting measurement instruments—both from the perspective of writing international translation guidelines, as well as from the country side as producer of national instrument versions.

Availability of data and materials
All presentations are available here http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm

Competing interests
The authors declare that they have no competing interests.

References
Acquadro, C., Patrick, D. L., Eremenco, S., Martin, M. L., Kuliś, D., Correia, H., et al. (2018). Emerging good practices for translatability assessment (TA) of patient-reported outcome (PRO) measures. *Journal of Patient-reported Outcomes, 2*(1), 8.

Behr, D. (2018a). Translating questionnaires for cross-national surveys – a description of a genre and its particularities based on the ISO 17100 categorization of translator competences. *Translation & Interpreting, 10*(2), 5–20.

Behr, D. (2018b). Translation and adaptation guidelines in large-scale studies - what they should, can and cannot achieve. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Boehnke, K. (2018). Latent variable equivalence: towards a quantitative emic approach to cross-cultural comparisons. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Boehnke, K., Arnaut, C., Bremer, T., Chinyemba, R., Kiewitt, Y., Koudadjey, A. K., et al. (2014). Toward emically informed cross-cultural comparisons: a suggestion. *Journal of Cross-Cultural Psychology, 45*, 1655–1670.

Braun, M., & Johnson, T. P. (2010). An illustrative review of techniques for detecting inequivalences. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 373–393). Hoboken: Wiley.

Dept, S. (2018). Two decades of translation verification: relevance and limitations of metrics, evaluation reports and standardization. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Dorer, B. (2018). Advance translations carried out in the European Social Survey (ESS) – a means for improving the translatability of the source questionnaire in a cross-cultural survey? Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Fetvadjiev, V. H., Meiring, D., Van de Vijver, F. J., Nel, J. A., & Hill, C. (2015). The South African Personality Inventory (SAPI): a culture-informed instrument for the country's main ethnocultural groups. *Psychological Assessment, 27*, 827–837.

Hambleton, R. K. (2018). Translating and adapting achievement tests: PISA guidelines, ITC guidelines, or a mixture? Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46–74). New York: Cambridge University Press.

Harkness, J. A., & Schoua-Glusberg, A. (1998). Questionnaires in translation. In J. Harkness (Ed.), *ZUMA-Nachrichten Spezial 3: Cross-Cultural Survey Equivalence* (pp. 87–127). Mannheim: ZUMA.

International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). [www.InTestCom.org]

International Test Commission. (2018). ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations. [www.InTestCom.org]

Lupsa, D.. (2018). Translation technology in ILSAs: striving to combine apps, discernment and capacity-building. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-

gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Olivieri, M. E.. (2018). Insights from the International Test Commission Guidelines on Designing and Developing Large-Scale Assessments of Culturally and Linguistically Diverse Populations. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Organization for the Economic Co-operation and Development (2018). Upcoming events. http://www.oecd.org/skills/piaac/events/ Accessed 12 Apr 2019.

Pettinicchi, Y, & Philip, Jeny Tony (2018). Automatic translation verification: can we rely on MOSES? Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Sireci, S. G. (2018). Statistical and qualitative approaches for facilitating comparability in cross-lingual assessments. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Smith, T. (2004). Developing and evaluating cross-national survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 431–452). Hoboken: Wiley.

Thorn, W. (2018). Welcome and introduction. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Upsing, B. (2018). How do country teams and translators work? The impact of guidelines. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Upsing, B., & Rittberger, M. (2018). The translator's perspective on translation quality control processes for international large-scale assessment studies. *Translation & Interpreting, 10*(2), 55–72.

van de Vijver, F. (2018). Emic and etic measurement in large-scale assessments: towards an integration. Paper presented at the OECD-GESIS Seminar on Translating and Adapting Instruments in Large-Scale Assessments, Paris. http://www.oecd.org/skills/piaac/events/oecd-gesisseminarontranslatingandadaptinginstrumentsinlargescaleassessments7-8june2018-presentations.htm Accessed 9 July 2019.

Publisher's Note