PROBLEMS AND ISSUES IN TRANSLATING

INTERNATIONAL ACHIEVEMENT TESTS

The paper reviews research and findings on problems and issues faced when translating international academic achievement tests. The purpose is to draw attention to the problems, to help to develop the procedures followed when translating the tests and to provide suggestions for further research. The problems concentrate on the following: the unique and demanding purpose of the translation task, the partly contradictory task specifications and translation instructions, the indecision as to whether to produce one or two target versions, the indecision as to whether to use one or two source versions, inadequate revision and verification, deficient translator competences, and a lack of time. To solve the problems, the paper suggests the following: ensuring that the translation guidelines provide a right, unequivocal and balanced picture of the purpose of the translation task; ensuring the equivalence of the two source versions; putting more emphasis on revision, and ensuring that the verification is sufficiently thorough; using only qualified translators, providing them with training in test translation, and including also subject matter and testing specialists in the translation teams; and allotting sufficient time to the translation work. However, the main lesson from the review is that more research in the field is badly needed.

INTRODUCTION

Recent years have witnessed a considerable increase in the interest in international academic achievement studies. Studies have been conducted, for example, by the International Association for the Evaluation of Educational Achievement (IEA), the Educational Testing Service (ETS), Statistics Canada (STATCAN), the Organization for Economic Co-operation and Development (OECD), the United Nations Educational, Scientific and Cultural Organization (UNESCO), and the Southern Africa Consortium for Monitoring Education Quality (SACMEQ). In all these studies, a common test has been used which has been translated[1] into the languages of the participating countries.

When translating international achievement tests into multiple languages, it is important to ensure that the versions are equivalent, or comparable, to each other – that they measure the same construct and are equally easy to answer. For this to be the case, the mental effort required by testees to respond to the items needs to remain the same across languages. No version must place a heavier cognitive load (Sweller**,** 1988; see also e.g., Rueda, 2011, pp. 93-4), or consume more of the limited (processing)

[1] In this paper, the term 'translation' is used to refer to the process of reproducing a text originating in one language and culture for use in another language and culture. The term thus covers all kinds of between-language meaning transfer, from close, or literal, translation to adaptation, or the making of major changes to the target version so as to make it more suitable for the target population (e.g., changes in currency or measurement units). The term 'adaptation', accordingly, refers to a special subtype of translation. The way the two terms are used in this paper coincides with how they are employed in both Translation Studies and international achievement studies. However, it differs from the use recommended, for example, by Hambleton (2005, p. 4) and preferred in psychological tests, where the term 'adaptation' often refers to the entire process of preparing a test constructed in one language and culture for use in another language and culture (from deciding whether a test can measure the same construct in a different language to checking the equivalence of the adapted test) and to making the target version not only linguistically but also culturally and psychologically appropriate for the target culture, whereas 'translation' only refers to literal translation.

capacity of testees' working, or short-term, memory compared to the other versions. This, in turn, requires, among other things, that all versions – not only the items but also the accompanying stimulus texts – be equally easy to understand. If this is not the case, if some items or stimuli are harder to understand than the others, more working memory is needed to decode and make meaning of them and less memory is left for actually responding to the items. Readers of these versions would then be at a disadvantage, which, in turn, would jeopardize the validity of inferences made on the basis of the test.

Thus, if, for example, the source version uses a literal match to link an item to its stimulus text (i.e., exactly the same word or expression in both), a literal match should also be used in all target versions, because other types of matches (e.g., synonyms) require the testee to do more inferencing and thus consume more memory capacity (cf. Kirsch, 2001; Mosenthal & Kirsch, 1991). Or if there are no explicit markings (e.g., conjunctions) to signal a link or meaning relationship (e.g., cause-and-effect) in the source version, no explicit signals should be used in the target versions either, because these would reduce the amount of inferencing and cognitive processing required of the testee (cf. Kemper, 1983). Or if natural, idiomatic and authentic language is employed in the source version, similar language should also be used in all target versions, because unnatural, odd and cumbersome language cannot be processed in as large chunks and as automatically and effortlessly as natural and idiomatic language (cf. Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983).

Clearly, translating international achievement tests and establishing their equivalence is a most responsible and demanding task. Rigorous translation procedures and have therefore been developed to ensure high-quality and equivalent translations. However, no unanimity seems to have been reached on how to best translate these tests, judging from the differing procedures followed in the organizations implementing the tests. Besides, both research and experience suggest that there have been problems when implementing the procedures and translating the tests (e.g., Hambleton, 2002, 2005; Harkness, Villar & Edwards, 2010) and that equivalence may

therefore not always have been attained (e.g., Bechger et al., 1998; Blum, Goldstein & Guérin-Pace, 2001; Bonnet, 2002; Ercikan & Koh, 2005; Karg, 2005).

PURPOSE AND SCOPE OF THE PAPER

This paper reviews research and findings on problems and issues encountered when implementing translation procedures and translating tests in international achievement studies. The purpose is to draw attention to these problems and issues, to help to develop translation procedures in these studies and to provide suggestions for further research in the field.

The paper limits itself to problems and issues of *translation*, or "the process of reproducing a text originating in one language and culture for use in another language and culture", and factors having an impact on it (e.g., translators, time), because it is in this area in particular where research is lacking. The problems are discussed within the frame of Translation Studies, whose principles should guide all translation work, including test translation; however, it seems that in test translation the principles have not always been heeded (see also Harkness, 2003, p. 36; Harkness, Villar & Edwards, 2010, p. 118). To a lesser degree, reference is also made to what can be learned from other cross-cultural studies, such as large-scale psychological and social surveys (e.g., the European Social Survey, [ESS]), where translation issues have been dealt with for a longer time than in academic achievement studies.

Problems and issues of adaptation (in the broad sense of the word; see Footnote 1) fall outside the scope of this paper. This involves, for example, methods for evaluating and testing the equivalence of the translations: having examinees take the test and comment on it (see e.g., Hambleton, 2005, pp. 14-6), or statistical techniques and issues related to their use in validating

and equating translated tests (see e.g., Angoff & Cook, 1988; Sireci, 1997) – such as their ability to evaluate only statistical item difficulty, not overall language or text difficulty[2]. Also ignored will be issues concerning the compiling, writing and quality of source materials, even though these are known to play a significant role in determining translation difficulty and quality (Hambleton, 2005, p. 26; Harkness, 2003, p. 46; see also ESS, 2010, pp. 6-7). A relatively large body of research already exists on source text characteristics that threaten translation equivalence (e.g., Allalouf, 2003; Author, 2007; Brislin, 1986, pp. 143-9; Elosua & López-Jaúregui, 2007; Gierl & Khaliq, 2001; Solano-Flores, Backhoff & Contreras-Niño, 2009); guidelines are also available on how to develop the source instruments (e.g., Harkness, Edwards, Hansen, Miller & Villar, 2010).

It is also important to note that the paper limits itself to the studies conducted by the OECD and the IEA, using as examples – or cases – the OECD PISA (Programme for International Student Assessment) and the IEA TIMSS (Trends in International Mathematics and Science Study) studies. PISA assesses 15-year-old students' knowledge and skills – acquired either in school or somewhere else – in reading, mathematics and scientific literacy (together with some cross-cultural competences) and examines how well students can apply their knowledge and skills to real-life problems and situations (for examples of PISA tasks, see e.g., OECD, 2009b). TIMSS, for its part, looks at how well students (in grades 4 and 8) master the knowledge taught in school mathematics and science curricula (for examples of TIMSS tasks, see e.g., http://www.erc.ie/?p=169). (Other studies arranged by the two organizations include, for example, the Programme for the International Assessment for Adult Competencies, or PIAAC,

[2] In this paper, the term 'difficulty', when talking about language, text or item difficulty, is not used as a statistical concept but as a broader, cognitive concept, describing the processing difficulty faced when reading and answering items (unless otherwise specified).

by the OECD; and the Progress in International Reading Literacy Study, or PIRLS, and the International Civic and Citizenship Education Study, or ICCS, by the IEA.)

The main reason for focusing on OECD PISA and IEA TIMSS and using them as cases is that for other studies, extremely little, if any, data are available on translation problems and issues (and often even on translation procedures). To collect data on translation problems in international achievement studies, a literature search was made, whereby literature was collected from electronic databases (e.g., ERIC and PsychInfo), from search engines (e.g., Google and Google Scholar), from reference sections of relevant studies, and by directly contacting organizations conducting international achievement studies. However, the search showed, not only that research on translation problems in international achievement studies is extremely limited, both in number and in generalizability and/or methodological rigor, but also that almost all existent findings concern PISA or TIMSS. The decision to focus on PISA and TIMSS and to use them as cases was further supported by the following (cf. the principle of purposeful sampling guiding qualitative research; Patton, 2002, pp. 230-43): They are the best known and most extensively documented international achievement studies; they bring together a huge array of cultural and linguistic backgrounds and may therefore be expected to provide a rich collection of translation problems; and they are modern studies that will be conducted also in the future and may therefore be expected to use the latest and most advanced translation procedures (e.g., those recommended in the International Test Commission [ICT] guidelines for translating and adapting tests; e.g., ICT, 2010) and to provide a good picture of current translation problems. From among PISA and TIMSS, more references will be made to the former, because, again, more data are available on it. For the same reason, the most attention, from among the participating countries, will be paid to Finland (and my own findings). Even though the paper largely focuses on OECD

and IEA studies, the translation principles, problems and issues discussed are, for the most part, common to all cross-national studies and can therefore help to develop translation procedures also in these studies.

The paper first briefly describes the translation procedures followed in OECD and IEA studies, then discusses translation problems and issues faced in these studies, and, finally, summarizes the lessons that can be learned from the review.

TRANSLATION PROCEDURES IN OECD AND IEA STUDIES

Forward Translation

In both OECD and IEA studies, the translations are made by following the forward translation procedure. However, the procedures differ, especially during the first phases of the translation process. The procedures are summarized in Table 1 (for a more detailed description of the procedures, see e.g., OECD, 2009a; Olson, Martin & Mullis, 2008).

TABLE 1
The recommended translation and verification procedure in OECD and IEA studies

Step	*Recommended OECD procedure*	*Recommended IEA procedure*
1	For every unit two translators produce two independent target language versions. In PISA (and partly also in IALS) one of the versions is translated on the basis of the English and the other on the basis of the French source version; in other studies	One translator produces the first target language version on the basis of the English source version. The translator may consult subject matter experts.
2	The two independent versions are merged into one national version by a reconciler. The reconciled version may be reviewed for appropriateness of content and terminology by a domain expert. The reconciler then decides on the "next-final" national version.	The national version is reviewed by a translation reviewer.
3	The "next-final" national version is verified by an independent translator (verifier) from the International Project Centre. The verifier makes suggestions for corrections and improvements, marking some of them as obligatory ("key corrections").	The reviewed version is verified by an independent translator (verifier) from the International Study Center. The verifier makes suggestions for corrections and improvements, using a "severity code" (from 1=serious error to 4=acceptable adaptation) to sho
4	The national translators decide on the final versions and have them compiled into test booklets.	The national translators decide on the final versions and have them compiled into test booklets.
5	The verifier checks that the obligatory corrections have been made. The booklets are checked optically for layout errors at the International Project Centre.	A quality control monitor checks whether the verifier's suggestions have been implemented. The booklets are checked optically for layout errors at the International Study Center.

The main differences between the OECD and IEA procedures are the following: First, in IEA studies, the translations are made from one source language (mainly English; however, in more recent TIMSS studies, also an Arabic source version has been produced), but in OECD studies, from one (English; in PIAAC) or two (English and French; in PISA). Second, in IEA studies, only one translation is produced from every source version, but in OECD studies two. Third, in IEA studies, the stage (Step 2) following the production of the first national version only involves reviewing and revising this version, whereas in OECD studies two national versions first have to be merged into one.

In addition to these differences, there is also variation in how the procedures are actually implemented in the participating countries. For example, in Finland, when translating PISA

materials, only one translation has been produced which has then been reworked and revised by two successive national translators. In PISA 2000, the revisions were made almost exclusively only against the English source versions, but in PISA 2009, also against the French versions.

In those OECD studies where two source versions have been provided, the English version has usually been prepared first and the French version has then been translated on the basis of it, the procedure having been largely the same as when translating the national versions. However, in more recent studies, the two versions have been produced more in parallel. In TIMSS, the Arabic version has been translated on the basis of the English version**.**

In practice, the translating takes place on screen.

Translation Teams and Translator Requirements

In both OECD and IEA studies, translation teams are used to translate the test materials. The teams consist of translators (national translators making the first drafts), reconcilers (national translators in OECD studies merging the two first versions into one) or reviewers (national translators in IEA studies checking the accuracy of the first drafts), and verifiers (international translators checking the equivalence of the source and target versions). If needed, the teams may also consult subject matter specialists (see Table 1).

The requirements set for the translators (henceforth used as a generic term for all members of the translation team – translators, reconcilers, reviewers, and verifiers – unless otherwise specified) vary somewhat according to their roles and between the studies. However, usually they are expected to have a perfect command of the target language, an excellent command of the source language, experience in the target culture and with students in the target

population, knowledge of the subject matter, and familiarity with test development. In addition, however, in OECD studies, reconcilers are also said to benefit from knowledge of the other source language and verifiers are even required to have a sufficient command of this language. In IEA studies, literary translators (in PIRLS studies) are expected to have experience in literary translation, translators making the first national versions are required to be experienced translators, and verifiers are expected to be certified translators.

Translation Guidelines and Translator Training

To help translators in the translation work, both the OECD and the IEA provide translators with translation and adaptation guidelines, which contain information on the goal of the translation and the translation procedures and also some more specific translation instructions. However, the guidelines differ between the studies in that in the OECD, they contain a great number of detailed examples of the most common translation problems (e.g., the layout of the translations, how to maintain the difficulty level of the vocabulary and the syntax of the text unchanged, and how to translate the question items) and advice on how to avoid them (e.g., OECD, 1999), whereas in the IEA, they are relatively general, with only a very few specific translation instructions.

In OECD studies, countries are also encouraged to offer training to their national translators, based on the translation guidelines. Verifiers are trained in both studies. This training is provided by the International Centre, and during it, verifiers familiarize themselves with the study, the test materials and the translation procedures, and get detailed instructions and practice on how to review the materials and what to do with deviations from the source versions.

PROBLEMS AND ISSUES IN TRANSLATING INTERNATIONAL ACHIEVEMENT TESTS

In research, several factors have been found to cause problems and issues when translating international achievement tests. Among the most common and intricate of these, are the following: the unique and demanding purpose of the translation task; the partly contradictory and controversial task specifications and translation instructions; the choice of the number of the target versions to be produced; the choice of the number of source versions to be used; inadequate revision and verification; deficiencies in translator competences; and a lack of time. These are discussed in more depth in the following. For each factor, the paper first lays out what Translation Studies has to say about it and its impact on translation, then discusses what problems or issues there have been in international achievement studies with respect to the factor, and, finally, suggests what can be done to alleviate the problems.

UNIQUE AND DEMANDING PURPOSE – EQUIVALENCE IN DIFFICULTY

Every translation has a purpose, goal or function. It is, moreover, this purpose which governs the entire translation work, determining how a text is to be translated (Reiss & Vermeer, 1984; Vermeer, 1989). For example, when translating a short story, the purpose is usually to produce a literary text, which, in turn, requires that special emphasis be put on aesthetic factors.

In addition to the more specific purposes, most translations also seek to be dynamically equivalent to their source texts, dynamic equivalence thus typically being the overriding purpose.

Dynamic equivalence, or idiomatic, translations are translations that aim at having a similar *effect* on the target text reader as the source text has on the source text reader (Nida, 1964, p. 159). They thus strive to read like normal and natural target language texts, intended for target readers, not like translations. Therefore, they often cannot be translated literally, but need to be rendered more freely and idiomatically. For example, a dynamic equivalence translation for the French *Défence de fumer* – literally, 'Interdiction to smoke' – is *No smoking*, which deviates syntactically and lexically from the source text rendering, but has the same pragmatic, intended meaning. Dynamic equivalence translations frequently also need improvement (e.g., linguistic refinement) and explicitation (e.g., clarification, additions) – so much so that the two have been termed universals of translation (Laviosa-Braithwaite, 1998; Séguinot, 1989).

A much rarer purpose is to produce a formal equivalence translation. *Formal equivalence*, or literal, translations are translations that strive to preserve also the formal and structural elements of the source text (Nida, 1964, p. 159). For example, the rendering *Interdiction to smoke* literally reproduces the meaning and form of *Défence de fumer*. However, the problem with formal equivalence is that it often leads to interference, undue influence by the source language (Toury, 1995, p. 275), or even translationese, artificial target language, and translations that are awkward and difficult to understand (e.g., Nida & Taber, 1969). Unduly literal translation, or interference, is also extremely common when translating. Therefore, it, too, is called a universal of translation (Toury, 1995, p. 275).

When pursuing formal equivalence, translation is much easier than when pursuing dynamic equivalence. This is because when translating, it is normally the literal and formal equivalence translation that comes to mind first (Englund Dimitrova, 2005). However, when pursuing dynamic equivalence, the translator first has to try to find out what the effect of the

source text might be on the source reader – which, of course, is at best an educated guess – and then find a formulation which might have a similar effect on the target reader (see e.g., Munday, 2001, p. 42). Pursuing dynamic equivalence often also means that the translator has to distance himself or herself from the literal translation and search for more idiomatic renderings. All this requires extra cognitive effort and time. (Englund Dimitrova, 2005.)

In international achievement studies, each text and item has its own individual translation purpose (varying greatly in e.g., reading tests). In addition, however, all translated versions also have a common, superior purpose – equivalence in difficulty. Equivalence in difficulty, as will be remembered, presupposes that the mental effort required of testees remain the same across languages. Equivalence in difficulty thus presumes equivalence in effect and falls under, or is a subtype of, dynamic equivalence. However, contrary to what is normally the case with dynamic equivalence, equivalence in difficulty also requires comparability in *difficulty*. When pursuing equivalence in difficulty, translators thus have an extra, important requirement that they need to take into consideration. This, in turn, seems to increase the difficulty of the translation task and make it hard for translators to reach the translation goal. Author (2012b) conducted discussions with the translators rendering the PISA 2009 materials into Finnish, asking them about the difficulties they faced while translating. As one of the complications, the translators mentioned the requirement for equivalence in difficulty. They also gave two reasons for this: First, when translating, the translator has no way of knowing how difficult the source and target versions truly are (cf. the requirement for equivalent effect). Second, equivalence in difficulty is typically not a purpose in other types of translation, and therefore translators are usually not trained for and used to pursuing it. The purpose of translations used in international achievement studies is thus especially demanding to pursue.

Measures are therefore needed to make the translation purpose more tangible and easier to grasp. This is best done by providing translators with clear instructions and training on the purpose and on how to pursue it. However, as shown by the following sections, providing such instructions and training has not always been easy, and therefore translators may not always have had a clear picture of what is involved in equivalence in difficulty (see also Author, 2012b).

TASK SPECIFICATIONS AND TRANSLATION INSTRUCTIONS

Since translations may have different purposes, translators need to be clearly informed about the purpose and specifics of the translation task at hand so that they know how to translate. To this end, translators are usually provided with written task specifications. The specifications contain, or should contain, information about the purpose of the translation and the conditions under which the purpose should be achieved, such as the context, use and audience of the translation, and even linguistic translation instructions (Nord, 1991, 2006). When made well, the specifications help the translator to solve translation problems (e.g., how literally or freely to translate) and to produce translations that are of high quality and fulfill their purpose (Sharkas, 2009, p. 47). However, if the specifications are missing or if they are not clear or explicit enough, the translator is left uncertain as to how to translate, which, in turn, easily results in inadequate translations (Nord, 2006).

Today, all achievement studies provide translators with written task specifications (cf. Harkness, 2003, p. 45). Typically these are supplied by means of specific translation and adaptation guidelines. Even though very little research exists on these guidelines, it seems that they have not always managed to provide a clear and unequivocal picture of the unique and demanding purpose of the translations used in achievement studies.

In OECD studies, the guidelines have contained a great number of linguistic translation instructions so as to help translators to judge the difficulty of the source and target versions. However, discussions with the Finnish PISA translators (Author, 2012b) suggest that because of the numerous detailed instructions, the guidelines have ended up being slightly contradictory and misleading: On one hand, the guidelines have defined the purpose as equivalence in effect and difficulty and advised translators to strive for as natural and authentic target versions as possible, the emphasis thus being on dynamic equivalence. On the other hand, however, the guidelines have mainly consisted of detailed instructions on how to remain lexically and syntactically as close to the source version as possible, the focus in practice being on formal equivalence.

This discrepancy, in turn, as commented by the Finnish translators (Author, 2012b), has made it extremely difficult for them to decide how freely or literally to translate. However, mainly the strong emphasis on formal and micro-level factors seems to have made them feel that they have had to stay close to the source text and translate literally. They have felt that they have had to accept translations which they have known were not the best and most idiomatic choices and which they would not have accepted in other contexts. In more recent OECD guidelines, slightly more attention has been paid to natural target language, and this has helped the translators to distance themselves somewhat from the source version. However, as pointed out by the translators, since the guidelines still consist largely of detailed instructions emphasizing faithfulness to the source version – the number of the instructions even having increased during the past years – these still tend to drown out the need to produce natural target versions.

Another problem with the specific linguistic instructions, brought up by the Finnish PISA translators, is that they do not apply equally well to all languages (Author, 2012b). Languages differ enormously, and therefore no universally valid linguistic instructions can be given. In

PISA, for example, the instructions have best applied to Indo-European languages. However, when translating into, say, Finno-Ugric (e.g., Finnish), Afro-Asiatic (e.g., Arabic), Sino-Tibetan (e.g., Chinese) and Altaic (e.g., Korean) languages, they have often not been of much help or may sometimes even have tempted into overly literal translations (Author, 2012b; Grisay, de Jong, Gebhardt, Berezner & Halleux-Monseur, 2007, p. 28).

In IEA studies, only a very few specific instructions have been given in the translation guidelines, and somewhat more emphasis has been put on idiomatic target language (and dynamic equivalence). However, it seems that also this practice has had its problems: When translating TIMSS 2011 materials, Finnish translators felt forced to ask for the opportunity to make use of the OECD instructions. This suggests that the IEA guidelines may have been even too general, not containing enough information to be of practical help.

Sometimes, however, even though much less often, problems seem to have been caused by the requirement for the target versions to be natural and authentic. Typically, the result has been unduly free, explicit, transparent and/or straightforward translations. For instance, Author (2007) conducted a text analytic study of Finnish PISA 2000 materials and found that in an attempt to make the Finnish versions natural, Finnish translators sometimes improved and explicated the versions by adding grammatical words or by using more concrete expressions (cf. Hambleton, 2001, p. 166). Usually, the reason appears to have been that the translators did not realize how much even seemingly small linguistic changes may affect item difficulty. In questionnaire translation too, strong emphasis on authenticity has led to overly free translations (Kleiner, Pan & Bouic, 2009).

Much more research is, of course, needed on the translation guidelines. However, on the basis of the above, it seems that the main problem with the guidelines has been a lack of balance

between faithfulness to the source text (formal equivalence) and idiomatic target language (dynamic equivalence). To find such a balance, the following suggestions might be given: Considering the uniqueness and difficulty of translating international achievement tests, it seems that linguistic translation instructions are needed to make it clear to translators what is involved in the task and to help them to judge item and text difficulty. However, to ensure that the linguistic instructions do not make the translators translate overly literally – which, moreover, is a universal tendency among them – it is necessary to emphasize that one of the prerequisites for equivalence in difficulty is that the target versions are in natural target language (see e.g., the translation guidelines for The European Social Survey; ESS, 2010, pp. 23-5, 28; cf. Jeanrie & Bertrand, 1999, p. 279); explicit warnings also are needed against unduly literal translations (ESS, 2010, pp. 37-8). Reminders are likewise needed that because of differences between languages, the instructions do not always apply and that, therefore, slavishly following them easily leads to awkward translations. Separate, customized instructions may need to be prepared for languages that are very far from English and French, as has already been done in PISA for translation into Arabic and Chinese. To make sure that the request for naturalness does not lead to unduly free and straightforward translations – and to help translators to resist their universal tendency to improve and explicate texts – reminders are needed that test translation differs from other types of translation, in that in it improvement and explicitation are not allowed (see also Hambleton, 2001, p. 166; Harkness, 2003, p. 46); translators should also be reminded that even apparently insignificant linguistic modifications may sometimes improve texts and bring about changes in difficulty (Author, 2007).

SINGLE- OR DOUBLE-TRANSLATION?

When more than one translation is made from the same source version, the reason is typically either to produce an updated translation of an already translated text (e.g., the Bible) or to assess the language skills of testees. In both cases, the translations may end up being quite different from each other in terms of vocabulary and style, for example. This has both its cons and pros. At their best, the different versions provide the reader with different perspectives into the source text, thereby helping him or her to get a deeper understanding of it (e.g., in Bible translation). They also present different ways of solving translation problems (e.g., in a test situation). However, a difficulty easily arises, if the different versions need to be merged into a single text: Whenever passages from different writers are combined, consistency and coherence easily suffer.

In international achievement studies, the procedures differ in that, for example, in OECD studies, two independent translations are produced and then reconciled into one (this is also the procedure recommended by e.g., Harkness, Villar and Edwards, 2010 and used e.g., in ESS), whereas in IEA studies, only one translation is made which is then reworked and revised. Not much research exists on the effectiveness of these procedures. Besides, the results of this research are ambiguous.

For example, Grisay (2002, 2003) compared the effectiveness of the different translation methods used in the PISA 2000 field trial: double-translation from two languages; double-translation from one language with cross-checks against the other language; double-translation from one language without cross-checks; single translation; and mixed methods, typically with at least some of the materials double-translated from two languages. When calculating the

percentages of flawed items (items with poor psychometric characteristics) in all national versions, she found no great differences between those that had been single-translated and those that had been double-translated from one language. These were the two least effective translation methods, single-translation, however, being slightly more effective than double-translation from one language.

However, the above findings are confounded by the fact that at least Finland was misclassified (apparently because of faulty reporting and misunderstanding) in the study (Grisay, 2002, p. 67; 2003, p. 236): Finland was classified as a country using double translation from English with cross-checks against French. In reality, however, Finnish materials were only single-translated and with very few cross-checks. Yet, the quality of the materials was very high, as shown by both the low number of flawed items in them and reports by verifiers.

Finland used the single-translation procedure also in PISA 2003, with, however, the exception that the cross-checks against the French version were systematic and extensive. This time the quality of the Finnish translations was even better than in 2000 (OECD, 2005). These findings run counter to those of Grisay, suggesting that single-translation can also be very effective, especially when combined with extensive cross-checks against the other source version.

Taken together, the above findings seem to suggest that single-translation may be at least as effective, if not even more effective than double-translation, if only one source version is used; when two source versions are used, double- and single-translation are both very effective. The findings (especially those on Finland) also suggest that there may be other factors that may be even more decisive for the efficiency of the translation method, such as the amount of

revising and finalizing done on the target versions (which seems to be a major difference between the Finnish and the recommended OECD procedure). Much more research, of course, is needed not only on single- and double-translation (and the pros and cons of each translation method), but also on all other aspects of the translation procedures and on the methods actually used in the participating countries to find the most effective way of translating international achievement tests.

ONE OR TWO SOURCE VERSIONS?

Reference materials (e.g., dictionaries, encyclopedias, parallel texts) are typically used to facilitate translating and to make better-quality translations (Nord, 1991). Making use of more than one source version serves the same purpose. The different words, structures and nuances used in the various versions provide extra clues that help the translator to better understand the source material and to convey its meaning more accurately. They also help the translator to see that typically there is not just one but several acceptable ways for expressing one and the same idea. This, in turn, can encourage the translator not to translate literally but idiomatically.

However, using several source versions can also create problems. Firstly, it means that merging – or reconciliation, or putting together of ideas or extracts from various sources – of some kind is needed (either right away, with all source versions directly reconciled and translated into one single target version; or after each source version has first been translated as an independent target version). No matter how the merging is done, it places heavy demands on the consistency and coherence of the text. Establishing the coherence of the text, in turn, requires

a considerable amount of time. Secondly, if the source versions are in different languages, extra requirements are set for the reconciler, who needs to be proficient in all the languages so as to be able to make full use of the versions. Thirdly, the different source versions may increase the risk of non-equivalent translations. This risk, of course, is the greater, the more different, or non-equivalent, the versions are. Since non-equivalences are typically larger between languages than within them – full equivalence hardly ever existing between languages (e.g., Chesterman, 1997; Pym, 1995) – the risk is especially great, when the source versions are in different languages.

Similar controversial results have also been obtained in international achievement studies. Research on PISA studies, where two parallel source versions have been used, shows that this procedure can both help to produce higher-quality and more equivalent translations and increase translation difficulty and the risk of non-equivalence (cf. Grisay, 2003). For example, the Finnish PISA 2009 translators (Author, 2012b) commented that the use of two versions is beneficial, because it provides two alternatives from which to choose, which, in turn, helps to see how much freedom is allowed in the translations and to avoid overly literal translations and to produce more natural translations (see also OECD, 2009a). Also, when a word in the source version has several meanings and could be translated in several ways (as is often the case), the other source version frequently helps to find the correct translation.

That the use of two source versions helps to produce more equivalent translations seems to be supported by several studies. For instance, in her text-analytic study of Finnish PISA 2000 translations, which were made almost exclusively on the basis of only the English versions, Author (2007) discovered that the Finnish versions sometimes contained mistranslations and overly literal and clumsy renderings and concluded that these could largely have been avoided, if, when translating, more extensive use had been made of the French versions. Also, verifiers in

the PISA 2000 field trial reported that those translations that were rendered from both source versions contained fewer mistranslations, fewer unduly literal translations and fewer flawed items than those that had been translated from only one source version (Grisay, 2002). Moreover, when Grisay (2002, 2003) calculated the proportion of flawed items in these versions, she found that those translations that had been rendered by using both source versions – especially by double translation from the two languages but also by using one of the source versions for double translation and the other for extensive cross-checks – contained significantly fewer flawed items than those that had been rendered from only one source version. (In this context, however, it has to be remembered that at least Finland was misclassified in Grisay's study: The Finnish materials were not double-translated from English with extensive cross-checks against French, but single-translated with very marginal cross-checks. Yet, the Finnish versions were of high quality.) In PISA 2003, too, (in a similar comparison) double translation from two languages yielded the best translations (OECD, 2005).

Another significant advantage of using the two source versions is that the translation of the other source version serves as a translation trial: It helps to spot and correct errors and ambiguities in the source versions and to anticipate translation problems in the target versions, which, in turn, makes the translation of the national versions and attaining equivalence easier (Grisay, 2002, p. 60; cf. Harkness, Villar & Edwards, 2010, p. 131).

However, the use of the two source versions also poses some problems. For example, it complicates the reconciliation and may have a negative effect on the language of the target versions. When translating materials for the International Adult Literacy Survey (IALS; conducted by STATCAN in cooperation with ETS) – from which PISA translation procedures have largely been inherited – Finland made two independent translations (from English and

French) and reconciled them into one. However, the translators (P. Linnakylä, personal communication, November 14, 2008) found the procedure complicated, remarking that when two translations are made from two different-language source versions, the translations often end up so different that merging them into one target version is extremely difficult. Therefore, extra effort and time is required of the reconciler to make the reconciled versions into a consistent, coherent, flawless and idiomatic whole. At the same time, however, less time is left for him or her to do so: Comparing the target versions to two different-language source versions, of course, takes more time and effort than comparing them to only one version (as also remarked by the Finnish PISA 2009 translators; Author, 2012b). Also, when having to operate with two source versions, the risk of interference and overly literal translations is even greater than when there is only one source version. All this suggests that for translation from two source versions to be truly effective and able to yield high-quality and natural national versions, sufficient time needs to be allotted to the reconciliation phase so that the reconciler, in addition to all his or her other responsibilities, also has time to refine and finalize the national versions.

Another problem when translating from two source languages is the difficulty of finding – at reasonable cost – translators, reconcilers and verifiers who would be competent in both languages, a problem suggested by the fact that the vast majority of PISA countries (e.g., 39 out of 45 in 2000, and 40 out of 55 in 2003; Grisay, 2002; OECD, 2005) have not followed the recommended translation procedure but have translated from only one language (although there may, of course, also be other reasons for the choices of the countries; see also Author, 2012b; B. Halleux-Monseur, personal communication, January 24, 2008). If all translators making the first national versions are competent in only one and the same source language, the two source versions, of course, cannot be used. If the reconciler or verifier is competent in only one source

language, the two versions can be used, but this complicates and reduces the effectiveness of the procedure. For example, if the reconciler only knows one source language, s/he has no way of knowing to what extent the other target version matches with its source version. This may be a problem, especially if the two target versions are very different, and may even result in the reconciler judging some translations as too free and therefore rejecting them and choosing more literal translations instead. The same may be true of the verifier, if s/he compares the national version to only one source version. In this way one of the obvious advantages of using two different-language source versions – the opportunity to see alternative ways of expressing ideas and to get more leeway in translation – may be forfeited during the reconciliation or verification phase. Evidently, therefore, for translation from two source versions to be beneficial, it is important that reconcilers and verifiers are competent in both source languages.

However, the most serious problem when using two different-language source versions is that the two versions do not always seem to have been fully equivalent to each other. This is a major complication, because non-equivalent source versions almost inevitably lead to non-equivalent translations. For example, in linguistic comparisons between English and French PISA 2000 materials (e.g., Author, 2012a; Grisay, 2004, Karg, 2005), differences have been found in the choice of key vocabulary, in the precision of terms, and in how literal the link between the stimulus texts and items has been (similar results have also been obtained for IALS; see e.g., Blum, Goldstein & Guérin-Pace, 2001). Also, Finnish PISA 2009 translators reported that sometimes when the Finnish version had been translated more in line with and was thus equivalent to the French source version, the translation was “corrected” during the verification so as to make it more equivalent to the English source version (Author, 2012b). Moreover, Grisay (2002, 2003) compared the length (number of words in the stimulus texts) and linguistic

complexity (measured by means of readability formulas) of the English and French stimulus texts used in the PISA 2000 field trial and found that the French texts tended to be significantly longer, which, in turn, (modestly) increased the difficulty of some French items. She also compared the psychometric quality of the national versions adapted from the English and those adapted from the French source versions and found that even though there were no significant differences between the groups in the number of flawed items, a few individual items functioned differentially across the groups. Similar results were also obtained in PISA 2003 (OECD, 2005). In addition, when Grisay and Monseur (2007) compared the proportion of DIF (differential item functioning) items within (e.g., all various English versions) and between language groups (e.g., all English vs. French versions) in the PISA 2000 main study, they found that it was considerably higher between groups, concluding that whenever a test is translated into another language – from English into French, for example – part of equivalence is always lost.

If, then, two different-language source versions are used, it is imperative to ensure that the versions are equivalent to each other. Of course, full equivalence cannot be expected, because, as widely acknowledged in both Translation Studies (e.g., Chesterman, 1997; Pym, 1995) and test translation (Grisay, Gonzalez & Monseur, 2009), absolute equivalence, or exact uniformity, does not exist. However, a sufficient level of equivalence needs to be ensured, and measures need to be taken to guarantee that this level is attained. This should involve not only a careful simultaneous production of the two versions by a team of translators and content and testing specialists (cf. Harkness et al., 2010, p. 47), as largely done today in PISA, but also thorough quantitative and judgmental comparisons between the English and French versions and preferably even pretesting – and a considerable amount of time to do all this. Then, on the basis of the comparisons and testing, a decision can be made on how equivalent the two versions are

and whether this level is sufficient to justify the use of the two versions. If a sufficient level cannot be ensured, it may be safer to use only one source version. Nevertheless, much more research is needed on the comparability of the two different-language source versions and the practice of using two versions.

REVISION AND VERIFICATION

Revision is an important part of the translation process and aims at correcting and improving the translation (Mossop, 2007, p. 17). Although some revising typically takes place while the translator is still drafting the target text, a more thorough revision is usually carried out as a final step in the translation process (see e.g., Englund Dimitrova, 2005). Properly revising a translation is a demanding and time-consuming task. It requires checking not only the semantic accuracy and grammaticality of the translation but also, for example, its idiomaticity, fluency, style and textuality. Usually, all this cannot be done successfully at the same time. This is because each of the aspects calls for a different type of reading and because focusing on one aspect generally blinds one to the others. (Larson, 1998; Mossop, 2007.) For example, when comparing the translation and the source text, the translator has to stay close to the source text, and this makes it impossible for him or her to pay full attention to the target text and its idiomaticity (Englund Dimitrova, 2005, pp. 32, 233; Mossop, 2007, p. 147).

Therefore, several separate revisions are typically needed to ensure a high quality translation, each of the revisions focusing on a different aspect (e.g., separate checks for comparing the translation to the source text and for assessing the idiomaticity of the translation). Other factors that help to improve revision include using several revisers and making the revision

on paper. Both make it easier for the reviser to spot errors and unidiomaticities. Making the revision on paper (instead of making it on screen) also makes it easier for the reviser to examine the text as a whole and to ensure its coherence. Conversely, if the revision is made in a hurry, with insufficient separate checks, by only one person, and on screen, the translation normally ends up containing more unidiomaticities, errors and text-level problems. (Mossop, 2007; see also Englund Dimitrova, 2005.)

In international achievement studies, making a revision is an even weightier and more demanding task. This is because in these studies, not only the overall linguistic quality of the translated versions (e.g., their grammatical, semantic, stylistic and textual accuracy and their fluency and idiomaticity), but also their equivalence in difficulty needs to be assessed. A separate phase, international verification, has therefore been designed, and it is implemented as the final translation stage in these studies. However, otherwise the practices of revision differ between the studies. For example, in IEA studies, there is at the end of the national translation process a stage (Step 2) during which the reviewer only concentrates on reviewing and revising the national versions. In OECD studies, no comparable stage exists. Instead, the reconciler first has to merge together the two target versions, which may differ considerably from each other. Then, in addition to this, s/he has to review and revise the resulting version.

Even though there is to date no research proper on the effects of revision on translated tests, there are some findings which suggest, not surprisingly, that the above facts about revision also apply in test translation and that not paying sufficient attention to revision also has a negative effect on the quality of translated tests (e.g., Solano-Flores, Backhoff & Contreras-Niño, 2009). For example, Solano-Flores, Contreras-Niño and Backhoff-Escudero (2006) examined the quality of the Mexican Spanish-language version of the TIMSS 1995 test and

found that it contained a significant number of translation errors, concluding that these were largely due to the version not having been properly revised nationally (in TIMSS 1995, the translation procedure was double-translation from English followed by reconciliation). By contrast, in Italy, a lot of time and effort was invested in revising and finalizing the PISA 2003 materials, and this was felt to be an important factor explaining why (according to verification reports) the materials only contained some minor problems (Siniscalco, 2006, p. 206). Also, when in Finland two target versions were made in IALS and reconciled into one, the translators found the procedure problematic, one of the main reasons being that it did not leave sufficient opportunity to work on, revise and finalize the Finnish versions (P. Linnakylä, personal communication, November 14, 2008). Finland therefore decided to follow a slightly different procedure in PISA, a procedure where only one translation is made which is then reworked and revised by two successive national translators (plus the international verifier). And yet, the quality of the Finnish PISA tests has usually been judged to be very high, one of the best: In psychometric comparisons, they have been found to contain very few flawed items (Grisay, 2002; OECD, 2005); and according to verification reports, they have often not needed any proofreading, because they have been linguistically and grammatically of so high quality. More research, of course, is needed to disentangle the reasons for the seemingly high quality of Finnish PISA tests. However, it appears that at least part of the credit goes to the fact that in Finland much more weight has been put on revision than in the recommended PISA procedure.

It thus seems important that in all international achievement studies sufficient attention be paid to revision (see also Hambleton, 2002, p. 67; Hambleton & Patsula, 1999; Siniscalco, 2006; Solano-Flores, Contreras-Niño & Backhoff-Escudero, 2006). In practice this might mean, for example, including in the translation guidelines a checklist of what needs to be revised (see e.g.,

Hambleton & Zenisky, 2011; Solano-Flores, Backhoff & Contreras-Niño; 2009, p. 82; cf. Jeanrie & Betrand, 1999, pp. 280-1) and reminding revisers of the importance of making several revision rounds. In OECD studies, it might also involve allotting more time to the reconciliation phase so that the reconciler really has time to do all the needed revision rounds and finalize the national versions. However, since reconcilers first need to examine the two target and the two source versions and merge the two target versions into one, it is more than likely that these different versions will continue to have an impact (interference) on them also while revising and blind them to errors and overly literal renderings. Therefore, a better option would be to add a separate phase dedicated to revision also to the OECD translation procedure (see also Author, 2012b). This would make it possible for at least one person to properly revise and finalize the national versions before they are verified. Making at least part of the revision on paper might also help to detect and correct more errors and idiomaticities, not only in OECD but also in IEA studies.

The above suggestions seem all the more important, since there appear to have been deficiencies also in the verification. For example, analyses of Finnish PISA 2000 translations (Author, 2007) and discussions with Finnish PISA 2009 translators (Author, 2012b) suggest that the verifications in PISA 2000 and 2009 may have lacked thoroughness and that, therefore, errors and non-equivalences were not always found. Also, in the Mexican TIMSS 1995 version, errors and non-equivalences were found that had not been detected during the verification (Solano-Flores, Contreras-Niño & Backhoff-Escudero, 2006).

No research exists on factors affecting the efficiency of verification. However, given the numerous aspects that have to be verified (the grammatical, semantic, stylistic and textual accuracy of the translations, their fluency and idiomaticity, literal and synonymous matches, etc.), it seems reasonable to assume that the lack of thoroughness in the verification and its

inability to detect errors and problems have been due to there not having been sufficient revision rounds. This, in turn, may have been because verifiers may have lacked certain competences, because they have not been fully aware of the need for several revisions, and/or because they have not had sufficient time to make the revisions. Part of the explanation may also have lain in the verification having been done on screen. (See also Author, 2012b.)

Given the decisive role the verification plays in the translation process, it is extremely important that it be carried out thoroughly and in sufficient depth. To ensure that this is the case, it is necessary to hire only qualified verifiers, to remind them (in the written instructions and training) of the need to make several revisions and of the advantages of making it at least partly on paper, and to allot them sufficient time to do all this.

TRANSLATORS

The translator is the key actor in the translation process. It is the translator who actually produces the translation, making the final decisions on how to translate. (Vermeer, 1989.) Therefore, the quality of the translation is, in the end, dependent on the translator.

The qualifications required of translators vary somewhat according to the purpose of the translation. However, the minimal requirement for any translator to be able to translate is good linguistic skills: To be able to produce a high-quality and natural target language text, the translator needs mastery of the target language; and to fully understand the source text, s/he needs mastery of the source language (see e.g., Shreve, 1997, p. 122). In addition to the linguistic skills, however, the translator also needs extra-linguistic knowledge: knowledge of the subject matter, context and cultures concerned (PACTE, 2005, p. 610). Subject matter knowledge is

extremely important (Ericsson & Kintsch, 1995; Kim, 2006), especially when translating factual texts. In literary translation, again, literary skills are needed (Lefevere, 1992). Finally, to be able to choose the (right and) best way to translate each text and to solve translation problems, the translator needs translational and strategic knowledge (PACTE, 2005, p. 610): S/he needs to know, for example, how the purpose and readers of the translation affect the way a text should be translated. Translational knowledge, in particular, cannot typically be gained without explicit translator training.

Translators with deficient translator competences typically produce lower-quality translations than competent translators: They easily focus too much on formal and word-for-word correspondence, translating overly literally (cf. Dansk & Griffin, 1997, pp. 171-2; Jensen, 2000, p. 166; Jääskeläinen, 2010, p. 221; Kim, 2006), which, in turn, often leads to interference, cumbersome translations and errors. In contrast, competent translators concentrate on translating ideas (e.g., Rydning & Lachaud, 2010, p. 107). In doing this, they take into account the purpose (Englund Dimitrova, 2005, pp. 14-5) and readers of the translation (e.g., Jääskeläinen, 2010, p. 221), aiming at comprehensible, idiomatic and readable target texts (Jensen & Jakobsen, 2000, p. 114). They also pay more attention to stylistic factors (see e.g., Englund Dimitrova, 2005, pp. 14-5; Lörscher, 2005, p. 606). When revising, competent translators also tend to detect more errors and problems than less competent translators (cf. Hayes, Flower, Schriver, Stratman & Carey, 1987, p. 233).

The above competences are also required when translating international achievement tests: linguistic skills, extra-linguistic knowledge – e.g., knowledge of school, school subjects, students, cognitive processes, and testing – and translational knowledge. However, it seems that there have been deficiencies in these competences (e.g., Hambleton, 2002, 2005). For example,

when Karg (2005) examined German PISA 2000 and 2003 field trial and main study materials, comparing them to the source versions and some national versions, she found several syntactic errors in them, seemingly pointing to deficiencies in the translators' mastery of the target language. Also, in Finland, when translating PISA 2000 materials, even the verifier was reported not to have had as good a command of the source language (English) as required (Author, 2012b). However, the most serious problem especially in PISA studies seems to have been that countries have not always managed to find translators who would have been able to translate, and especially to reconcile and verify, materials from the two source languages (B. Halleux-Monseur, personal communication, January 24, 2008; Grisay, 2002, p. 62), and that therefore both the translation process and the quality of the translations have suffered (Author, 2012b).

Deficiencies have also been found in subject matter knowledge and literary translation skills. Deficient literary translation skills have mainly caused problems in reading tests (Author, 2011, a). However, somewhat surprisingly, subject matter knowledge, too, seems to have been lacking not only in, say, mathematics and science tests (e.g., TIMSS; Hambleton, 2005, p. 25) but also in reading tests, at times leading to erroneous and misleading translations (Author, 2012b; Karg, 2005). That both the above competences have posed problems in reading tests appears to be due to the wide variety of text types (e.g., literary, expository) and topics (e.g., technology, biology, medicine, law) covered in these tests, because of which it is impossible for any one translator – translators typically specializing in only one field – to master all of them. Finnish PISA 2009 reading literacy translators, for example, felt that they were not as experienced in translating biology and literary texts as, say, educational texts (Author, 2012b).

Translators have also lacked familiarity with cognitive tests and test translation. This was the case, for example, in TIMSS 1995 (Hambleton, 2005, p. 25). Also, in Finland, two translators

were used to translate the PISA 2009 materials into Finnish (the first drafts). One of them was a translator with ample experience in test translation but without academic translator training. The other was an academically trained translator, with, however, no training and experience in testing and test translation. S/he had read the translation guidelines, and s/he had even been offered the opportunity to have training in test translation. However, pleading his/her academic training, s/he declined. When the drafts made by these two translators were analyzed, those of the translator with no experience in test translation contained much more violations of principles of test translation (e.g., literal matches, when the source version used synonyms) (Author, 2012b).

There seem to have been deficiencies in translational and strategic knowledge, too. For example, Karg (2005) mentions several points in the German PISA 2000 and 2003 translations where translators appear to have failed to take sufficiently into consideration, for example, stylistic factors and German readers and to make the German versions dynamically equivalent to the source versions. Also, Author (2007) found that the Finnish PISA 2000 translations contained excessively literal renderings and concluded that these were largely due to most of the Finnish translators, including the verifier, not having had training in translation theory and strategies. Similarly, in PISA 2009, the Finnish translations made by the translator who was not acquainted with translation theory were clearly more literal and cumbersome than those of the academically trained translator (Author, 2012b).

Without further study, it is, of course, impossible to say what the reasons for the above deficiencies are. However, it seems reasonable to assume that especially in OECD studies, the deficiencies may have been at least partly due to deficiencies in the requirements (as outlined in the translation guidelines): Reconcilers have not been required to be proficient in both source languages; reading literacy translators have not been expected to have knowledge of the subject

matter or experience of literary translation; and translators have not been required to have formal credentials. Another reason may have been a lack of training in test translation: In IEA studies, training has been lacking altogether (except for verifiers), and in OECD studies, translators have not always taken part in the training. Finally, in both studies, the deficiencies may have been due to the high costs of hiring competent translators.

Considering the highly responsible role of translators in international achievement studies and the deficiencies there seem to have been in their competences, it appears that more care is needed when hiring the translators (as strongly emphasized also in e.g., Hambleton, 2001, p. 166; 2005, pp. 24-5; Harkness, Villar & Edwards, 2010). This may even necessitate testing the prospective translators (for more suggestions for such tests, see e.g., ESS, 2010, pp. 12-5). For example, it is important to ensure that the translators have a good command of the target and source languages. If the test is translated from two source languages, the reconciler and verifier need to be proficient in both the languages. At least some translators in each team need knowledge of the subjects assessed, not only, for example, in mathematics and science tests, but also in reading tests. Translators of literary texts need experience of literary translation. And all translators need training in what is involved in cognitive tests (see also ESS, 2010; Hambleton, 1994; 2002, p. 66; Harkness, Villar & Edwards, 2010) and a good knowledge of the principles and strategies of translation. However, since translating international achievement tests requires so many widely differing and specific skills, it may be expected that it will not always be possible to find translators who would fill all the requirements. Deficiencies may be anticipated in subject matter knowledge and familiarity with testing, in particular. To rise to this challenge, it is good to involve also subject matter and testing specialists more closely in the translation teams and to have regular discussions between all team members (as also suggested in e.g., ESS, 2010;

Harkness, Villar & Edwards; see also Hambleton, 2001, p. 166; 2005, pp. 24-5). The requirements for the translators may also need to be revised.

TIME

Quality takes time (Mossop, 2007, p. 114). This is also true of translation, which has been described as a complicated and time-consuming cognitive problem-solving process (Jensen, 2000, p. 40). However, when under time pressure or in a hurry, the translator lacks cognitive resources, which, in turn, easily leads to errors (cf. Zakay, 1993). Also, when under time pressure, the translator has not sufficient time for problem-solving (e.g., Jensen, 2000): S/he does not have time to consult and interact with others (Mackenzie, 1998); and s/he does not have time to be creative (see e.g., Fontanet, 2005, p. 444) and to elaborate on the text (Jääskeläinen, 1996). Instead, s/he has to be satisfied with the solutions that first come to mind (Jensen, 2000). These, however, are typically the most literal translations, which follow closely the formulation of the source text (see also Chesterman, 1997; Krings, 1986, p. 507), sometimes even to the point of interference (cf. Neubert, 1997, p. 20). If, moreover, the translator is not qualified or familiar with the subject matter, the negative effects of time pressure are even more serious (Jensen & Jakobsen, 2000, p. 114; Kim, 2006).

Time pressure and the lack of time have repeatedly been reported as causing problems also when translating achievement tests. This was the case, for example, in TIMSS 1995 (Hambleton & Berberoglu, 1997) and PISA 2000 (Hambleton, 2002, 2005). One of the Finnish PISA 2009 translators even commented that “time is always a problem in this context” (Author, 2012b). Because of time pressure, translators have not had sufficient time to be creative and to

discuss with others, and errors and unduly literal translations have ensued (Author, 2012a). However, the problems do not appear to have affected all translators to a similar degree. Rather, at least in OECD studies and in Finland, the greatest sufferers seem to have been reconcilers (Author, 2012b).

Given the serious negative effects time pressure and the lack of time have on the quality of translated tests, it is important to ensure that translators in international achievement studies have sufficient time to do their job properly – that they have time to discuss, elaborate and be creative and meticulous. One way to ensure this is to see to it that there are enough translators for each step. However, since at least in OECD studies reconcilers, in particular, seem to have the most time pressure and since their requirements, in particular, are so high, this option may not be easy to realize. Other options, then, would include decreasing the duties of the reconcilers by making, for example, proofreading and revision a phase of its own and a responsibility of another translator, and allotting more time to translation in the assessment schedule. These options, however, would necessitate changes in the assessment procedures and probably even in the assessment cycles.

CONCLUSION

This paper reviewed research and findings on problems and issues encountered when translating international achievement tests and seeking to ensure their equivalence. The problems concentrated on the following: the unique and demanding purpose of the translation task, the partly contradictory and controversial task specifications and translation guidelines, the indecision as to whether to make one or two target versions and whether to use one or two source

versions, the two source versions not having been fully equivalent to each other, inadequate revision and verification, translators not having been fully competent, and a lack of time.

The paper also suggested solutions to the problems, some of them destined for test developers, some for practitioners, and some for both. These may be summarized as follows:

- Ensuring that the translation guidelines provide a right, clear and unequivocal picture of the purpose of the translation task and that there is in the guidelines a balance between the need for idiomatic target language and linguistic translation instructions. Partly customizing the instructions.
- Ensuring the equivalence of the two source versions, and using two source versions only if they are equivalent to each other.
- Putting more emphasis on the revision and refining of the national versions, and ensuring that the verification is sufficiently thorough: Providing a checklist of what needs to be revised, and reminding of the need for several revision rounds. Encouraging revision on paper (partly, at least). If two parallel versions are used, making the revising and finalizing of the translations a phase of its own.
- Using only qualified translators and revisers to translate and revise the tests, providing them with training in cognitive tests and test translation, and partly revising the translator requirements. Involving subject matter and testing experts in the translation teams.
- Ensuring that the translators, revisers and verifiers have sufficient time to do their jobs. Allotting sufficient time to translation and verification in the assessment schedule.

However, the most important lesson from this paper is that research in the field is extremely limited and that more research is therefore badly needed. The fact that the translation procedures in international achievement studies have differed both between the organizations and between the participating countries as well as the partly contradictory (and sometimes even faulty) findings and confusing experiences that have been gained thus far suggest that further studies are needed to find the best and most effective way to translate these tests. This would include research, for example, on the following:

- The translation guidelines (e.g., what is the ideal number of specific linguistic translation instructions and the ideal way of presenting them?).
- The number of target versions to be produced (e.g., which produces better translations and is more cost-effective: making one or two target versions?).
- The number of source versions to be used (one or two?).
- The equivalence of the two source versions.
- The ideal way of using the two source versions (e.g., double-translation followed by reconciliation, or translation from one source version with cross-checks against the other?).
- Revision and refining when following the different translation procedures (e.g., when only one target version is made, or when there are two target versions that are reconciled into one).

To tackle these issues, comparisons are needed between the translation procedures recommended in the various organizations conducting international achievement studies and followed in the participating countries (e.g., Finland) and, especially, between the translations made when following each of the procedures.

In addition to the above, however, there are also other issues that are related to the translation of achievement tests but which were not discussed in this paper. These include, for example, the evaluation of language difficulty and text difficulty across languages, the formulation of "translatable" source materials, the functioning of translation teams, statistical and judgmental analyses for evaluating and validating translation quality, and the use of web-based translation platforms and the translation of electronic materials. More research is needed also on all these.

REFERENCES

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education 16* (1), 55-73.

Angoff, W., & Cook, L. (1988). *Equating the scores of the prueba de aptitud academica and the scholastic aptitude test* (College Board Report No. 88-2). New York: College Entrance Examination Board.

Arffman, I. (2007*). The problem of equivalence in translating texts in international reading literacy studies. A text analytic study of three English and Finnish texts used in the PISA 2000 reading test* (Research Reports 21). Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Arffman, I. (2012a). International education studies: Increasing their linguistic comparability by developing judgmental reviews. *ISRN Education 2012*. doi:10.5402/2012/179824.

Arffman, I. (2012b). *Translating international achievement tests: Translators' view* (Finnish Institute for Educational Research, Reports 44). Jyväskylä: Finnish Institute for Educational Research. Retrieved from http://ktl.jyu.fi/img/portal/22708/g044.pdf.

Bechger, T., van Schooten, E., de Glopper, C., & Hox, J. (1998). The validity of international surveys of reading literacy: The case of the Reading Literacy Study. *Studies in Educational Evaluation 24*, 99-125.

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

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment [*Review of the book Knowledge and skills for life: First results from PISA 2000*]. *Assessment in Education 9* (3), 387-399.

Brislin, R. (1986). The wording and translation of research instruments. In W. Lonner & J. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Beverly Hills: Sage.

Chesterman, A. (1997). *Memes of translation.* Amsterdam: Benjamins.

Danks, J., & Griffin, J. (1997). Reading and translation. A psycholinguistic perspective. In J. Danks, G. Shreve, S. Fountain & M. McBeath (Eds.), *Cognitive processes in translation and interpreting* (pp. 161-175). Thousand Oaks, CA: Sage.

Elosua, P., & López-Jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing 7* (1), 39-52.

Englund Dimitrova, B. (2005). *Expertise and explicitation in the translation process*. Amsterdam: Benjamins.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing 5* (1), 23-25.

Ericcson, K.A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review 102* (2), 211-245.

European Social Survey. (2010). *ESS round 5 translation guidelines*. Mannheim, European Social Survey GESIS.

Fontanet, M. (2005). Temps de créativité en traduction [Time for creativity in translation]. *Meta 50*, 432-447.

Gierl, M. J., & Khaliq, S. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement 38* (2), 164-187.

Grisay, A. (2002). Translation and cultural appropriateness of the test and survey material. In R. Adams, & M. Wu (Eds.), *PISA 2000 technical report* (pp. 57-70). Paris: OECD.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing 20* (2), 225-240.

Grisay, A. (2004). *PISA 2000: Differences in item difficulty between English, French and German countries*. Unpublished manuscript.

Grisay, A., de Jong, H., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA Countries. *Journal of Applied Measurement 8* (3), 249-266.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions in PIRLS and PISA reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments 2*, 63-84.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33* (1), 69-86.

Hambleton, R. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment 10*, 229-224.

Hambleton, R. (2001). The next generation of ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment 17*, 164-172.

Hambleton, R. (2002). Adapting achievement tests into multiple languages for international assessments. In A. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58-79). Washington: National Academy Press.

Hambleton, R. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. Hambleton, P. Merenda & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Erlbaum.

Hambleton, R., & Berberoglu, G. (1997). *TIMSS instrument adaptation process: A formative evaluation*. Amherst: University of Massachusetts, School of Education.

Hambleton, R., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technilogy 1*, 1-30.

Hambleton, R., & Zenisky, A. (2011). Translating and adapting tests for cross-cultural assessment. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46-73). Cambridge: Cambridge University Press.

Harkness, J. (2003). Questionnaire translation. In J. Harkness, F. van de Vijver & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35-56). Hoboken, NJ: Wiley.

Harkness, J., Edwards, B., Hansen, S.E., Miller, D., & Villar, A. (2010). Designing questionnaires for multipopulation research. In J. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler, B.-E. Pennell & T. Smith (Eds.), Survey methods in multinational, multiregional, and multicultural contexts (pp. 33-57). Wiley: Hoboken, NJ.

Harkness, J., van de Vijver, F., & Johnson, T. (2003). Questionnaire design in comparative research. In J. Harkness, F. van de Vijver & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 19-34). Hoboken, NJ: Wiley.

Harkness, J., Villar, A., & Edwards, B. (2010). Translation, adaptation and design. In J. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler, B.-E. Pennell & T. Smith (Eds.), Survey methods in multinational, multiregional, and multicultural contexts (pp. 117-140). Wiley: Hoboken, NJ.

Hatim, B., & Mason, I (1997). *The translator as communicator*. London: Routledge.

Hayes, J., Flower, L., Schriver, K., Stratman, J., & Carey, L. (1987). Cognitive processes in revision. In S. Rosenberg (Ed.), *Advances in psycholinguistics. Vol. 2. Reading, writing, and language learning* (pp. 176-240). Cambridge: Cambridge University Press.

International Test Commission. (2010). International test commission guidelines for translating and adapting tests. [http://intectcom. org]

Jeanrie , C., & Betrand, R. (1999). Translating tests with the International Test Commission's guidelines: Keeping validity in mind. *European Journal of Psychological Assessment 15* (3), 277-283.

Jensen, A., & Jakobsen, A. (2000). Translating under time pressure. In A Chesterman, N. Gallardo San Salvador & Y. Gambier (Eds.), *Translation in context. Selected contributions from the EST congress, Granada 1998* (pp. 105-116). Amsterdam: Benjamins.

Jensen, A. (2000). *The effects of time on cognitive processes and strategies in translation*. Copenhagen: Copenhagen Business School, Faculty of Modern Languages.

Jääskeläinen, R. (1996). Hard work will bear fruit? A comparison of two think-aloud protocol studies. *Meta 16*, 60-74.

Jääskeläinen, R. (2010). Are all professionals experts? In G. Shreve & E. Angelone (Eds.), *Translation and cognition* (American Translators Association Scholarly Monograph Series, XV, pp. 213-227). Amsterdam: Benjamins.

Karg, I. (2005). Mythos PISA. Vermeintliche Vergleichbarkeit und die Wirklichkeit eines Vergleichs [The myth PISA. Alleged comparability and the reality of a comparison]. Gottingen: V&R unipress.

Kemper, S. (1983). Measuring the inference load of a text. *Journal of Educational Psychology 75* (3) 391-401.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review 5* (5), 363-394.

Kim, R. (2006). Use of extralinguistic knowledge in translation. *Meta 51*, 284-303.

Kirsch, I.S. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured*. Princeton, NJ: Educational Testing Service.

Kleiner, B., Pan, Y., & Bouic, J. (2009). The impact of instructions on survey translation: An experimental study. *Survey Research Methods 3* (3), 113-122.

Krings, H. P. (1986). *Was in den Köpfen von Übersetzern vorgeht. Eine empirische Untersuchung der Struktur des Übersetzungsprozesses an fortgeschrittenen Französischlernern* [What goes on in the translator's head. An empirical study of the structure of the translation process in advanced students of French]. Tübingen: Narr.

Larson, M. (1998). *Meaning-based translation. A guide to cross-language equivalence* (2nd rev. ed.). Lanham, MD: University Press of America.

Laviosa-Braithwaite, S. (1998). Universals of translation. In M. Baker (Ed.), *Routledge encyclopedia of translation studies* (pp. 288-291). London: Routledge.

Lefevere, A. (1992). *Translating literature. Practice and theory in a comparative literature context*. New York: Modern Language Association of America.

Lörscher, W. (2005). The translation process: Methods and problems of its investigation. *Meta 50*, 597-607.

Mackenzie, R. (1998). Creative problem-solving and translator training. In A. Beylard-Ozeroff, J. Králová & B. Moser-Mercer (Eds.), *Translator's strategies and creativity* (pp. 201-206). Amsterdam: Benjamins.

Mosenthal, P., & Kirsch, I. (1991). Toward an explanatory model of document literacy. *Discourse Processes 14*, 147-180.

Mossop, B. (2007). *Revising and editing for translators*. Manchester, UK: St. Jerome.

Munday, J. (2001). *Introducing translation studies. Theories and applications*. London: Routledge.

Neubert, A. (1997). Postulates for a theory of translation. In J. Danks, G. Shreve, S. Fountain & M. McBeath (Eds.), *Cognitive processes in translation and interpreting* (pp. 1-24). Thousand Oaks, CA: Sage.

Nida, E. (1964). *Toward a science of translating*. Leiden: Brill.

Nida, E., & Taber, C. (1969). *The theory and practice of translation*. Leiden: Brill.

Nord, C. (1991). *Text analysis in translation. Theory, methodology, and didactic application of a model for translation-oriented text analysis*. Amsterdam: Rodopi.

Nord, C. (2006). Loyalty and fidelity in specialized translation. *Confluéncias – Revista de Tradução Cientifica e Técnica 4*, 29-41.

OECD. (1999). *National project manager's manual*. Paris: Author.

OECD. (2005). *PISA 2003 technical report*. Paris: Author.

OECD, (2007, September). *PISA 2009 translation and adaptation guidelines*. Paper presented at the National Project Managers' Meeting, Dubrovnik, Croatia.

OECD. (2009a). *PISA 2006 technical report*. Paris: Author.

OECD. (2009b). *Take the test. Sample questions from OECD's PISA assessments*. Retrieved from http://www.oecd.org/document/31/0,3343,en_32252351_32236191_41942687_1_1_1_1,00.html

Olson, J., Martin, M., & Mullis, I. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

PACTE. (2005). Investigating translation competence: Conceptual and methodological issues. *Meta 50*, 609-619.

Patton, M.Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Pym, A. (1995). European translation studies, Une science qui dérange, and Why equivalence needn't be a dirty word. *TTR 8,* (1), 153-176.

Reiss, K., & Vermeer, H. (1984). *Grundlegung einer allgemeinen Translationstheorie* [Groundwork for a general theory of translation]. Tübingen: Niemeyer.

Rueda, R. (2011). Cultural perspectives in reading: Theory and research. In M. Kamil, P.D. Pearson, E.B. Moje & Afflerbach, P. (Eds.), *Handbook of reading research IV* (pp. 84-104). New York: Routledge.

Rydning, A.F., & Lachaud, C. (2010). The reformulation challenge in translation: Context reduces polysemy during comprehension, but multiplies creativity during production. In G. Shreve & E. Angelone (Eds.), *Translation and cognition* (American Translators Association Scholarly Monograph Series, XV, pp. 85-108). Amsterdam: Benjamins.

Séguinot, C. (1989). The translation process: An experimental study. In C. Séguinot (Ed.), *The translation process* (pp. 21-54). Toronto: H. G. Publications.

Sharkas, H. (2009). Translation quality assessment of popular science articles. Corpus study of the Scientific American and its Arabic version. *Trans-kom 2* (1), 42-62.

Shreve, G. (1997). Cognition and the evolution of translation competence. In J. Danks, G. Shreve, S. Fountain & M. McBeath (Eds.), *Cognitive processes in translation and interpreting* (pp. 120-136). Thousand Oaks, CA: Sage.

Siniscalco, M. (2006). What are the national costs for a cross-national study? In K. Ross & I Genevois (Eds.), *Cross-national studies of the quality of education: planning their design and managing their impact* (pp. 185-209). Paris: UNESCO, International Institute for Educational Planning.

Sireci, S. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice 16*, 12-19.

Solano-Flores, G., Contreras-Niño, L. & Backhoff-Escudero, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales [Translation and adaptation of tests: Learned lessons and recommendations for participant countries in TIMSS, PISA and other international

comparisons]. *Revista electrónica de investigación educativa 8* (2). http://redie.uabc.mx/vol8no2/contents-solano2.html.

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

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science 12* (2), 257–285.

Toury, G. (1995). *Descriptive translation studies and beyond*. Amsterdam: Benjamins.

van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Orlando, FL: Academic Press.

Vermeer, H. J. (1989). Purpose and commission in translational action. In A. Chesterman (Ed.), *Readings in translation theory* (pp. 173-187). Helsinki: Finn Lectura.

Wilss, W. (1990). Cognitive aspects of the translation process. *Language & Communication 10* (1), 19-36.

Zakay, D. (1993). The impact of time perception processes on decision-making under time stress. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 59-72). New York: Plenum.