Janet Harkness, Brita Dorer, and Peter Ph. Mohler, 2016

Introduction
Guidelines
Assessment and survey translation quality
- **1. Assessment as part of team translation.**
- **2. Translation assessment using external translation assessors and verification procedures in a quality con framework paradigm.**
- **3. Translation assessment using survey quality prediction (SQP) coding.**
- **4. Translation assessment using focus groups and cognitive interviews with the target population.**
- **5. Translation assessment using quantitative analyses.**

References
Appendices: A | B

Introduction

This chapter on translation assessment will consider different forms of qualitative and quantitative assessment relate translation, and present the current state of research and relevant literature as available. It is useful to distinguish bet procedures that assess the quality of translations *as translations* and those that assess how translated questions perfo questionnaire instruments. Survey instrument assessment must address both translation and performance quality .

Evaluations of the translations focus on issues such as whether the substantive content of a source question is captur translation, where there are changes in pragmatic meaning (i.e., the meaning perceived by respondents), and whethe technical aspects are translated and presented appropriately (e.g., linguistic and survey appropriateness of response s Approaches combining translation, review, and adjudication, as part of the TRAPD model of translation, are seen to most useful ways to evaluate and improve translation quality and implicitly underscore the relationship between des translation.

Assessments of performance can focus on how well translated questions work for the target population, how they pe comparison to the source questionnaire, and how data collected with a translated instrument compares with data col with the source questionnaire. In the first case, assessment may indicate whether the level of diction is appropriate f sample population; in the second, whether design issues favor one population over another; and in the third, whethe response patterns for what is nominally 'the same question' differ (or not) in unexpected ways across instruments an populations.

Translation quality and performance quality are obviously linked, but good translation does not suffice to ensure tha questions will function as desired in performance. Thus, well-translated questions may work better for a well-educat population than for a less educated population of the same linguistic group, either because the vocabulary is too diff those less educated or because the questions are less salient or meaningful for this group. Problems of question desig as asking questions not salient to the target population, should be addressed at the questionnaire design level; they a difficult to resolve in terms of translation. As testing literature points out, question formats also affect responses if th chosen format is culturally biased and more readily processed by respondents in one culture than in another .

Assessment and evaluation of translation and performance quality assume that there are criteria of evaluation availa which to assess the quality of given translation products and benchmarks, and that standards exist against which tra products can be 'measured.' In the survey research field, there is only limited consensus on what these criteria and benchmarks might be and on what translations that meet these criteria might then look like.

However, items are measurement instruments in comparative survey research. From this follows that in the end, the measurement properties of items must be comparable within well-defined limits in comparative research across cou cultures, or regions. There are a number of statistical methods available that allow the researcher to test for statistica

comparability (or equivalence), ranging from Cronbach's Alpha to Structural Equation Models (see Statistical Analy
Within the Total Survey Error framework, other quality issues must also be dealt with (see below).

The guidelines below include several different qualitative and quantitative approaches for translation assessment, id
criteria of obvious relevance for survey translations, and specifying which may or may not be of relevance in a give
context. It is unlikely that any one project would employ all of the techniques discussed; it is most appropriate for th
and target population to guide researchers in choosing the most efficient methods of assessment.

↑ Ba

Guidelines

Goal: To assess whether the translation of the survey instrument in the target language accurately reflects all aspect
source language instrument.

↑ Ba

Assessment and survey translation quality

Generally, assessment and evaluation require criteria and benchmarks against which to assess the quality of the trans
However, in the field of survey methodology, there is little consensus on the components of such standards. This sec
deal with these issues. It will identify criteria of obvious relevance for survey translations as well as others which m
may not be of relevance in a given context.

↑ Ba

1. Assessment as part of team translation.
Rationale Qualitative assessment of initial translations as they are being developed is an integral and essential comp
team translation procedures (see Translation: Overview). ***Procedural steps***

(See Translation: Overview)

Lessons learned

1.1 The TRAPD model is one effective method of detecting translation errors. See for a discussion of the kinds
mistakes discovered at different stages of translation review in projects based on the TRAPD model.

↑ Ba

**2. Translation assessment using external translation assessors and verification procedures in a quality control
framework paradigm.**
Rationale

Various models use external reviewers and external verification procedures in survey translation efforts. Some proje
currently rely on external review teams to provide most of their assessment; others combine internal assessment pro
with outside quality monitoring.

The word 'verification' in this context refers to a combination of checking the linguistic correctness of the target ver
checking the 'equivalence' of that target version against the source version. 'Equivalence' refers to linguistic equivale
including equivalence in quality and quantity of information contained in a stimulus or test item as well as equivalen
register or legibility for a given target audience . See for more information.

The role of verifiers is to: (a) ensure linguistic correctness and cross-country equivalence of the different language v
of the source instrument; (b) check compliance with the translation annotations provided in the source questionnaire
achieve the best possible balance between faithfulness and fluency; and (d) document all changes for all collaborati
countries and any overall project or study coordinators. Verifiers should ideally have prior experience in verifying (c
producing) questionnaire translations for other cross-cultural social surveys.

Procedural steps

2.1 An external translation verification firm (e.g., cApStAn) uses a monitoring tool—such as the Translation an Verification Follow-up Form (TVFF) used in the European Social Survey (ESS)—to assess translation and ad decisions and to ensure appropriate documentation (see Appendix A; see also Translation: Overview, Append a discussion of the TVFF independent of its utility in assessment).

2.2 The verifier uses the TVFF (or a similar tool) to label each 'intervention' (i.e., recommendation for change c notation) as necessary for each survey item in question.

2.2.1 Examples of intervention categories are 'minor linguistic defect,' 'inconsistency in translation of repeat 'untranslated text,' 'added information,' 'annotation not reflected,' etc. See Appendix B for complete list o intervention categories used in verification of translations of Round 6 of the ESS. See also complete ESS 7 Translation Guidelines .

2.3 The verifiers may prioritize their interventions using the TVFF (or a similar tool):

2.3.1 Interventions are categorized as 'key' (an intervention that could potentially have an impact on how the questionnaire item works) or 'minor' (a less serious intervention that could improve the translation).

2.3.2 This categorization can help translation adjudicators and other team members to identify which errors more/less serious.

2.4 The verifiers may instead be asked to require followup on all interventions by the national teams, as is the c ESS Round 7. The idea behind this decision is that no intervention should stay without followup by the natior teams, or else important corrections may not get made if the national teams don't feel the necessity .

2.5 The TVFF (or other documentation form used) is returned to the national team. Each notation by the verifie be reviewed, and any comments/changes/rejections of suggested changes should be marked accordingly. It ma advisable to require the national teams to get back to the verifiers in order to either confirm acceptance of the verification intervention or, in case these interventions are not incorporated, to justify this decision.

Lessons learned

2.1 The purpose of documenting adaptations and other issues in the TVFF is not only to record such issues, but provide the external verifier with all the relevant background information they will need for the verification assignment, to avoid unnecessary comments and changes, and to be as time-efficient as possible.

2.2 The requirement that national teams provide feedback on whether they incorporate verification intervention TVFF] or not provides better control over how verifiers' suggestions are implemented. In addition, the differe between the verifiers, national teams, and translation experts within the survey may trigger interesting discuss about translation and verification issues.

2.3 Recent use of the verification system by cApStAn in ESS translation assessments has found that verificatio

2.3.1 Enhances understanding of translation issues for:

- The ESS translation team, for languages they do not understand.
- National teams, when choosing a translation by encouraging reflection on choices made.
- Source question designers, enabling them to have a better understanding of different country contexts a in translation.

2.3.2 Enhances equivalence with the source questionnaire and across all language versions, especially for problematic items.

2.3.3 Gives the ESS translation team a better idea of translation quality/efforts/problems in participating cou

2.3.4 Prevents obvious mistakes, which otherwise would lead to nonequivalence between countries, from be fielded.

2.4 Systematic external verification streamlines overall translation quality.

↑ Ba

3. Translation assessment using survey quality prediction (SQP) coding.

Rationale

SQP can be used to prevent deviations between the source questionnaire and the translated versions by checking the characteristics of the items. SQP coding is meant to improve translations by making target country collaborators mo of the choices that are made in creating a translation, and of the impact these choices have on the comparability and reliability of the question. The ESS has been using SQP coding as an additional step of translation assessment since .

Procedural steps

3.1 Provide each study country team with access to the SQP coding system .

3.2 A team member from each study country uses the SQP program to provide codes for each item in the target country's translated questionnaire.

3.2.1 SQP codes refer to formal characteristics of items including:

- Characteristics of the survey question, including the domain in which the variable is operating (e.g., wo health, politics, etc.), the concept it is measuring (e.g., feeling, expectation, etc.), whether social desirab is present, the reference period of the question (past, present, future), etc.
- The basic response or response scale choices (e.g., categories, yes/no scale, frequencies, level of extrem etc.).
- The presence of optional components (e.g., instructions of interviewers and/or respondents, definitions, additional information, motivation).
- The presence of an introduction in terms of linguistic characteristics such as number of sentences, word adjectives, subordinate clauses, etc.
- Linguistic characteristics of the survey question.
- Linguistic characteristics of the response scale.
- The characteristics of any show cards, if used.

3.3 SQP coding can also be used in the process of designing the source questionnaire.

3.4 The team dealing with SQP coding will then compare the SQP codes in the target language(s) and the sourc language.

3.4.1 Differences in SQP coding resulting from mistakes should be corrected.

3.4.2 No action is needed for true differences that are unavoidable (e.g. number of words in the introduction

3.4.3 True differences that may or may not be justified necessitate discussion between the central team and t national team, with possible change in translation necessary.

3.4.4 True differences that are not warranted (e.g., a different number of response categories between the so target language versions) require an amendment to the translation as submitted.

Lessons learned

3.1 In Round 5 of the , SQP coding produced valuable information that helped to detect deviations in translatio had they gone undetected—would have affected the quality of the items, as well as the design of experiments.

3.2 See the ESS Round 6 SQP Guidelines and Codebook for further detail.

↑ Ba

4. Translation assessment using focus groups and cognitive interviews with the target population.

Rationale Various pretesting methods using both focus groups and cognitive interviews can be used to gain insight i appropriateness of the language used in survey translations.

Procedural steps

4.1 Focus groups can be used to gain target population feedback on item formulation and how questions are per They are generally not suitable for assessment of entire (lengthy) questionnaires. To optimize their efficiency, materials pertinent for many items can be prepared (fill-in-the blanks, multiple choice, etc.), and participants asked to explain terms and rate questions on clarity. At the same time, oral and aural tasks are more suitable th written when target population literacy levels are low or when oral/aural mode effects are of interest.

4.2 Cognitive interviews allow for problematic issues to be probed in depth, and can identify terms not well und across all sub-groups of the target population.

4.3 Protocols should be developed and documented for all types of pretests, with particular care toward designs investigate potentially concerning survey items (see Pretesting).

4.4 Interviewer and respondent debriefings can be used after all types of pretests, with full documentation of de to collect feedback and probe comprehension of items or formulations.

Lessons learned

4.1 Focus groups and cognitive interviews are useful for assessing questions in subsections of the target popula example, focus groups conducted to validate the Spanish translation of the U.S. National Health and Sexual B Study (NHSB) revealed that participants did not know certain terms, considered unproblematic up to that poir related to sexual organs and behaviors .

4.2 Interviewer and respondent debriefing sessions are valuable opportunities for uncovering problematic areas translations. Debriefing sessions for the 1995 ISSP National Identity module in Germany revealed comprehen problems with terms covering ethnicity and confirmed cultural perception problems with questions about “tak pride” in being German .

4.3 Tape recording of any pretesting allows for behavioral coding for particular questions of interest. provide ev that behavior coding accurately identifies questions that are believed a priori to have clear comprehension and mapping flaws. Based on their study on the comparability of behavior coding across respondents of different groups (non-Hispanic African-American, Korean-American, Mexican-American, and non-Hispanic White), th authors found, in general, a high degree of consistency in how survey respondents from various cultural backg express difficulties processing survey questions as part of respondent-interviewer interactions that can be dete through behavior coding.

4.4 If computer-assisted pretesting is used, paradata such as timestamps and keystroke data can be used to ident items that are disrupting the flow of the interview and may be present due to translation issues .

↑ Ba

5. Translation assessment using quantitative analyses.

Rationale

Textual assessment of translation quality does not suffice to indicate whether questions will actually function as requ across cultures; statistical, quantitative analyses are required to investigate the measurement characteristics of items assess whether translated instruments perform as expected. The central aim is to detect biases of different types that measurement systematically. Statistical tests can vary depending on the characteristics of an instrument, the sample available, and the focus of assessment (for general discussion, see , , , , , , and).

Procedural steps

5.1 Variance analysis and item response theory can be used to explore measurement invariance and reveal diffe item functioning, identifying residual translation issues or ambiguities overlooked by reviewers .

5.2 Factor analysis (adapted for comparative analyses: exploratory factor analysis or confirmatory factor analys multidimensional scaling can be used to undertake dimensionality analyses . See Statistical Analysis for more information.

5.3 For the evaluation of individual items, item bias can be estimated using multitrait-multimethod (MTMM) procedures, as described in and .

Lessons learned

5.1 Some procedures, like SQP used in the ESS , rely on intensive analyses of questions collected, like a corpus linguistics. However, the questions accepted as input in the corpus were not systematically evaluated using sta quality inspection, such as checking for double-barreled, double negation, response scales that do not fit the q etc. Thus, the scores obtained might be biased, and researchers should carefully use such systems.

5.2 Where scores are relevant (e.g., in credentialing tests), a design is needed to link scores on the source and ta versions .

5.3 The emphasis placed on quantitatively assessing translated instruments and the strategies employed differ a disciplines.

5.3.1 Instruments that are copyrighted and distributed commercially (as in health, psychology, and educatio also often evaluated extensively in pretests and after fielding.

5.3.2 Some quantitative evaluation strategies call for a large number of items (e.g., item response theory), ar thus unsuitable for studies that tap a given construct or dimension with only one or two questions.

5.3.3 Small pretest sample sizes may rule out strategies such as multidimensional scaling and factor analysis

5.3.4 Some assessment techniques are relatively unfamiliar in the social sciences (e.g., multitrait-multimeth

5.4 Post hoc analyses that examine translations on the basis of unexpected response distributions across langua usually intended to help guide interpretation of results, not translation refinement. Caution is required in using procedures for assessment because bias may also be present when differences in univariate statistics are not.

5.5 For multi-wave studies, document any post hoc analyses for consideration when carrying out future translat

↑ Ba

References

{2265844:WIZ9W4KX};{2265844:ANG6TTIF},{2265844:A7AZ5TUL},{2265844:NBGFDWNT};
{2265844:25BJAQUB},{2265844:BGCZP7E8};{2265844:TNMYF2MY};{2265844:CERE8EDX};
{2265844:J3QVK33R};{2265844:883WBJP7};{2265844:883WBJP7};{2265844:JUGNDB7L};{2265844:LJ6JG5
{2265844:JUGNDB7L};{2265844:2P4ZV8BA};{2265844:V57GHQF6};{2265844:XXQ5X5HX};
{2265844:XXQ5X5HX};{2265844:WIZ9W4KX};{2265844:YPJN7XRB};{2265844:T58BKUJR};
{2265844:ANG6TTIF};{2265844:KQ7E4IUP};{2265844:V2JU79JA};{2265844:JIJSW3HW};{2265844:P3M6I6
{2265844:YFYE79Z4};{2265844:BGCZP7E8};{2265844:KUBWSP7R},{2265844:GJIFFCYD},{2265844:L2A9
{2265844:VITEC3CE};{2265844:EDREPDNB},{2265844:VDXKMEZK},{2265844:BGCZP7E8};
{2265844:AAVHWBGE};{2265844:9MRNKMU3};{2265844:LJ6JG5RQ};{2265844:ANG6TTIF} apa creator asc

↑ Ba