Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 172–190.

Methodological considerations for survey research: Validity, reliability, and quantitative analysis

Christopher D. Mellinger
University of North Carolina at Charlotte, United States
cmelling@uncc.edu
https://orcid.org/0000-0003-4915-8821

Thomas A. Hanson
Butler University, United States
tahanson@butler.edu
https://orcid.org/0000-0001-9074-8829

Abstract

As translation and interpreting studies continue to develop cognitive theories of translator and interpreter behavior and processing, there has been increased emphasis on research methods and data collection methodologies to glean new insights into the translation process. This article presents a critical review of survey research methods in Cognitive Translation Studies and argues for their inclusion as a means of better understanding translator and interpreter attitudes, behaviors, perceptions, and values. The article begins with a reflection on measurement and the need for alignment with theoretical frameworks and constructs; then it reviews important considerations when developing theoretically-grounded, empirically-based survey instruments, namely, validity, reliability, measurement invariance, and quantitative analysis. The article concludes with a call for additional methodological reflection on developing and using survey instruments.

Keywords: survey; validity; reliability; measurement invariance; Cognitive Translation Studies

1. Introduction

Cognitive translation and interpreting studies (CTIS) encompass an array of research areas that share a common interest in the cognitive behaviors and processes of translators, interpreters, and users of language services (Risku, 2012). This definition overlaps with substantial portions of reception studies (Kruger et al., 2016), translator studies (Chesterman, 2009), and translation process research (Muñoz Martín, 2016; Shreve & Angelone, 2010a), while also engaging with theories and methods from adjacent disciplines (Halverson, 2010). Regardless of the specific topic or focus of research in this area, however, empirical studies of cognitive aspects of translation and interpreting necessitate indirect measurement of cognitive processes, which cannot be directly observed (Dancette, 1997). Therefore, valid and

reliable measurement of theoretical, latent constructs has been a principal concern for the purpose of quantifying unobservable variables.

Consequently, the history of Cognitive Translation Studies is replete with techniques to procure numerical data that reflect various cognitive processes, and many authors have outlined the history of innovation in such research methods (e.g., Alves, 2015; Alves & Hurtado Albir, 2017; Jääskeläinen, 2011; Muñoz Martín, 2017; Shreve & Angelone, 2010b). These histories generally trace the first phase of cognitive studies to the mid-1980s and the predominance of descriptive research based on data collected with think-aloud protocols (TAPs), a method drawn from the field of cognitive psychology (Ericsson & Simon, 1984). Over time, concerns emerged regarding ecological validity, the limits of concurrent and retrospective verbalizations, and the lack of rigorous methodological implementation that might hinder the utility of TAPs (e.g., Bernardini, 2001; Jääskeläinen, 2010; Li, 2004). In response, a second phase emerged that emphasized product-oriented research through keystroke logging, screen recording, and triangulation of multiple research methods (Alves, 2003; Jakobsen, 2003). Methodological innovation in the form of eye-tracking and pupillometry later ushered in a third phase marked by these new methods to provide reflective measures of cognitive processes (Hvelplund, 2014; O'Brien, 2009). Therefore, the overarching history is often presented as a series of innovations in data collection methods and related technology.

If one hopes to identify a current, fourth phase in the field, it might be best characterized by ongoing methodological innovation and triangulation (see Alves & Hurtado Albir, 2017). However, the apparent consensus about the previous three phases can also obscure the diversity of methods, the interdisciplinarity, and the overlapping adoption, which have always characterized Cognitive Translation Studies (Alvstad et al., 2011; O'Brien, 2013). Additional data collection methods – corpus analysis, imaging technologies (e.g., electroencephalography [EEG], functional magnetic resonance imaging [fMRI]), and other psychophysiological measures (e.g., heart rate, blood pressure, stress hormones) – suffice as examples to demonstrate the available range of quantitative methods. These methods have been further augmented by qualitative approaches (Risku, 2014) and triangulation that combines various forms of evidence.

Concomitant with the recent flourishing of innovation in methods is a renewed emphasis on theory building and more formal definition of constructs (e.g., House, 2013; Muñoz Martín, 2016, 2017; Shreve & Angelone, 2010b). Perhaps with further hindsight a future generation of scholars will condense the first three phrases of research in Cognitive Translation Studies into one longer era of method-driven innovation (in which TAPs, keystroke-logging software, and eye-tracking hardware drove changes in empirical research) and a currently emerging second era including greater attention to theory building, consolidation, and testing. One such attempt to identify broad theoretical paradigms is Muñoz Martín's (2017) distinction between *computational translatology* and *cognitive translatology*. The latter emphasizes that cognition is embodied, embedded, enactive, extended, and affective (4EA; Clark, 1996) and recognizes the role of context and situated cognition (Pöchhacker, 2005; see also Muñoz Martín, 2016).

Notably lacking in this history of methodological innovation and theoretical development is the rigorous use of psychometric testing as a supplement to other measurement techniques. Whereas surveys do appear in the published literature, they have not played a prominent role in theory development and testing, despite their utility in such related fields as psychology and communication studies (Boyle et al., 2015; Groth-Marnat & Wright, 2016; Rubin et al., 2011). Measures of cognition and related constructs are necessarily proxies that presume a high degree of correlation between an assumed, underlying construct and an observable phenomenon. Therefore, constructs in Cognitive Translation Studies can be measured by responses to questionnaires in addition to the other common data collection methods outlined above.

There have been occasional calls for the development of valid and reliable questionnaires (e.g., Alves & Hurtado Albir, 2017), and some discipline-specific scales are available (e.g., Lee, 2014). However, the traditional Likert-type scale has not received the same rigorous treatment and wide application of other methods, such as eye-tracking and keystroke logging. Therefore, the present study provides a critical review of survey instruments as a theoretically-grounded measure that can help with understanding the various traits and characteristics of translators and interpreters, on the one hand, and the users of language services, on the other. This review focuses on validity as an explicit link between theory and survey instrument development, the importance of establishing reliability and measurement invariance, and the analysis of quantitative survey data. The overall aim is to link theoretical and methodological work from survey design and statistics to cognitive research on translation and interpreting (T&I).

2. Measurement and Likert-type scales

Surveys can be used to assess latent constructs, which are well-defined, theory-based concepts that yield testable hypotheses. Whereas any measurement technique is susceptible to misuse, survey instruments may be particularly vulnerable because they can be created, circulated, and analyzed with relative ease. Surveys (and especially online surveys) have the additional allure of a potentially larger sample size that can span multiple geographic areas and reach a diverse and scattered sample of respondents (Mellinger, 2015). However, poorly designed measurement scales can invalidate statistical analysis, leading to errors in inference and implications of empirical research. This issue is compounded by a dearth of critical reflection on survey methods in translation and interpreting studies. Therefore, this section briefly reviews the philosophy of measurement and the specific format and construction of Likert-type scales[1] that, we argue, can possibly be of use in Cognitive Translation Studies.

The act of measurement presupposes a theoretical framework, and Borsboom (2005) advocates that psychometrics – specifically, latent variable analysis – demands ontological realism of attributes in order to justify the effort to measure them. This philosophical stance is rarely explained directly in discussions of research methods, but it is implicitly embraced by many empirical scholars. Consider, for example, House's (2013) question of whether observable behavior, such as keystroke logging, is truly informative about unobservable cognitive processes. Similar questions have been raised about the effectiveness of TAPs,

fMRIs, and every available measurement tool, and also about whether the observed data correlate with cognitive processes or other attributes of the individual. However, such critiques implicitly assume the existence of underlying cognitive and/or affective phenomena in theorizing about their nature and in attempting to describe valid and reliable means of measuring them for use in quantitative analysis.

The key issue for any measurement is whether the manifest, observable variables are useful and adequate proxies for the underlying phenomenon of interest. For example, much of the eye-tracking research rests on the eye–mind assumption (Just & Carpenter, 1980; see also Hvelplund, 2014, 2017) to justify the use of specific eye movements and visual attention as indicators of cognitive processes. The use of surveys similarly presupposes the existence of an underlying construct and the correlation of the measurement with its degree or intensity. For instance, Angelelli's (2004) Interpreter Interpersonal Role Inventory (IPRI) measure makes two contentions: first, that visibility is a meaningful and stable construct; second, that the IPRI survey is able to distinguish different levels of that construct among respondents. The first claim is one of ontology, while the second is one of validity.

By advocating ontological realism, we follow Borsboom (2005) by considering solely reflective surveys – in which an underlying attribute is assumed to be the source of variation in responses – while omitting consideration of formative surveys (cf. Edwards & Bagozzi, 2000). The difference might be most easily illustrated in reception studies – for example, satisfaction with subtitles. A formative construct (also sometimes referred to as an *emergent construct* or an *index*) might use items concerning the size, color, placement, and pacing of text on the screen; satisfaction on each of these individual aspects would be summed to a total satisfaction score. In contrast, a reflective scale would consider the implications of satisfaction and possibly include items about enjoyment, understanding, and/or intention to view more subtitled content.

Reflective latent variables are favored for several reasons. First, the mathematics of factor analysis relies on the assumption that the observed items covary due to their joint causation by the underlying construct. Second, an ongoing debate among philosophers, statisticians, and applied researchers across disciplines questions the distinction, value, and even the legitimacy of formative measurement scales. Wilcox et al. (2008) review this debate and conclude that formative measures are problematic when used for measuring latent variables.[2] Finally, almost all scales used in psychology and social science research are reflective scales (Bollen, 2002). Therefore, for statistical, theoretical, and practical reasons, theory-driven research in Cognitive Translation Studies should favor reflective measurement scales by first defining the latent trait to be studied and then considering which observable items will reflect that latent attribute.

A fuller discussion of the ontology and epistemology of surveys lies beyond the scope of this article, but these issues deserve at least brief acknowledgement before proceeding to the practical problems of survey design. The agenda of cognitive translation scholars with an interest in surveys should begin with theoretical matters and the operationalization of discipline-specific constructs as well as the recognition and reuse of extant measures and

constructs from neighboring disciplines. Then, scale development can proceed with item writing, pilot testing, and factor analysis to provide a sound basis for applied research.

One example of the need for the alignment of philosophy, definition of constructs, and practical measurement is the concept of translation competence. Multiple approaches exist for operationalizing competence, with studies having been conducted in an effort to identify various subcompetences of the translation task (e.g., Hurtado Albir, 2017) and to understand the extent to which competence has been acquired by students or novices. More recently, a questionnaire has been developed to examine competence using self-report data (Schaeffer et al., 2020). These types of instruments hinge on the existence of an underlying construct that can be measured. However, recent scholarship has called into question the utility of competence as a theoretical construct that is grounded in the extant literature in psychology (e.g., Shreve et al., 2018). The potential disconnect between instrument development and the theoretical status of the underlying construct may raise questions about the construct validity of these instruments. Moreover, this lack of alignment demonstrates the iterative nature of research and the imperative for continuous refinement, with research studies serving as the foundation for theory development, which can then be examined through empirical work. As debate and empirical research continue, a pressing issue will be harmony among theory, construct definitions, and measurement tools.

Surveys can take many forms, but this review concentrates on Likert-type scales as a means to measure attitudes (Likert, 1932). The format of such a scale consists of individual items in the form of statements (not questions) to which respondents mark their level of agreement or disagreement. The responses are quantified (including any necessary reverse coding) and then summed to yield a respondent's score. Not all surveys that use a multiple-choice response format are properly called Likert-type surveys. That designation should be reserved for any survey that is conceived as a unified instrument to reflectively measure a construct with the intention that the item scores be summed for analysis. Likert-type scales can be used in a wide range of applications to measure latent constructs that indicate attitudes, knowledge, perceptions, and values (Vogt & Johnson, 2016).

Cognitive translation scholars have previously used Likert-type scales to explore the relationships among various constructs and traits in the context of translation and interpreting. Two of the many available examples are personality traits (Hubscher-Davidson, 2009) and self-efficacy (e.g., Bolaños-Medina, 2014; Jiménez Ivars et al., 2014; Lee, 2014, 2018). There are also examples of scales developed to measure constructs directly related to issues of language, translation, and interpreting, such as interpreter visibility (Angelelli, 2004) and language learning motivation (Csizér & Dörnyei, 2005). Unfortunately, other published literature does not always exhibit the same level of rigor demonstrated in these studies. Common modeling and statistical errors can lead to confounded research instruments and undermine the researcher's ability to draw conclusions.

Increasing the utility and legitimacy of survey scales in Cognitive Translation Studies requires recognition that the purpose of a scale is to quantify latent constructs. The true relationships among theoretical constructs cannot be directly observed, but the statistical relationships

among the measured variables from survey scales can be computed to test hypotheses. The power of the statistical tests and the legitimacy of the conclusions rely on the presumption of both validity and reliability of a reflective survey scale; unstable measurement contributes to smaller estimated effect sizes through attenuation bias and a corresponding increase in Type II statistical error (see Mellinger & Hanson, 2017). Given the philosophical importance and practical implications of survey quality, empirical researchers in Cognitive Translation Studies need to follow best practices in survey design and implementation to ensure accurate measurement. In order to contribute toward that end, the remainder of this article proceeds by discussing issues of validity, reliability, and quantitative analysis of surveys. These methodological discussions are then examined in light of cognitive translation and interpreting studies as a call for their inclusion in the methodological repertoire of T&I researchers interested in cognition.

2.1 Validity

Validity can be described as the property of a scale to produce a measurement that accurately reflects an underlying construct. In other words, the scale measures what it intends to measure (Litwin, 1995). Validity can also be thought of as alignment between a measure and theoretical definitions, relationships, and predictions (Messick, 1995). Therefore, validity is the primary concern for any scale development and for the evaluation of scales for reuse (AERA, 1999). In addition to creating and validating scales specifically for use in Cognitive Translation Studies, there is an opportunity to contribute to other social sciences that acknowledge the importance and influence of translation in adapting scales for multiple languages (Hambleton & Patsula, 1998; Smith, 2010). This section discusses the philosophy and terminology of validity while providing several examples from translation and interpreting studies.

Validity is a holistic evaluation that a scale is appropriate, useful, and meaningful in measuring a construct (Kane, 1994) and has traditionally been conceived of in three broad categories: content, construct, and criterion validity (Cronbach & Meehl, 1955). However, modern scholarship stresses that validity is a single property of a test (e.g., AERA, 1999). In particular, Messick (1995) proposed consolidating all validity under the umbrella of construct validity while also describing six aspects to be considered in evaluating validity, notably including the impact of a survey's use on respondents. In a succinct definition, Borsboom (2005) argues that a test is valid if and only if it measures an existing, underlying attribute that causes observable variation in the measurement outcome. Recent standards stress the unitary nature of validity, but for the applied researcher the traditional tripartite division (i.e., content, construct, and criterion validity) can still provide a useful scheme for accumulating and describing evidence in the process of validating a scale (e.g., Goodwin & Leech, 2003). Indeed, discussions of these so-called types of validity can be found in handbooks on T&I research methods and research studies as forms of evidence supporting claims of measurement validity.

Content validity describes the extent to which a survey covers all aspects of a construct and also subsumes the more superficial standard of face validity, which is the extent to which the items in a survey appear relevant to a reader familiar with the construct being measured (DeVellis,

2017). Therefore, content validity relies on theory to describe the construct to be measured. In particular, theory provides insight into the relevant wording, concepts, and dimensionality of a construct. For example, Lee (2014) bases the Interpreting Self-Efficacy scale on social cognitive theory and on related scales that define self-efficacy as encompassing three factors: self-confidence, self-regulatory efficacy, and preference for task difficulty.

Content and construct validity are established, in part, by including questions that align with each of these dimensions. Moving forward, an important task of scholars in Cognitive Translation Studies is to develop and probe theories for construct definitions and their associated dimensionality to create and test measurement scales that provide valid inferences.

While there is no statistical test for content validity, correlation coefficients are often employed as partial evidence for the other two traditional types of validity: criterion and construct validity. Criterion validity considers the extent to which a scale aligns with an observable trait and encompasses the subcategories of predictive and concurrent validity. Examples of predictive validity often arise in studies related to student performance, screening, and proficiency tests (e.g., Bontempo & Napier, 2011; Lee, 2018).

Meanwhile, construct validity involves correlations with other latent variables in a nomological network (Cronbach & Meehl, 1955). Validating a scale for cognitive translation theories involves collecting evidence of correlations with both manifest variables and a web of relationships with other constructs. Ongoing research on default translation (Halverson, 2019) illustrates these multiple evidential processes. While product-oriented research examines the output of the translation process, other research considers the psychological and cognitive processes that might lead to the existence of a default translation. Therefore, both direct observation and theoretical constructs are considered, which is an example of the types of multiple validation techniques needed for surveys.

Validity is a characteristic of a scale in its particular use and context (Chan, 2014). To illustrate this point with an admittedly extreme example, a scale to measure introversion might be well-conceived and valid for that purpose, but that same scale would clearly be invalid and useless as a measure of translation competence. Adaptation and borrowing of scales from adjacent disciplines is useful, but the practice demands reflection on the instrument's validity and theoretical alignment if the underlying construct is not identical. To date, survey development in CTIS has been too ad hoc and has lacked sufficient theoretical motivation (Muñoz Martín, 2017; Shreve & Angelone, 2010b). Some exceptions do exist (e.g., Angelelli, 2004; Csizér & Dörnyei, 2005; Lee, 2014), but the advancement of the discipline requires explicit alignment of theory and survey scales to provide valid measurement and to aid replication.

2.2 Reliability and measurement invariance

The reliability of a survey instrument refers to its ability to produce consistent and reproducible results. For a reliable survey scale, the observed variation in numerical measures is presumed to arise from measurement error (Nunnally, 1978), and the results should be

stable across time (test–retest reliability), items (internal reliability), and groups (measurement invariance).[3]

The purpose of establishing reliability is to separate variability due to measurement error from true differences attributable to the underlying construct. Similarity of multiple measurements decreases error in the measurement tool and improves the power and interpretation of subsequent statistical analysis. Moreover, reliability allows results obtained from survey instruments to be compared more confidently across research studies, in this way facilitating theory building through replication.

Perhaps the most widely-used method for reporting reliability is Cronbach's alpha, which is a measure of internal reliability based on the proportion of variance that can be attributed to a latent variable (DeVellis, 2017). Alpha is appropriate for latent variable analysis (though not for formative scales; see Streiner, 2003), and the statistic is often described as the average of all split-half reliabilities (Warrens, 2015). Common lore among applied researchers is that Nunnally (1978) justified 0.70 as the standard level for acceptability. However, as with any statistical rule of thumb, this figure is only one benchmark, and the evaluation of reliability should consider multiple factors in a more complete assessment of reliability, including the number of items in the scale and its intended use (Cortina, 1993; Peters, 2014).

For several reasons, the property of reliability cannot be fully established by reporting a single statistic (e.g., Cronbach's alpha) in the initial development of a scale. First, any computed reliability coefficient is a function of the sample data and not an established quality of the survey instrument itself, so researchers must report Cronbach's alpha every time that the survey is administered (DeVellis, 2017). A lack of survey instruments in the field of T&I research makes this somewhat uncommon to date. However, there are examples in the extant literature. For example, Mellinger and Hanson (2018) reported alpha coefficients from published examples in previous studies along with the figures from their sample as part of their methodological discussion of several survey instruments. In addition, a confidence interval for Cronbach's alpha can be reported to provide further information about the likely range of the true value (Mellinger & Hanson, 2017).

Yet, Cronbach's alpha has some notable statistical shortcomings, including variations due to survey length, inter-item correlation structure, and sample characteristics (Agbo, 2010). For this reason, additional techniques should be coupled with reporting the single statistic. The assessment of reliability can also include item analysis, which could be informal assessment of language, leave-one-out analysis using Cronbach's alpha, or item response theory employing item characteristic curves. Alternative measures, such as omega, have also been proposed (Dunn et al., 2014). Software implementation and widespread adoption often lag statistical innovations, which reinforces the need to remain current with one's reading and training in quantitative methods and/or collaborate with statisticians and psychometricians in conducting empirical research.

Internal reliability considers only the relationship among responses to the items of a scale, but nearly every aspect of survey design has been examined for the possibility of both the

introduction of bias and the influence on data quality (Choi & Pak, 2005; DeCastellarnau, 2018). Brown (1996) categorizes many of the possible influences into five sources: (1) the test itself, (2) scoring procedures, (3) administration procedures, (4) the test environment, and (5) the individual examinees. Other factors that could affect reliability include respondent motivation and the thoroughness and comprehensibility of the instructions. The diversity of possible influences implies the need for thoughtful choices in all aspects of survey design and administration. The discussion below highlights three issues that are especially common in Cognitive Translation Studies: online administration, translation of surveys, and cross-cultural differences.

First, online surveys are a common modality to conveniently reach a larger sample of the geographically-dispersed population of professional translators and interpreters (Mellinger, 2015). However, by using this data collection technique the researcher relinquishes control of the testing environment and cannot answer any questions, to name just two potential threats to reliability. Whereas much of the commentary related to online surveys has focused on data security and ethics (e.g., Buchanan & Hvizdak, 2009), recent years have seen an increased interest in the psychometric properties of online surveys. Generally, results have shown that online administration of a previously-developed survey does not damage its internal reliability (e.g., Zlomke, 2009). Still, researchers should report how the data were collected, explain whether the survey had been developed or validated for that modality, and describe any potential problems with reliability as a result of the data collection method.

A second threat to reliability is the possible effect of translation on survey responses (Harkness et al., 2004; Harkness et al., 2010). For instance, lexical choices that increase ambiguity or alter the valence of the items can affect responses and reliability, whereas mistranslations may undermine the researcher's ability to measure any potential underlying construct. These challenges can also manifest when adapting materials into signed languages (Graybill et al., 2010).

A third issue that can influence reliability is data collection across different cultural groups (e.g., McGorry, 2000). Reliability can be degraded due to a lack of familiarity with the format of Likert-type scales and cultural bias. For instance, Flaskerud (2012) documents the influence of a respondent's literacy on survey data, and Lee et al. (2002) reveal cross-cultural differences in respondents' willingness to select extreme answers at the endpoints of the scale. Translation and interpreting studies researchers are typically attuned to the challenges of working with multiple, distinct groups; however, explicit reflection on this topic is often taken up by those outside of the discipline. Consequently, this aspect of T&I research methods may be an area worth greater attention as the field continues to evolve.

If a scale is tested and found to behave similarly across a range of samples, it can be said to possess measurement invariance. More formally, measurement invariance concerns the factor structure (configural invariance), factor loadings (metric invariance), mean comparisons (scalar invariance), and equality of variance and error (strict invariance). Measurement invariance has received less attention than validity and reliability in T&I research; its importance is perhaps more recognized in psychology (e.g., Kankaraš & Moors,

2010; Lubke et al., 2003; Milfont & Fischer, 2010). For example, the Beck Depression Inventory and Children's Depression Inventory both measure depression but for adult and youth populations respectively; meanwhile, IQ testing has a long and troubled history with cross-group comparisons and issues of measurement invariance (Wicherts & Dolan, 2010). Strict measurement invariance is necessary for direct comparisons across groups, and it is a difficult standard for any scale to meet. CTIS naturally involves multicultural, multilinguistic samples. Scales that lack measurement invariance could be interpreted differently across these groups and yield non-comparable results (Coulacoglou & Saklofske, 2017). As the development and use of Likert-type scales expands, establishing measurement invariance will only become more important.

Multiple approaches have been developed to deal with the issues of creating reliable and invariant surveys across diverse samples (e.g., King et al., 2004). Because psychometric properties are established, in part, through data collection, the nature of the respondents influences the structure and properties of a survey. Therefore, the measurement provided by a scale can be presumed to be valid and reliable only for respondents who are similar to the original sample used to develop the scale. Larger samples, increased replication, and the adoption of best practices in survey methods will allow for the valid use of scales and their widespread adoption.

2.3 Quantitative analysis

Rigor in quantitative analysis in translation and interpreting studies continues to improve in terms of statistical design, analysis, and reporting. Scholars who examine large datasets derived from eye-tracking, keystroke logging, and corpus studies have explored a sophisticated range of quantitative tools (e.g., Balling, 2008; Oakes & Ji, 2012), and general volumes on research methods have further contributed to this trend (e.g., Angelelli & Baer, 2016; Mellinger & Hanson, 2017; O'Brien & Saldanha, 2014). In this section, we highlight three common errors in survey analysis. The first two errors were selected because of their prevalence in reported research, whereas the third error relates to the underlying mathematical structures involved in the analysis of surveys.[4]

One common error in survey analysis is conducting single-item comparisons. Because a Likert-type scale is conceived and constructed as a unified instrument, only the summed scores should be subject to statistical analysis. In particular, comparisons of the means of single items are almost never appropriate (Carifio & Perla, 2007). Reported results must maintain the distinction between single items and scales: individual items can be summarized and described only qualitatively, whereas summed scales are appropriate for statistical testing. Such is the case across disciplines; however, T&I research that draws on survey data has unfortunately, at times, relied on single-item comparisons to draw larger conclusions. Researchers should always be cautious of overgeneralization based on a single test or result, and survey results are no exception. Our intent here is not to single out studies that have conducted single-item analysis; rather, we hope to cast a more critical eye on results from survey research and present ways by which the methods can be improved.

A second error is the belief that standard parametric analyses cannot be applied to data collected with Likert-type scales. While it is true that some scholars raise concerns about the level of measurement of a survey scale, arguing that nonparametric methods are more appropriate (e.g., Jamieson, 2004), the consensus among statisticians is that treating the summed scales as continuous data and conducting traditional parametric analysis will yield acceptable statistical results (Carifio & Perla, 2008; Norman, 2010). Only in unique cases such as severe departures from normality, one-sided tests, moderate sample sizes, or considerable differences in sample size among groups is nonparametric analysis likely to be required (Harpe, 2015).[5]

A third error that is sometimes made in the use of survey data is using a statistical method that is inappropriate for latent factor analysis. One specific example is the use of principal components analysis, which should be supplanted by exploratory factor analysis in determining the factor structure of scales (Fabrigar et al., 1999). Additional examples are the incorrect use of path analysis, which should be reserved for use with manifest variables, and partial least squares, which is less powerful than factor analysis of latent variables when sufficient sample sizes are collected (Cole & Preacher, 2014; Rönkkö & Evermann, 2013). The primary issue is that applied research needs to select the appropriate statistical model that aligns with the data and the research question, while understanding the relationships among the variables (Edwards & Bagozzi, 2000).

Scholars who wish to develop new survey scales must pay particular attention to the proper use of factor analysis, although full treatment of the topic is impossible within the confines of a single article.[6] The mathematics of factor analysis is typically distinguished between exploratory and confirmatory models, although these models are nested within the overarching topic of structural equation modeling (SEM), which comprises measurement and structural models and unifies such disparate approaches as path analysis, factor analysis, and item response theory models (Beaujean, 2014). Exploratory factor analysis (EFA) involves a number of decisions, both practical and statistical. The practical decisions include study design, construct definition, and sample selection. The statistical choices consist of selecting a model fitting procedure, identifying the number of factors, determining a rotation methodology, and deciding which items to retain (Fabrigar et al., 1999). As a complement to the selection procedures of EFA, confirmatory factor analysis (CFA) involves comparisons of data with *a priori* models (Beaujean, 2014). The primary outputs of CFA are model fit indices with which to assess alignment with the theoretical construct. Once measures are determined to fit the hypothesized definitions and to possess acceptable validity and reliability, the relationships among various constructs and other observed variables can combine the measurement model of factor analysis with a structural model to test hypotheses.

In this section, the discussion has centered on latent factor analysis, which is a subject-focused approach to modeling, in contrast to item response theory (IRT) and the Rasch model, which consider both the subject and the survey items. Both approaches (i.e., latent factor analysis and IRT) have their advocates, strengths, and weaknesses. For instance, IRT is favored in educational testing and in any setting with dichotomous (e.g., right/wrong) responses (Wirth & Edwards, 2007). However, factor analysis is an appropriate tool for initial scale

development to measure continuous latent constructs and for application in social scientific research. Furthermore, innovations and developments in statistical methods is always ongoing (e.g., van Bork et al., 2019). Methodological and analytical diversity can strengthen a field, and researchers must select statistical methods that are appropriate to their research questions.

3. Conclusion

A wide range of research methods are available in cognitive translation and interpreting studies with increasingly more refined approaches to measurement and triangulation. Having multiple tools available enables researchers to explore constructs and hypotheses that were previously more difficult to observe, with the results now providing insights into cognitive theories of translation and interpreting. The present critical review of survey instruments explains some important aspects of Likert-type scales and suggests their utility in translation and interpreting studies. Several examples from the field are provided to illustrate their potential use, given their ability to examine underlying latent constructs that may inform our understanding of behaviors, attitudes, and perceptions during the translation and interpreting task.

The discussion here has emphasized the creation and analysis of new survey instruments specific to CTIS, given the need to align directly with theory development and testing. We have also addressed how the adoption or adaptation of existing scales from neighboring disciplines can provide researchers with useful sources of measurement. The topic of survey translation has been largely omitted, although translation scholars could play an important role in developing the scholarship on that topic. Over-reliance on back-translation and notions of equivalence are problematic in much of the literature on this topic (e.g., Behr & Shishido, 2016). The perspective here is also limited to that of the researcher, although a substantial body of work exists on the survey response process (Schwarz, 2007). All surveys require consideration of validity, reliability, and rigorous quantitative analysis, which is the motivation for their selection here.

CTIS has developed in parallel with new data collection methods, including TAPs, keystroke logging, eye-tracking, and other innovative technologies. A current shift in emphasis in the field should encourage the development and refinement of theories in tandem with improvements in research methods and cross-discipline collaboration to allow for generalization and the advancement of Cognitive Translation Studies as a rigorous science (House, 2013). Surveys can be one important tool contributing to the definition and use of latent constructs that will develop along with theory and empirical work in the field.

Whereas the suggestion is made that survey instruments should be added to the repertoire of translation process research, it is done in full recognition of some limitations of these instruments. No single research tool is optimal for all measurements or for all studies, and surveys are not without their challenges and detractors. Reid (1990) writes in a narrative style to admit the challenges of designing, translating, pilot testing, and analyzing survey data.

Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 172–190.

Furthermore, Gorard (2010) presents a skeptical view of the use of Likert-type scales for measuring latent constructs. The present article highlights several similar aspects of survey design – philosophical, practical, and quantitative – that need to be considered in the process of creating and adapting scales. In particular, we have addressed and summarized issues related to validity, reliability, and quantitative analysis in order to provide guidance and examples. These three areas are by no means exhaustive, and more work is needed with regard to sampling theory, triangulation, item writing, and adaptation of existing instruments. However, an emphasis on validity, reliability, and quantitative analysis can serve as a foundation for more rigorous research that develops and employs surveys in Cognitive Translation Studies.

References

Agbo, A. A. (2010). Cronbach's alpha: Review of limitations and associated recommendations. *Journal of Psychology in Africa, 20*(2), 233–239. https://doi.org/10.1080/14330237.2010.10820371

Alves, F. (Ed.). (2003). *Triangulating translation: Perspectives in process-oriented research*. John Benjamins. https://doi.org/10.1075/btl.45

Alves, F. (2015). Translation process research at the interface: Paradigmatic, theoretical, and methodological issues in dialogue with cognitive science, expertise studies, and psycholinguistics. In A. Ferreira & J. W. Schwieter (Eds.), *Psycholinguistic and cognitive inquiries into translation and interpreting* (pp. 17–40). John Benjamins. https://doi.org/10.1075/btl.115.02alv

Alves, F., & Hurtado Albir, A. (2017). Evolution, challenges, and perspectives for research on cognitive aspects of translation. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 537–554). Wiley. https://doi.org/10.1002/9781119241485.ch29

Alvstad, C., Hild, A., & Tiselius, E. (2011). Methods and strategies of process research: Integrative approaches in translation studies. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research* (pp. 1–9). John Benjamins. https://doi.org/10.1075/btl.94

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Angelelli, C. V. (2004). *Revisiting the interpreter's role: A study of conference, court, and medical interpreters in Canada, Mexico, and the United States*. John Benjamins. https://doi.org/10.1075/btl.55

Angelelli, C. V., & Baer, B. J. (Eds.). (2016). *Researching translation and interpreting.* Routledge. https://doi.org/10.4324/9781315707280

Balling, L. W. (2008). A brief introduction to regression designs and mixed-effects modelling by a recent convert. *Copenhagen Studies in Language, 36*, 175–192.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide.* Routledge. https://doi.org/10.4324/9781315869780

Behr, D., & Shishido, K. (2016). The translation of measurement instruments for cross-cultural surveys. In C. Wolf, D. Joye, T. W. Smith, & Y.-C. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 269–287). SAGE. https://doi.org/10.4135/9781473957893.n19

Bernardini, S. (2001). Think-aloud protocols in translation research: Achievements, limits, future prospects. *Target, 13*(2), 241–263. https://doi.org/10.1075/target.13.2.03ber

Bolaños-Medina, A. (2014). Self-efficacy in translation. *Translation and Interpreting Studies, 9*(2), 197–218. https://doi.org/10.1075/tis.9.2.03bol

Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 172–190.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*(1), 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Bontempo, K., & Napier, J. (2011). Evaluating emotional stability as a predictor of interpreter competence and aptitude for interpreting. *Interpreting, 13*(1), 85–105. https://doi.org/10.1075/intp.13.1.06bon

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Boyle, G. J., Saklofske, D. H., & Matthews, G. (Eds.). (2015). *Measures of personality and social psychological constructs*. Academic Press.

Brown, J. D. (1996). *Testing in language programs.* Prentice Hall.

Buchanan, E. A., & Hvizdak, E. E. (2009). Online survey tools: Ethical and methodological concerns of human research ethics committees. *Journal of Empirical Research on Human Research Ethics, 4*(2), 37–48. https://doi.org/10.1525/jer.2009.4.2.37

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences, 3*(3), 106–116. https://doi.org/ 10.3844/jssp.2007.106.116

Carifio, J., & Perla, R. J. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education, 42*(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 9–24). Springer. https://doi.org/10.1007/978-3-319-07794-9_2

Chesterman, A. (2009). The name and nature of translator studies. *Hermes, 22*(42), 13–22. https://doi.org/10.7146/hjlcb.v22i42.96844

Choi, B. C. K., & Pak, A. W. P. (2005). A catalog of biases in questionnaires. *Preventing Chronic Disease, 2*(1), A13.

Clark, A. (1996). *Being there: Putting brain, body, and world together again.* MIT Press. https://doi.org/10.7551/mitpress/1552.001.0001

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*(2), 300–315. https://doi.org/10.1037/a0033805

Cortina, J. M. (1993). What is coefficient alpha?: An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and psychological assessment: Principles and applications.* Academic Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Csizér, K., & Dörnyei, Z. (2005). The internal structure of language learning motivation and its relationship with language choice and learning effort. *Modern Language Journal, 89*(1), 19–36. https://doi.org/10.1111/j.0026-7902.2005.00263.x

Dancette, J. (1997). Mapping meaning and comprehension in translation: Theoretical and experimental issues. In J. H. Danks, G. M. Shreve, S. B. Fountain, & M. K. McBeath (Eds.), *Cognitive processes in translation and interpreting* (pp. 77–103). SAGE.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity, 52*(4), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE.

Diamantopoulos, A. (Ed.). (2008). Formative indicators [Special issue]. *Journal of Business Research, 61*(12).

Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 172–190.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* MIT Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Flaskerud, J. H. (2012). Cultural bias and Likert-type scales revisited. *Mental Health Nursing, 33*(2), 130–132. https://doi.org/10.3109/01612840.2011.600510

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181–191. https://doi.org/10.1080/07481756.2003.11909741

Gorard, S. (2010). Measuring is more than assigning numbers. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The SAGE handbook of measurement.* SAGE. https://doi.org/10.4135/9781446268230.n20

Graybill, P., Aggas, J., Dean, R. K., Demers, S., Finigan, E. G., & Pollard Jr., R. Q. (2010). A community-participatory approach to adapting survey items for deaf individuals and American Sign Language. *Field Methods, 22*(4), 429–448. https://doi.org/10.1177/1525822X10379201

Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment*. Wiley.

Halverson, S. L. (2010). Cognitive translation studies: Developments in theory and method. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 349–369). John Benjamins. https://doi.org/10.1075/ata.xv.18hal

Halverson, S. L. (2019). 'Default' translation: A construct for cognitive translation and interpreting studies. *Translation, Cognition & Behavior, 2*(2), 187–210. https://doi.org/10.1075/tcb.00023.hal

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45*(1–3), 153–171. https://doi.org/10.1023/A:1006941729637

Han, C. (2018). Latent trait modeling of rater accuracy in formative peer assessment of English–Chinese consecutive interpreting. *Assessment & Evaluation in Higher Education, 43*(6), 979–994. https://doi.org/10.1080/02602938.2018.1424799

Harkness, J. A., Pennell, B.-E., & Shoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). Wiley. https://doi.org/10.1002/0471654728.ch22

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 115–140). Wiley. https://doi.org/10.1002/9780470609927.ch7

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning, 7*(6), 836–850. https://doi.org/10.1016/j.cptl.2015.08.001

House, J. (2013). Towards a new linguistic-cognitive orientation in translation studies. *Target, 25*(1), 46–60. https://doi.org/10.1075/target.25.1.05hou

Hubscher-Davidson, S. (2009). Personal diversity and diverse personalities in translation: A study of individual differences. *Perspectives, 17*(3), 175–192. https://doi.org/10.1080/09076760903249380

Hurtado Albir, A. (Ed.). (2017). *Researching translation competence by PACTE Group*. John Benjamins. https://doi.org/10.1075/btl.127

Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 172–190.

Hvelplund, K. T. (2014). Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data. *MonTI Special Issue—Minding Translation, Special Issue 1*, 201–223. https://doi.org/10.6035/monti.v0i0.292854

Hvelplund, K. T. (2017). Eye tracking in translation process research. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 248–264). Wiley. https://doi.org/10.1002/9781119241485.ch14

Jääskeläinen, R. (2010). Think-aloud protocol. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of translation studies* (Vol 1, pp. 371–373). John Benjamins. https://doi.org/10.1075/hts.1.thi1

Jääskeläinen, R. (2011). Studying the translation process. In K. Malmkjær & K. Windle (Eds.), *The Oxford handbook of translation studies* (pp. 123–135). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199239306.013.0010

Jakobsen, A. L. (2003). Effects of think aloud on translation speed, revision, and segmentation. In F. Alves (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 69–95). John Benjamins. https://doi.org/10.1075/btl.45.08jak

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1212–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jiménez Ivars, A., Pinazo Catalavud, D., & Ruiz i Forés, M. (2014). Self-efficacy and language proficiency in interpreter trainees. *The Interpreter and Translator Trainer, 8*(2), 167–182. https://doi.org/10.1080/1750399X.2014.908552

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixation to comprehension. *Psychological Review, 87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425–461. https://doi.org/10.3102/00346543064003425

Kankaraš, M., & Moors, G. (2010). Researching measurement equivalence in cross-cultural studies. *Psihologija, 43*(2), 121–136. https://doi.org/10.2298/PSI1002121K

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191–207. https://doi.org/10.1017/S000305540400108X

Kruger, J.-L., Soto Sanfiel, M. T., Doherty, S., & Ibrahim, R. (2016). Towards a cognitive audiovisual translatology: Subtitles and embodied cognition. In R. Muñoz Martín (Ed.), *Reembedding translation process research* (pp. 171–194). John Benjamins. https://doi.org/10.1075/btl.128.09kru

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. (2002). Cultural differences in response to a Likert scale. *Research in Nursing & Health, 25*(4), 295–306. https://doi.org/10.1002/nur.10041

Lee, S.-B. (2014). An interpreting self-efficacy (ISE) scale for undergraduate students majoring in consecutive interpreting: Construction and preliminary validation. *The Interpreter and Translator Trainer, 8*(2), 183–203. https://doi.org/10.1080/1750399X.2014.929372

Lee, S.-B. (2018). Exploring a relationship between students' interpreting self-efficacy and performance: Triangulating data on interpreter performance assessment. *The Interpreter and Translator Trainer, 12*(2), 166–187. https://doi.org/10.1080/1750399X.2017.1359763

Li, D. (2004). Trustworthiness of think-aloud protocols in the study of translation processes. *International Journal of Applied Linguistics, 14*(3), 301–313. https://doi.org/10.1111/j.1473-4192.2004.00067.x

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 1–55.

Litwin, M. (1995). *How to measure survey reliability and validity.* SAGE. https://doi.org/10.4135/9781483348957

Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 172–190.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*, 543–566. https://doi.org/10.1016/S0160-2896(03)00051-5

McGorry, S. Y. (2000). Measurement in a cross-cultural environment: Survey translation issues. *Qualitative Market Research, 3*(2), 74–81. https://doi.org/10.1108/13522750010322070

Mellinger, C. D. (2015). On the applicability of Internet-mediated research methods to investigate translators' cognitive behavior. *Translation & Interpreting, 7*(1), 59–71.

Mellinger, C. D., & Hanson, T. A. (2017). *Quantitative research methods in translation and interpreting studies*. Routledge. https://doi.org/10.4324/9781315647845

Mellinger, C. D., & Hanson, T. A. (2018). Interpreter traits and the relationship with technology and visibility. *Translation and Interpreting Studies, 13*(3), 366–392. https://doi.org/10.1075/tis.00021.mel

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111–121. https://doi.org/10.21500/20112084.857

Muñoz Martín, R. (2016). Of minds and men–computers and translators. *Poznań Studies in Contemporary Linguistics, 52*(2), 351–381. https://doi.org/10.1515/psicl-2016-0013

Muñoz Martín, R. (2017). Looking toward the future of cognitive translation studies. In J. W. Schwieter & A. Ferreira (Eds.), *The handbook of translation and cognition* (pp. 554–572). Wiley. https://doi.org/10.1002/9781119241485.ch30

Norman, G. (2010). Likert scales, levels of measurement and the 'laws' of statistics. *Advances in Health Sciences Education, 15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.

Oakes, M. P., & Ji, M. (Eds.). (2012). *Quantitative methods in corpus-based translation studies*. John Benjamins. https://doi.org/10.1075/scl.51

O'Brien, S. (2009). Eye tracking in translation process research: Methodological challenges and solutions. In I. M. Mees, F. Alves, & S. Göpferich (Eds.), *Methodology, technology, and innovation in translation process research: A tribute to Arnt Lykke Jakobsen* (pp. 251–266). Samfundslitteratur.

O'Brien, S. (2013). The borrowers: Researching the cognitive aspects of translation. *Target, 25*(1), 5–17. https://doi.org/10.1075/target.25.1.02obr

O'Brien, S., & Saldanha, G. (2014). *Research methodologies in translation studies.* Routledge. https://doi.org/10.4324/9781315760100

Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist, 16*(2), 56–69.

Pöchhacker, F. (2005). From operation to action: Process-orientation in interpreting studies. *Meta, 50*(2), 682–695. https://doi.org/10.7202/011011ar

Reid, J. (1990). The dirty laundry of ESL survey research. *TESOL Quarterly, 24*(2), 323–338. https://doi.org/10.2307/3586913

Risku, H. (2012). Cognitive approaches to translation. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0145

Risku, H. (2014). Translation process research as interaction research: From mental to socio-cognitive processes. *MonTI Special Issue—Minding Translation, Special Issue 1*, 331–353. https://doi.org/10.6035/MonTI.2014.ne1.11

Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 172–190.

Rönkkö, M., & Evermann, J. (2013). A critical examination of common beliefs about partial least squares path modeling. *Organizational Research Methods, 16*(3), 155–174. https://doi.org/10.1177/1094428112474693

Rubin, R. B., Rubin, A. M., Graham, E. E., Perse, E. M., & Seibold, D. R. (2011). *Communication research measures II: A sourcebook*. Routledge. https://doi.org/10.4324/9780203871539

Schaeffer, M., Huepe, D., Hansen-Schirra, S., Hofmann, S., Muñoz, E., Kogan, B., Herrera, E., Ibáñez, A., & García, A. (2020). The Translation and Interpreting Competence Questionnaire: An online tool for research on translators and interpreters. *Perspectives, 28*(1), 90–108. https://doi.org/10.1080/0907676X.2019.1629468

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*(2), 277–287. https://doi.org/10.1002/acp.1340

Shreve, G. M., & Angelone, E. (2010a). *Translation and cognition*. John Benjamins. https://doi.org/10.1075/ata.xv

Shreve, G. M., & Angelone, E. (2010b). Translation and cognition: Recent developments. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 1–13). John Benjamins. https://doi.org/10.1075/ata.xv.01shr

Shreve, G. M., Angelone, E., & Lacruz, I. (2018). Are expertise and translation competence really the same?: Psychological reality and the theoretical status of competence. In I. Lacruz & R. Jääskeläinen (Eds.), *Innovation and expansion in translation process research* (pp. 37–54). John Benjamins. https://doi.org/10.1075/ata.18.03shr

Smith, T. W. (2010). Survey across nations and cultures. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 733–763). Emerald.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*(3), 217–222. https://doi.org/10.1207/S15327752JPA8003_01

Van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (2019). Latent variable models and networks: Statistical equivalence and testability. *Multivariate Behavioral Research*, 1–24. https://doi.org/10.1080/00273171.2019.1672515

Vogt, W. P., & Johnson, R. B. (2016). *The SAGE dictionary of statistics and methodology* (5th ed.). SAGE.

Warrens, M. J. (2015). On Cronbach's alpha as the mean of all split-half reliabilities. In R. Millsap, D. Bolt, L. van der Ark, & W. C. Wang (Eds.), *Quantitative psychology research* (pp. 293–300). Springer. https://doi.org/10.1007/978-3-319-07503-7_18

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice, 29*(3), 39–47. https://doi.org/10.1111/j.1745-3992.2010.00182.x

Wilcox, J. B., Howell, R. D., & Breivik, E. (2008). Questions about formative measurement. *Journal of Business Research, 61*(12), 1219–1228. https://doi.org/10.1016/j.jbusres.2008.01.010

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Zlomke, K. R. (2009). Psychometric properties of internet administered versions of Penn State Worry Questionnaire (PSWQ) and Depression, Anxiety, and Stress Scale (DASS). *Computers in Human Behavior, 25*, 841–843. https://doi.org/ 10.1016/j.chb.2008.06.003

[1] We adopt the term *Likert-type scale* throughout this paper to emphasize that most survey scales do not strictly follow Likert's initial formulation (1932). For instance, a Likert scale must be symmetrical in its responses and measure level of agreement, with an odd number of responses indicated by both an integer and a verbal label. Since many scales alter at least one of these characteristics, they are more properly identified as Likert-type scales.

Mellinger, C. D., & Hanson, T. A. (2020). Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 19, 172–190.

---

[2] For a more complete discussion of this debate from multiple perspectives, see the special issue of the *Journal of Business Research* (Diamantopoulos, 2008).

[3] Due to this article's topic of surveys, inter-rater reliability is not discussed, though it is a separate topic worthy of attention in any study that involves assessment or categorization by multiple judges (e.g., Han, 2018).

[4] There are certainly other errors that could be discussed, but for considerations of space, we have chosen three that are representative of issues that often appear in T&I scholarship.

[5] That is not to say that all parametric tests are appropriate in all cases. For instance, the use of Student's *t*-test in translation and interpreting studies should be scrutinized given the difficulty in meeting its strict assumptions. Instead, Welch's *t*-test is the preferable parametric test. For an extended discussion on this topic, see Mellinger and Hanson (2017).

[6] Many volumes provide more details of the theory and practice of factor analysis, including two approachable accounts by DeVellis (2017) and Bandalos (2018).