Methodological issues in cross-linguistic and multilingual advertising research

Bert Weijters

Stefano Puntoni

Hans Baumgartner

Bert Weijters (Ph.D., Ghent University) is assistant professor of market research, Department of Personnel Management, Work and Organizational Psychology, Ghent University, B-9000 Ghent, Belgium, +3292646296, bert.weijters@ugent.be.

Stefano Puntoni (Ph.D., London Business School) is professor of marketing, Rotterdam School of Management, Erasmus University, PO Box 1738, 3000 DR, Rotterdam, the Netherlands, +31104081184, spuntoni@rsm.nl.

Hans Baumgartner (Ph.D., Stanford University) is the Smeal Professor of Marketing, Smeal College of Business at The Pennsylvania State University, Department of Marketing, 482 Business Building, University Park, PA 16802, 814 863 3559, hansbaumgartner@psu.edu.

Methodological issues in cross-linguistic and multilingual advertising research

This paper discusses methodological issues related to language in advertising research. We introduce a framework that distinguishes between cross-linguistic research settings, where several languages are used in the study and different samples of respondents are studied in their own language, and multilingual research settings, where only a single language is used and multilingual respondents are studied either in their native or non-native language. We review key principles that govern cross-linguistic and multilingual effects in advertising research to formulate guidelines for research design and data analysis. In the cross-linguistic context, these principles address non-uniform cross-linguistic differences in responses (related to non-equivalence of individual questionnaire items) versus uniform response effects (related to non-equivalence of verbal response category labels). In the multilingual context, we bring together evidence that shows how—even when comprehension is not a problem—stimuli, questions and response categories may be processed differently in respondents' native versus non-native language.

Key words: Language; survey research; multilingual research; translation; measurement equivalence.

The increasing interconnection of a variety of social, economic, and technological processes across geographic areas—a process usually referred to as globalization—is one of the defining trends of our time. Globalization is also transforming important aspects of the process and content of advertising research. For example, a large body of advertising research examines the impact of the continuing rise of global brands and global consumer culture on how people around the world consume advertising (Akaka and Alden 2010, Okazaki, Mueller and Diehl 2013). An important way in which globalization is changing advertising research, both in academia and in practice, is by increasing the relevance of linguistic issues, the topic of this paper. For example, advertising researchers often want to understand if and how responses to a particular message or type of appeal differ across consumers in different countries, who tend to speak different languages. In addition, even when advertising researchers are not interested in international aspects, they often sample groups of consumers who speak more than one language. In today's multicultural societies, it is no longer safe to assume that the language used in advertising stimuli and measurement instruments will be the native language of all respondents. Therefore, it is becoming ever more common for researchers to work with cross-linguistic samples (consisting of multiple groups of respondents who use different languages) and/or multilingual respondents (who speak multiple languages and use one of these languages).

In contrast to (other) cultural differences, which are often the explicit focus of advertising research, language effects are often not the primary focus of the research and are, rather, a challenge to be dealt with. Typically, in cross-cultural research, the goal is to discover and explain differences between groups of respondents. Instead, when advertising researchers face linguistic issues, the goal is usually to be able to generalize across heterogeneous groups of respondents. This is not to mean, of course, that advertising researchers never explicitly focus on linguistic issues (Luna and Peracchio 2001, Luna and

Peracchio 2005, Puntoni, De Langhe and Van Osselaer 2009, Tavassoli and Han 2002). However, the number of advertising researchers who treat language as a potential biasing influence is much larger than that for whom language is the focal issue. Despite the increasing relevance of linguistic issues in advertising research, and marketing more generally, researchers lack guidance about effective research design and data analysis. Therefore, this paper reviews and organizes recent research on methodological issues in cross-linguistic and multilingual contexts.

Advertising research has traditionally been led by North American scholars, for whom linguistic issues were often not highly relevant due to the monolingual make-up (at least until recently) of most of North America. For example, bilingual or multilingual consumers were traditionally considered a niche (e.g., they may have been of interest to researchers focusing on ethnic minorities) and linguistic considerations were often not an important aspect of research design. However, the majority of consumers around the world speak more than one language and multilingual issues are now relevant to many, if not most, research settings. Also, many brands and organizations operate across multiple countries and languages. It is therefore important that advertising researchers understand how research findings depend on the language in which the research is conducted and the linguistic make-up of the sample. The main goal of this paper is to develop a list of general principles and to formulate guidelines aimed at helping researchers understand how to best conduct advertising research with multilingual participants and/or across multiple languages. To organize these principles, we first propose a typology of relevant research contexts.

Our focus is on research that involves questionnaires in experiments and/or surveys. We concentrate on the questionnaires used to collect responses from participants, although some of the principles we discuss are applicable to the design of stimuli such as advertising messages as well. Linguistic issues are relevant to other research methods, such as content analysis (Lerman and Callow 2004, Okazaki and Mueller 2007) and qualitative methods (Eckhardt, Dholakia and Belk 2013). We decided to focus on research employing questionnaires because this context is relevant to the majority of advertising researchers and because space limitations prevented covering broader ground.

A TYPOLOGY OF LINGUISTIC RESEARCH CONTEXTS

To introduce our typology, consider how globalization is impacting diversity across and within countries. Globalization affects diversity in two opposing ways. First, globalization leads to a *decrease in diversity between countries*. Whereas only a few decades ago, people in different countries lived very different lives, we can now observe a remarkable cultural convergence. For example, to a large extent, teenagers today listen to the same music, dress in the same way, and play the same games regardless of whether they live in New York, Rotterdam, or Beijing. One major consequence of this process of cultural homogenization is that companies increasingly treat the world as one market (Alden, Steenkamp and Batra 1999, Levitt 1983). As a result, advertising research increasingly features participants from multiple countries who answer questions about the same advertisements or brands in their own native language. Researchers must thus establish the comparability of the results.

Second, globalization leads to an *increase in diversity within countries*. Contemporary societies are vastly more diverse than they used to be, as can be easily noticed by taking a walk in the centre of Rotterdam or in other cities across North America and Western Europe. This increase in diversity within countries means that advertising researchers must often deal with participant pools composed of native speakers of different languages. Even when it is possible to conduct advertising research using a single language for constructing stimuli and collecting measurements, for some of the participants the language used will not be their native language.

In sum, a paradox of globalization is that it both increases and decreases diversity. On the one hand, you can now eat sushi or Indian food in a sleepy Italian town. On the other hand, these restaurants look pretty much the same as those found in similarly sleepy towns in other countries. These opposing trends raise new issues for advertising researchers. Our typology of research settings builds on these broad trends to identify specific research contexts relevant to language-specific methodological issues.

Table 1 provides an overview of possible situations with regard to language in research situations. The two rows refer to whether the stimuli and/or research instruments used in a study are (1) all in the same language or (2) in multiple languages. This dimension speaks to the decreasing diversity between countries and the associated increase in cross-border advertising research. The two columns refer to whether (A) the language used in the research is the native language of all respondents or (B) at least some participants work with a non-native language. Following standard convention in linguistics, we refer to respondents' native language as L1 and to respondents' non-native language as L2.

PLACE TABLE 1 ABOUT HERE

Cell A1 was historically the default situation: researchers used one language for all materials (stimuli and measurements) and participants were all native speakers of this language. This situation does not involve multilingual or cross-linguistic issues and our interest lies therefore in situations other than this default. Cell A2 describes the situation where respondents belong to different language groups (e.g., multi-country samples are used) and each group completes the study in their own native language ("Context I"). Cell B1 describes the situation where the research is conducted using only one language, but this language is some respondents' L1 and other respondents' L2 ("Context II"). We will refer to Context I as *cross-linguistic research* and to Context II as *multilingual research*. Cell B2 describes the combination of the previous two situations: multiple languages are used in the research and each of these languages is L1 for some respondents and L2 for others ("Context III"). We will not discuss Context III separately, since insights from both Context I and Context II apply here. Finally, even though it is possible to conceive of an additional situation in which all respondents answer using L2, we do not consider this context in the present paper because of its rarity and lack of relevant research.

CONTEXT I: CROSS-LINGUISTIC RESEARCH

Examples of Relevant Research Contexts

The use of cross-linguistic samples in advertising research serves multiple purposes. Perhaps most importantly, advertising researchers testing theories across cultures and language groups grapple with questions of generalizability: does a theory or model initially developed and validated in a specific culture hold in other cultures as well (Dawar and Parker 1994). Advertising researchers often hope to establish 'strong theories' that are generally valid and are not limited to a specific context (Laczniak 2015). Cross-linguistic validation of advertising stimuli and measurement scales (such as recipients' beliefs about and attitudes toward advertising) form an important prerequisite for cross-cultural and cross-linguistic validation of advertising theories (Andrews, Durvasula and Netemeyer 1994).

Once cross-linguistically validated, the scales can then also be used in applied advertising research, where standardization (rather than theoretical generalizability) poses a key challenge. Traditionally, advertising for products sold in multiple countries used standardization most often in strategy, less often in execution, and least often in language

(Duncan and Ramaprasad 1995). However, for global brands it has become increasingly common and feasible to standardize execution and even content, not merely strategy (Taylor and Okazaki 2015). For cross-national standardization to be effective, successfully creating cross-linguistically equivalent advertising content and measurement scales becomes ever more important. Thus, equivalence in advertising meaning and measurement scales across different languages is a recurrent issue in advertising practice. In sum, generalizability and standardization are key issues in cross-linguistic research.

Differentiation is a second key rationale for using cross-linguistic samples. With differentiation we refer to research that aims to identify differences (rather than similarities or universals) in advertising-related variables or multivariate relations of interest across different groups of consumers, where the grouping is defined by national culture (Okazaki and Mueller 2007). An important challenge in this type of research is that national culture is usually (though not necessarily) confounded with language (Harkness et al. 2010). This poses additional challenges for research design, data collection and data analysis, because group differences may be attributed to cultural differences when in fact they are due to linguistically non-equivalent stimuli and/or measures. To avoid such misattributions, advertising researchers commonly employ translation/back-translation procedures (De Meulenaer, De Pelsmacker and Dens 2015, Minton et al. 2012, Rose, Bush and Kahle 1998).

Literature review and principles

Although cross-linguistic advertising research may often involve challenges related to data collection standardization (e.g., training interviewers) and sampling equivalence (e.g., obtaining matched samples in different countries), our focus here is on language-related issues that arise during stimulus design, instrument development, and data analysis. In line with common practice in advertising research, we focus on studies where researchers collect data from different language groups using equivalent stimuli, questions and response formats (Craig and Douglas 2005). In what follows, the focus will be on questions and response scales. Questions are, in essence, stimuli as well, and many of the insights concerning questions can also be applied to other stimuli, including advertising copy. For instance, back-translation procedures that have been developed mainly in questionnaire design are also applied to advertising copy translation (De Meulenaer, De Pelsmacker and Dens 2015).

Instrument Design

Equivalent questionnaires (including instructions, questions, and response options) are traditionally obtained through the use of translation/back-translation procedures (Brislin 1970). With this procedure, an instrument is designed in a source language, translated to a target language by a bilingual native speaker of the target language, and the result is translated back to the source language by a bilingual native speaker of the source language. Based on a comparison of the initial and the back-translated instrument, incidental differences are resolved. If need be, additional iterations are run. Back-translation is a common approach to help identify translation problems in advertising research, but it does not necessarily ensure equivalence in meaning in each language (Douglas and Craig 2007, Okazaki and Mueller 2007).

Instead, Douglas and Craig (2007) propose collaborative and iterative translation as an alternative approach. In this approach, a committee first establishes the equivalence of key concepts to be assessed in the questionnaire. Next, two independent translators, working in parallel, translate the instrument into the target language. These translations are then pre-tested and iteratively revised until satisfactory versions are attained. Importantly, the whole process involves a team or committee that brings together the necessary skills and knowledge related to questionnaire design and the languages involved.

Many of the problems related to translation can optimally be addressed in an early stage of questionnaire development, by adopting, adapting and/or creating items that are easily translatable to the languages of interest. This requires researchers to step out of their own reference frame and take a decentered approach, as opposed to what has been called ‘research imperialism or safari research’ (Smith 2004). Specifically, decentering refers to the simultaneous development of the same instrument in several languages (or cultures) from the very start (instead of designing the instrument in a source language). A decentered approach typically calls for cooperation of researchers involved in the research project who have a background in each of the languages (Douglas and Craig 2007, Smith 2004, Van de Vijver and Leung 1997).

Principle I.1: Decentering reduces source language dependence and thus facilitates equivalence in translation.

Brislin (1986) offers some readily implementable guidelines to make items easier to translate, summarized in Table 2.

PLACE TABLE 2 ABOUT HERE

Typical questionnaire items have two parts: the stem of the item presenting the substance and stimulus, and the response scale used for recording the answers. Translation issues occur for both parts, but some issues and related solutions are specific to each part.

Question design

Literal translations of a word may not map to exactly the same concept in different languages. For instance, the Spanish word ‘educacion’ may have different connotations (including associations with socially correct behavior) than the English word ‘education’ (which is more strongly related to the academic domain) (Greenfield 1997). Similarly, a commonly used verb in attitude measurement such as ‘(to) like’ may not have equivalent counterparts in some other languages, including the French alternative ‘aime(r)’ (which could be translated to ‘like’ or ‘love’, thus creating ambiguity) (http://visual.ly/facebook-translated-around-world).

The use of multiple indicators to measure latent constructs gets around the problem that an item will seldom if ever perfectly coincide with the construct it aims to measure. As such, it also enables researchers to accommodate cross-linguistic variations in the meaning of words and questions. It does so by averaging out such variations and by making it possible to detect systematic deviations in the way an indicator relates to the construct it aims to tap (Steenkamp and Baumgartner 1998). An important caveat in this context is the repeated use of a central word or concept in several or even all items that make up a measurement scale. For instance, if a satisfaction scale uses the word ‘satisfaction’ in each item, incidental differences in meaning across languages will permeate responses to each item (Smith 2004). To address this problem, Smith (2004) suggests that three linguistically distinct measures of the same construct are desirable, and the indicators should not be minor variations of the same underlying question stem, such that group-differences can be triangulated. Absent such triangulation, one can never be sure whether cross-linguistic differences are not an artefact of item translation non-equivalence. As a matter of fact, for a latent construct to have validity, using more (different) indicators is generally better, since the indicators are supposed to be a representative sample of a hypothetical population of possible indicators (Marsh et al. 2013).

Principle I.2: Wording key concepts in multiple distinct ways makes it possible to triangulate cross-linguistic variations in meaning.

Steenkamp and Baumgartner (1998) proposed a systematic procedure to examine configural, scalar, and metric measurement invariance, by performing multi-group confirmatory factor analysis. First, configural invariance implies that specific indicators relate to the same factor across groups (i.e., the factor structure is equivalent across groups). Second, metric invariance implies that the relationship between specific indicators and their underlying factor is the same across groups (i.e., the item loadings are equal). Third, scalar invariance implies that, for a given factor score, the means of the indicators are the same across groups (i.e., the item intercepts are equal). The Steenkamp and Baumgartner (1998) approach focuses on cross-national comparisons, but is directly applicable to cross-linguistic comparisons (which often coincide). When making cross-linguistic comparisons, configural and metric invariance are required in order to meaningfully compare variances and relations between variables (covariances, regression or path coefficients). For comparisons of means, scalar invariance is required as well.

Principle I.3a: Cross-linguistic (co)variance comparisons require configural and metric invariance.

Principle I.3b: Cross-linguistic mean comparisons require configural, metric and scalar invariance.

It has been repeatedly suggested in the past that measurement invariance testing has been underutilized in cross-national business research (He, Merz and Alden 2008, Hult et al. 2008), including advertising research (Okazaki and Mueller 2007). More recently, measurement invariance testing seems to have become more common. If researchers neglect to test for measurement invariance, this might be due to a lack of understanding of the approach among some researchers, as well as a supposed limitation of invariance testing: If measurement invariance is rejected, it may be perceived as putting an end to the analysis as no meaningful conclusions can be drawn from the data. Clearly, this is not a motivating prospect for most researchers. This is only partially true, however, and at least three caveats should be mentioned.

First, if full measurement invariance is rejected, partial measurement invariance may still hold. For cross-linguistic comparisons to be meaningful for a given construct, at least two indicators (per construct) need to exhibit invariance (Steenkamp and Baumgartner 1998).

Second, recent developments using Bayesian modeling approaches allow for some (limited) amount of across-group variation in measurement parameters. These advances include hierarchical Item Response Theory (IRT) models (De Jong and Steenkamp 2010, De Jong, Steenkamp and Fox 2007), as well as Bayesian Structural Equation Modeling (Muthén and Asparouhov 2012, Muthén and Asparouhov 2013). For now, the diffusion of these more advanced methods may be hampered by their analytical sophistication (Baumgartner and Weijters 2015).

Third, even within a standard CFA framework, measurement invariance testing can be approached from a modeling perspective, rather than a strict null hypothesis significance testing perspective. The latter approach might work if the model is simple and there are few groups to be compared. But for more complex models and a larger number of groups, a modeling perspective is probably more meaningful.

In cross-linguistic Structural Equation Modelling, the cross-linguistic equivalence of specific parameter estimates can be evaluated by means of nested model tests. A statistical test of the invariance hypothesis (the χ^2 difference test for nested models) examines whether the difference between two parameters is *exactly* zero (i.e., the two parameters are identical). This is problematic in two ways: first, identity is an unrealistic and probably unnecessary ideal, and second, statistical null hypothesis testing can establish non-identity but not identity, and is therefore necessarily inconclusive (Nickerson 2000). Instead, for more complex models and a larger number of groups the focus should be on model optimization, where one assesses whether corresponding model parameters are sufficiently invariant across different language groups. Thus, invariance evaluation should typically focus more on practical fit indices such as BIC, CAIC and RMSEA (Baumgartner and Steenkamp 2006, Steenkamp and Baumgartner 1998). These indices trade off closeness of fit and model parsimony. RMSEA is a fit index for which confidence intervals can be constructed (MacCallum, Browne and Sugawara 1996). A major advantage of BIC and CAIC is that they aid in selecting an optimal model (Williams and Holahan 1994). This allows the researcher to identify an optimal model, rather than a muddle of possible models that are all significantly but negligibly worse than an unconstrained (and usually rather non-parsimonious) baseline model.

> *Principle I.4: For more complex models and a larger number of groups, measurement invariance testing requires a modeling perspective rather than a null hypothesis significance testing perspective.*

Response Scale Design

A key difference between the item stem and the response scale is that item non-equivalence can partly be accommodated through the use of multiple items per construct. In contrast, response scales are typically used for multiple (if not most) items in the same instrument (Podsakoff et al. 2003, Rindfleisch et al. 2008), which makes for a potentially more systematic impact of non-equivalent translations. For instance, it is quite common to measure multiple constructs by means of five-point Likert-type items with response categories labeled ‘strongly disagree,’ ‘disagree,’ ‘neither agree nor disagree,’ ‘agree,’ ‘strongly agree’.

While the use of such standard verbal response category labels for many items in the same questionnaire is convenient for both the researcher and the respondent, it also carries risks. Most importantly, if the verbal anchors are translated to another language and lead to differential scale usage in different languages, the resulting bias will be uniform. That is, responses to all items that use the same format will be similarly affected (Podsakoff, MacKenzie, Lee and Podsakoff 2003, Rindfleisch, Malter, Ganesan and Moorman 2008). Although many researchers are not aware of this, such uniform response bias cannot be detected by standard measurement invariance testing, as invariance testing only detects item-specific biases. The reason is that, to the extent that the bias is uniform across items, it will consistently inflate or deflate all measurement model parameters (i.e., item intercepts and/or loadings). Standard Confirmatory Factor Analysis models cannot distinguish between a uniform change in intercepts and a change in factor mean, nor can they distinguish between a uniform change in factor loadings and a change in the factor variance (Little 1997, Little 2000, Weijters, Schillewaert and Geuens 2008). If, on the other hand, only one item is affected by a non-equivalent translation, this will typically show up as a non-equivalent intercept or factor loading for this item (Steenkamp and Baumgartner 1998).

Principle I.5a: In multi-item instruments, item-specific non-equivalence leads to non-uniform measurement bias (metric and/or scalar non-invariance).

Principle I.5b: In multi-item instruments, response scale label non-equivalence leads to uniform response bias, which is undetectable with standard measurement invariance testing.

Translating verbal anchors can be surprisingly challenging because languages often offer different possibilities in terms of syntax and semantics (Harkness, Pennell and Schoua-Glusberg 2004). The methodological literature offers three possible approaches that aim to maximize cross-linguistic equivalence of scale anchors (Douglas and Craig 2007, Smith 2004). First, one can use nonverbal scales, including visual or numerical analogs. This proposal has its own problems, however, including possible linguistic differences in the processing and interpretation of visual cues (Tavassoli and Han 2002) and numerical scales (Göbel, Shaki and Fischer 2011). But most importantly, the instructions that assign meaning to the numerical or visual scale still need to be translated, so this approach does not solve the basic problem.

Second, one can use dichotomous scales with responses such as yes/no or agree/disagree, which may be more likely to be equivalent across languages. This seems a viable solution for scales for which this format is useful and meaningful, but it leads to a loss of information (Cox III 1980, Garner 1960, Green and Rao 1970), and respondents may not like it that they cannot express gradations of liking or opinion (Preston and Colman 2000). Finally, the assumption that dichotomies are simple and equivalent across societies has been called into question. For example, “agree/disagree” in English can be translated into German in various ways, which may all lead to different measurement consequences (Harkness, Pennell and Schoua-Glusberg 2004).

A third solution is to calibrate the response scale to obtain equivalent verbal scales. Douglas and Craig (2007) make a distinction between endpoint labeled and fully labeled scale formats. Endpoint labeled formats have the advantage of simplicity, and only two anchors need translating. But survey methods research has shown that if only the endpoints are labeled, the non-labeled categories may be hard to interpret for some respondents (Arce-Ferrer 2006). In contrast, when all scale positions are fully labeled, all categories are more or less equally clear to respondents and this leads to more substantively consistent responses (Cabooter et al. 2010, Moors, Kieruj and Vermunt 2014). It is therefore preferable to work with fully labeled scale formats. This then poses the challenge of coming up with multiple equivalent response category labels.

The verbal label used for response categories affects the likelihood of respondents selecting the corresponding scale position. Two mechanisms have been proposed to explain this phenomenon: the intensity account and the familiarity account (Weijters, Geuens and Baumgartner 2013). According to the intensity account, verbal labels that denote greater intensity imply a more extreme position on the attribute to be measured (Smith 2004) and are consequently less likely to be selected (de Langhe et al. 2011). For instance, if a Likert-type scale is used, the endpoint labels ‘strongly disagree’ and ‘strongly agree’ suggest more intense (dis)agreement than the endpoint labels ‘disagree’ and ‘agree’. Because respondents are less likely to endorse more extreme positions, using extremely worded endpoints can lead to lower endorsement rates of the endpoints of the response scale.

Smith (2003) reviews three methods for having respondents assess category label intensity: ranking, rating and magnitude-estimation techniques. The rating method, where respondents rate each verbal anchor on a numerical scale, is suggested to be most useful as it is not very demanding but still measures absolute strength and the distance between terms, thus facilitating the design of equal-interval scales. Evidence suggests that the technique is robust and reliable (Smith 2003). Weijters, Geuens and Baumgartner (2013) also find that intensity measures based on direct ratings show convergent validity when compared to intensity measures based on paired comparisons.

According to the familiarity account, response categories are endorsed more frequently if the labels are more common in day-to-day language (Weijters, Geuens and Baumgartner 2013). For example, if ‘entirely (dis)agree’ is less common than ‘strongly (dis)agree’, the former will be endorsed less than the latter. Weijters, Geuens and Baumgartner (2013) found that differences in extreme responding between two languages disappeared when equally familiar category labels were used in both languages. Thus, researchers need to carefully select response category labels that are matched in terms of familiarity across languages.

Principle I.6a: More intense translations of response scale category labels are endorsed less frequently.
Principle I.6b: Less familiar translations of response scale category labels are endorsed less frequently.

The familiarity of alternative verbal category labels can be evaluated in different ways. The approaches discussed for measuring intensity (ranking, rating, magnitude-estimation and pairwise comparison) can be adapted to have respondents assess familiarity. Two additional approaches, which were demonstrated to show convergent validity with direct ratings and pairwise comparisons, are available as well (Weijters, Geuens and Baumgartner 2013). First, verbal category labels can be used as stimuli in a lexical decision task, with faster responses indicating greater familiarity. Second, the number of search engine hits can be used as a proxy for familiarity (Weijters, Geuens and Baumgartner 2013). This approach is particularly efficient and works as follows, using as an example a situation in which a positive endpoint label is needed for a 5-point Likert scale in a survey conducted in English and French (see Table 3): (1) formulate several English and French endpoint labels; (2) look up the number of verbatim search engine hits (e.g., in Google) for each expression in the specific language of interest; (3) divide the count for each endpoint label by the number of search engine hits for the label without the modifier (e.g., completely) and take the natural logarithm of the ratio; (4) select a label pair that has similar meaning and a low discrepancy in familiarity scores (i.e., adjusted search engine hits) across languages. As a general rule, if labels need to be defined in more than two languages, selecting the label with minimal cross-linguistic variation in relative familiarity scores represents the best option. It is also important to note that specific expressions can at times vary in frequency of use even between countries that share the same language. For example, it may not be appropriate to assume equivalence of category labels between Brazilian and Portuguese samples answering the same survey in Portuguese.

PLACE TABLE 3 ABOUT HERE

Calibration of verbal anchors in terms of intensity and familiarity should ideally be done at an early stage of questionnaire design. If existing research is available in which verbal anchors have been calibrated in the languages of interest, researchers can use this information to select appropriate calibrated verbal labels (under the assumption that the intended samples are linguistically comparable). Unfortunately, verbal anchors can simultaneously show cross-linguistic intensity equivalence and familiarity non-equivalence or vice versa. If a tradeoff needs to be made between intensity versus familiarity equivalence, evidence suggests that familiarity equivalence should get priority (Weijters, Geuens and Baumgartner 2013).

Principle I.7a: Rather than being literal translations, verbal response category labels need to be cross-linguistically calibrated and should ideally be equally familiar and equally intense.

Principle I.7b: When selecting verbal response scale category labels for cross-linguistic use, familiarity matching is more important than intensity matching.

So far, we have focused on calibrating verbal anchors for cross-linguistic equivalence (during questionnaire design). But this may not always be possible, given the often conflicting implications posed by equivalence based on intensity versus familiarity, as well as other considerations such as closeness of translation. Therefore, verbal anchor calibration in the preparatory research phase may often need to be complemented with a post hoc approach in which responses are reweighted in a way that corrects for cross-linguistic non-equivalence during data analysis.

Principle I.8: If verbal response scale category labels are non-equivalent in different languages, responses should be differentially weighted to take into account differences in endorsement likelihood that are not based on item content.

The calibrated sigma calibrated sigma method is designed to eliminate the non-comparability of responses across different groups of respondents in general and different language groups more specifically (Weijters, Geuens and Baumgartner 2011, Weijters, Baumgartner and Geuens 2016). The method uses information derived from a carefully-selected set of control variables to reweight the responses to substantive items in a group-specific way. Instead of assigning the same consecutive integers to the scale positions in all groups (e.g., in the case of a 5-point scale, ‘strongly disagree’ is usually coded as 1, ‘disagree’ as 2, ‘neither agree nor disagree’ as 3, ‘agree’ as 4, and ‘strongly agree’ as 5), the response categories are converted to numerical values in a language-specific way. Specifically, the numbers assigned to the response categories are based on the distribution of responses to control items which serve no purpose other than assessing the content-free endorsement frequencies of the response categories in different groups (i.e., these calibration items are not used for substantive purposes). Thus, instead of arbitrarily assuming an equal-interval scale, the scale scores are chosen based on how the different groups respond to a set of items that share no obvious common content (e.g., ‘strongly agree’ might be coded as 5 in English, whereas ‘tout à fait d’accord’ is coded as 4.5 in French, corresponding to the different endorsement rates of the fifth option in response to the control items across the two languages).

Response patterns are evaluated by including control (calibration) items in the questionnaire with the sole purpose of measuring response patterns. This calibration is based on the idea that response patterns observed across items that are highly heterogeneous in content can be assumed to be due to method bias (Baumgartner and Steenkamp 2001). In the measurement literature, scales have been developed that contain such heterogeneous items, most prominently a 16-item scale developed by Greenleaf (1992). Alternatively, researchers can randomly sample control items from various item inventories (De Beuckelaer, Weijters and Rutten 2010).

Table 4 provides a brief worked example of how to compute calibrated sigma values based on hypothetical responses to the 16 five-point Greenleaf items by two matched samples of respondents using two different languages. In step 1, the mean frequency with which each scale category is chosen across the control items has to be computed. In step 2, the frequencies are recoded as proportions. In step 3, the cumulative proportions are computed.

Step 4 involves computing the midpoint of each category proportion. Finally, in step 5, these midpoint proportions per category are transformed into calibrated sigma codes (which correspond to standardized z-scores). The sigma values obtained for the two languages can then be used to recode the responses to the substantive items of respondents in groups A and B, respectively. In group A, for instance, a ‘strongly disagree’ response would be coded as -1.96.

PLACE TABLE 4 ABOUT HERE

CONTEXT II: MULTILINGUAL RESEARCH

Examples of Relevant Research Contexts

Oftentimes, researchers have an incentive to use the same language across participants who are not equally proficient in the target language. There are multiple reasons for this, including cost (no need to translate/adapt the survey), general efficiency, and the fact that it is often hard to tell a priori what the native language of each participant will be (e.g., online settings or multicultural cities). Underlying the common choice of standardizing language in a given study is the assumption that, as long as participants are sufficiently proficient in the selected language to understand the materials, the actual language used in the materials does not matter. For example, many cross-cultural advertising studies on self-construal compare answers to materials in English provided by participants in an individualistic culture (e.g., the USA) with the answers provided by participants in a collectivistic culture (e.g., Singapore) to the same English-language materials. In such a design, there is a confound between self-construal and language because for half of the respondents the target language is L1, whereas for the remaining half it is L2. This confound is usually ignored. In this section of the current paper, we challenge the assumption that language does not matter, even if we can assume that comprehension is not a serious issue.

Literature review and principles

Several studies in advertising research have focused on the mixed use of multiple languages in the same ad. Luna and Peracchio (2005) and Bishop and Peterson (2010) investigate how bilingual consumers interpret and evaluate ads that make combined use of English and Spanish. Ahn and Ferle (2008) explore how foreign and local languages influence recall and recognition for brand name and body copy messages in a South Korean advertising context. Kubat and Swaminathan (2015) study the effects of using English in combination with Spanish, Chinese or Hindi in ads on brand liking. This line of research not only has implications for its focal study object (i.e., multilingual advertising effects), but can also inform multilingual advertising research methodology by pointing toward the differential impact of using L1 versus L2 in stimuli (which, as pointed out previously, include questionnaire items) and response scales. Additional research has explicitly focused on multilingual effects in the area of response scale format (de Langhe, Puntoni, Fernandes and van Osselaer 2011). Below, we review several important classes of effects.

Comprehension

Almost self-evidently, if research participants cannot properly understand the advertising stimuli or the measurement instruments, then responses cannot be valid. But even among bilinguals who are proficient in L2, subtle differences in comprehension may occur (Luna 2011). Comprehension of advertising messages is not all or nothing; there are different levels of comprehension (Mick 1992) and whether information is presented in L1 or L2 can change comprehension in subtle ways that may be hard to anticipate. Although the impact of language on comprehension processes is moderated by a variety of factors, such as visuals and motivation (Luna and Peracchio 2001, Wyer 2002), it should not be controversial to make a general statement about the impact of language on comprehension and memory.

Principle II.1: L1 texts lead to greater comprehension and memory than L2 texts.

Even if a respondent’s language proficiency is sufficiently high to understand what is being asked in a study, there are bound to be differences in the extent to which respondents answering in L1 versus L2 feel confident about their interpretation of the textual information provided in either the stimuli or the measurement instruments. Harzing (2006) investigated whether completing a questionnaire in L1 or in English (as L2) influenced stylistic responding and whether English language competence had an impact. The English questionnaire led to lower extreme response style (ERS) and higher midpoint response style (MRS) than the L1 questionnaire. In addition, for the English questionnaire, self-rated ability to understand written English was positively related to ERS and negatively related to MRS. This suggests that language competence makes respondents more willing to respond more extremely. Although more research in this area is needed, based on existing findings, it seems possible to conclude that respondents should in general be more confident in their answers when completing a survey in L1 than in L2. This confidence difference may in turn imply a number of additional consequences, which at this point remain speculative. For example, people may express greater preferences for simpler messages or safer options when faced with L2 stimuli, especially when justifiability is important

Principle II.2: Respondents will have more confidence in their answers in L1 than in L2.

Emotional intensity

Research across fields as diverse as advertising (Puntoni, De Langhe and Van Osselaer 2009), psycholinguistics (Harris, Aycicegi and Gleason 2003), and psychoanalysis (Javier 1989) shows that messages expressed in consumers' native language tend to be perceived as more emotional than messages expressed in their second language (Pavlenko 2007). The effect of language on emotionality holds even when one controls for language-specific stereotypes and associations. For example, Puntoni et al. (2009) asked Dutch-French Belgian bilinguals to read a series of advertising slogans, some in Dutch and some in French. For half of the volunteers, the native language was French and for half it was Dutch. Regardless of whether their native language was French or Dutch, native language slogans were perceived as more emotional than second language slogans. The emotional advantage of L1 words also holds when controlling for differences in comprehension. For example, Puntoni et al. (2009) document the effect in the case of simple single words, and even in the case of cognates (words that are almost identical in L1 and L2).

What, then, explains this difference in the emotional intensity of words? Everyday language use impacts perceived word emotionality by associating lexical representations with autobiographical memories (Harris, Gleason and Aycicegi 2006, Pavlenko 2007, Puntoni, De Langhe and Van Osselaer 2009). Thus, reading or hearing a word (unconsciously) triggers personal memories of situations in which that word played a role. These personal memories evoke emotions, making the words in L1 feel more emotional than words in L2. Two alternative accounts have been proposed for this process of association, and both are likely to contribute to the effect of language on emotionality. First, the emotional advantage of consumers' native language depends on the number of personal experiences with a language. Because consumers usually have more personal memories with words in their native language than in their second language, messages in their native language tend to be perceived as more emotional (Puntoni, De Langhe and Van Osselaer 2009). Second, the context of language learning tends to differ between L1 and L2. L1 is learned earlier in life in highly emotional contexts (primarily via interactions with primary care takers), whereas L2 is typically learned later in an instructional context, for most people secondary school (Dewaele 2004). This difference in learning contexts results in a difference between the emotional intensity of words in L1 and L2 (Altarriba 2003, Harris, Gleason and Aycicegi 2006). One important consequence of this language difference is that advertising stimuli tend to generate more intense emotions when they are expressed in L1 than in L2.

Principle II.3: L1 ads generate higher ratings of emotional intensity than L2 ads.

Interestingly, this principle is reversed when one looks at the effect of the language of the ratings scales used to elicit responses to advertising stimuli. de Langhe, Puntoni, Fernandes and van Osselaer (2011) show that completing a questionnaire in one’s native language or in a second language (e.g., English) introduces a systematic effect on the results. The authors demonstrate the tendency among multilingual respondents to report more intense emotions when evaluating consumption experiences and products on rating scales that are not expressed in their native language. The authors term this phenomenon the Anchor Contraction Effect.

The effect occurs because bilinguals perceive emotional scale anchors in their non-native language as less intense than the same emotional anchors in their native language. Because ratings are typically provided relative to these scale anchors, L2 rating scales yield more extreme ratings. To illustrate, imagine rating your response to an advertisement based on a rating scale using the word “ecstatic” versus “glad” as the anchoring point. For the same experience, you would likely provide a lower score for a target ad when rating it against "ecstatic" than "glad", as giving the same ratings would imply a much more intense reaction in the former than in the latter case. This same difference occurs in a multilingual research contexts, when bilingual respondents answer a question using an anchoring point expressed in L1 (more intense) versus L2 (less intense).

> *Principle II.4: L1 rating scales generate lower ratings of emotional intensity than L2 rating scales.*

In our experience, the effect of the language of the rating scales on emotional intensity (Principle II.4) tends to be stronger and more robust than that of the language of target stimuli (Principle II.3; e.g., see the difference in the effect size of the two main effects in de Langhe et al.'s Study 4). We speculate that the reason for this difference is that the effect of the language of rating scales competes less with other psychological processes that contribute to appraisals of emotional intensity. The effect of low-order processes tends to be stronger when they fly under the radar, so to speak, and rating scales are not usually the focus of participants' attention. They are merely a device used to express a belief or a mental state.

> *Principle II.5*: *The effect of the language of the rating scales on reported emotional intensity tends to be stronger and more robust than that of the language of target stimuli*.

What steps should advertising researchers take to control for the Anchor Contraction Effect? If possible, all respondents should answer items in their native language, which leads to Context I as discussed earlier. However, this introduces other problems and it may often not be feasible. Examples of the latter are situations when costs are too high or when the number of native languages in the final sample cannot be predicted beforehand (e.g., when a global audience answers questions online).

When the translation approach is not feasible, researchers can use corrective techniques based on the concomitant presentation of verbal and nonverbal cues, such as emoticons and colors. De Langhe et al. (2001) show that these cues can be effective in removing the effect of language on ratings of emotional intensity. Emoticons can be used when measuring specific emotions, in particular basic emotions that can be easily portrayed with stylized facial expressions. Emoticons are also especially appropriate in online settings and whenever poor comprehension is a potential concern (e.g., in the case of children, low levels of L2 proficiency, or low literacy). In contrast, colors are especially suitable in the case of abstract or complex emotional concepts (e.g., “emotional”, “pity”). Unfortunately, the associations between colors and emotions are partly universal, partly culture-specific, and colors may consequently be vulnerable to cross-cultural differences in interpretation (Hupka, Zaleski, Otto, Reidl and Tarabrina 1997).

More generally, the use of visual cues to mark response categories may be an interesting option to explore in situations where advertising researchers are concerned about different responses to ratings scales by participants who are L1 versus L2 speakers of the language of the survey instrument. It seems likely that, as in the case of perceived emotional intensity, visual cues may often restrict the range of interpretations and limit systematic inter-group differences.

Associations

Languages can often activate associations and stereotypes associated with a particular culture. For example, when Dutch students were asked to generate associations for a series of languages, mentions to “business-like” were twice as frequent in the case of German than in the case of French (Hornikx, van Meurs and Starren 2007). These differences in language associations can impact how consumers judge brands and advertisements. For example, French-sounding brands are perceived as more hedonic than English-sounding brands (Leclerc, Schmitt and Dubé 1994). The prevalence of language associations is culture-dependent and may often differ between native and non-native speakers of the target language. For example, simple Italian words may in general sound more romantic to an L2 speaker of Italian drawn to the language by a passion for Italian culture than by an L1 speaker for whom Italian is the language of everyday interactions. Conversely, language associations among L1 speakers may depend less on cultural stereotypes and more on the everyday use of language. For example, L1 texts are more likely than L2 texts to elicit thoughts about family, friends, home, or homeland among Hispanic Americans (Noriega and Blair 2008).

> *Principle II.6: Some concepts and conceptual domains are more accessible in L1 than in L2 (and vice versa), potentially changing in systematic ways the interpretation of textual information and associated judgments.*

In addition to subtly changing the interpretation of textual information, differential association of languages with concepts can impact answers on self-reported items due to fluency effects. If some concepts and conceptual domains are more accessible in L1 versus L2, language and conceptual domain can be either matched or mismatched. Matched language use increases fluency, which in turn may lead to more positive evaluations (Carroll and Luna 2011). Stimuli and questions related to these domains may thus lead to more positive evaluations.

Principle II.7: Language-domain matching leads to more positive responses.

CONCLUSION

Language issues in advertising research can emerge in two broad situations. They can lead to cross-linguistic differences, when respondents are exposed to advertising stimuli or survey instruments expressed in different languages ("Context I"), or they can lead to multilingual differences, when participants who are native speakers of different languages are exposed to advertising stimuli and survey instruments in a single language ("Context II"). In this paper, we reviewed recent research in the area of language-related methodological issues to generate principles and develop guidelines for researchers facing language issues in their research.

In the cross-linguistic context, a key distinction was made between non-uniform cross-linguistic differences in responses (related to non-equivalence of individual questionnaire items) versus uniform response effects (related to non-equivalence of verbal anchors). Iterative and collaborative translation from a decentered perspective facilitates the design of equivalent stimuli and questions. The use of multiple items and multiple terms to refer to key constructs, in combination with measurement invariance testing, enables advertising researchers to identify potential non-uniform bias issues. To address the risk of uniform bias, careful calibration of verbal anchors is required during design and data analysis.

In the multilingual context, stimuli, questions and response categories may be processed differently in respondents' native versus non-native language, even when comprehension is not an issue. We discussed several ways in which associations will differ as a function of whether L1 or L2 is used in stimuli and/or rating scales. We pointed out that researchers tend to pay more attention to target stimuli than to rating scales when designing studies but often the same effect is stronger when it results from differences in rating scales than from differences in stimuli, at least for processes that do not require much System 2 processing. This observation is likely to hold in the cross-linguistic context as well, although further research is needed to evaluate this possibility.

Furthermore, additional work is needed for some of the principles described above; for example about language effects on confidence and on language associations, and about possible additional effects relevant to Context III (research settings that are concurrently cross-linguistic and multilingual). Moreover, additional language influences will surely be uncovered in future investigations. Many areas are ripe for insight. Here we mention only three:

- *The impact of multi-language stimuli, such as subtitled and multilingual advertising*. For example, Brasel and Gips (2014) show that, among L1 speakers of the language used in a commercial, subtitles in the same language can influence attention and increase recall.
- *The influence of relatively small variations in language, such as accents and dialects*. Language is a social construction and the difference between a language and a dialect is often grounded more in politics than in linguistics. For example, the difference between Norwegian and Swedish is much smaller than the variation in Italian across Italy's regions (e.g., Sicilian versus Venetian). Social psychologists have studied the impact of accents and dialects on stereotyping (Fuertes et al. 2012), but this interest has not translated into attention to potential methodological consequences.
- *Politics of language use in multilingual markets*. In many multilingual societies, language use also has political overtones. For instance, Van Vaerenbergh and Holmqvist (2013) found that consumers in Belgium and Finland were more likely to tip if they were served in L1 compared to when they were served in L2. This

relationship did not depend on consumers' perceived second language proficiency, but was influenced by political considerations. Findings in Belgium, Canada and Finland suggest that bilingual consumers find it particularly important to be served in their native language in high-involvement services (Holmqvist and Van Vaerenbergh 2013). Similar considerations may apply to language use in advertising and advertising research instruments.

Globalization is fast changing both the context and the content of advertising around the world. As a result, the process of globalization is one of the main sources of new research questions for advertising researchers. At the same time, advertising research and practice are not merely responding to globalization. They are a key driver of it. The rise of global brands and the increase in advertising standardization are often identified by commentators and scholars as being among the main factors accelerating the process of cultural homogenization that is such a crucial element of globalization (e.g., the widespread use of English in the ads of countries where English is not an official language). Although less often highlighted within the advertising research community, this bidirectional causal link between advertising practice and globalization is another reason for studying linguistic issues in advertising research. Cross-linguistic and multilingual contexts are common today and are bound to become more common in the future. Understanding how linguistic factors affect the validity of inferences drawn by advertising researchers based on experimental and survey data is thus similarly bound to increase over time. We hope that the typology we proposed and our survey of current knowledge in this area will help improve the quality of advertising research, as well as stimulate other researchers to advance our understanding of linguistic issues in the advertising research process.

REFERENCES

Ahn, Jungsun and Carrie La Ferle (2008), "Enhancing recall and recognition for brand names and body copy: A mixed-language approach," Journal of Advertising, 37 (3), 107-117.

Akaka, Melissa Archpru and Dana L Alden (2010), "Global brand positioning and perceptions: International advertising and global consumer culture," International Journal of Advertising, 29 (1), 37-56.

Alden, Dana L, Jan-Benedict EM Steenkamp and Rajeev Batra (1999), "Brand positioning through advertising in Asia, North America, and Europe: The role of global consumer culture," The Journal of Marketing, 75-87.

Altarriba, Jeanette (2003), "Does cariño equal "liking"? A theoretical approach to conceptual nonequivalence between languages," International Journal of Bilingualism, 7 (3), 305-322.

Andrews, J. C., S. Durvasula and R. G. Netemeyer (1994), "Testing the cross-national applicability of United-States and Russian advertising belief and attitude measures," Journal of Advertising, 23 (1), 71-82.

Arce-Ferrer, Alvara J. (2006), "An Investigation into the Factors Influencing Extreme-Response Style," Educational and Psychological Measurement, 66 (3), 374-392.

Baumgartner, Hans and Jan-Benedict E.M. Steenkamp (2001), "Response Styles in Marketing Research: A Cross-National Investigation," Journal of Marketing Research, 38 (May), 143-156.

Baumgartner, Hans and Jan-Benedict E. M. Steenkamp (2006), "An extended paradigm for measurement analysis of marketing constructs applicable to panel data," Journal of Marketing Research, 43 (3), 431-442.

Baumgartner, Hans and Bert Weijters (2015), "Response Biases in Crosscultural Measurement," in Handbook of Culture and Consumer Psychology, Sharon Ng and Angela Y. Lee eds., Oxford Oxford University Press USA, 370.

Bishop, Melissa M and Mark Peterson (2010), "The impact of medium context on bilingual consumers' responses to code-switched advertising," Journal of Advertising, 39 (3), 55-67.

Brasel, S Adam and James Gips (2014), "Enhancing television advertising: same-language subtitles can improve brand recall, verbal memory, and behavioral intent," Journal of the Academy of Marketing Science, 42 (3), 322-336.

Brislin, Richard W (1970), "Back-translation for cross-cultural research," Journal of Cross-Cultural Psychology, 1 (3), 185-216.

------ (1986), "Research instruments," Field methods in cross-cultural research, 159-162.

Cabooter, Elke, Bert Weijters, Maggie Geuens and Iris Vermeir (2010), "Who said that looks do not matter? The effects of rating scales on response styles," in The 6 Senses-The Essentials of Marketing (EMAC 2010), 157-157.

Carroll, Ryall and David Luna (2011), "The Other meaning of Fluency: Content Accessibility and Language in Advertising to Bilinguals," Journal of Advertising, 40 (3), 73-84.

Cox III, Eli P (1980), "The optimal number of response alternatives for a scale: A review," Journal of Marketing Research, 407-422.

Craig, C Samuel and Susan P Douglas (2005), International marketing research: John Wiley & Sons Chichester.

Dawar, Niraj and Philip Parker (1994), "Marketing universals: Consumers' use of brand name, price, physical appearance, and retailer reputation as signals of product quality," The Journal of Marketing, 81-95.

De Beuckelaer, Alain, Bert Weijters and Anouk Rutten (2010), "Using ad hoc measures for response styles. A cautionary note," Quality and Quantity, 44, 761-775.

De Jong, Martijn G. and Jan-Benedict E.M. Steenkamp (2010), "Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research," Psychometrika, 75 (1), 3-32.

De Jong, Martijn G., Jan-Benedict E.M. Steenkamp and Jean-Paul Fox (2007), "Relaxing Measurement Invariance in Cross-national Consumer Research Using a Hierarchical IRT Model," Journal of Consumer Research, 34 (22), 260-278.

de Langhe, Bart, Stefano Puntoni, Daniel Fernandes and Stijn M. J. van Osselaer (2011), "The Anchor Contraction Effect in International Marketing Research," Journal of Marketing Research, 48 (2), 366-380.

De Meulenaer, Sarah, Patrick De Pelsmacker and Nathalie Dens (2015), "Have No Fear: How Individuals Differing in Uncertainty Avoidance, Anxiety, and Chance Belief Process Health Risk Messages," Journal of Advertising, 44 (2), 114-125.

Dewaele, Jean-Marc (2004), "The emotional force of swearwords and taboo words in the speech of multilinguals," Journal of multilingual and multicultural development, 25 (2-3), 204-222.

Douglas, Susan P and C Samuel Craig (2007), "Collaborative and iterative translation: An alternative approach to back translation," Journal of International Marketing, 15 (1), 30-43.

Duncan, Tom and Jyotika Ramaprasad (1995), "Standardized multinational advertising: the influencing factors," Journal of Advertising, 24 (3), 55-68.

Eckhardt, Giana M, Nikhilesh Dholakia and Russell Belk (2013), "Visual and projective methods in Asian research," Qualitative Market Research: An International Journal, 16 (1), 94-107.

Fuertes, Jairo N, William H Gottdiener, Helena Martin, Tracey C Gilbert and Howard Giles (2012), "A meta-analysis of the effects of speakers' accents on interpersonal evaluations," European Journal of Social Psychology, 42 (1), 120-133.

Garner, Wendell R (1960), "Rating scales, discriminability, and information transmission," Psychological review, 67 (6), 343.

Göbel, Silke M, Samuel Shaki and Martin H Fischer (2011), "The cultural number line: a review of cultural and linguistic influences on the development of number processing," Journal of Cross-Cultural Psychology, 42 (4), 543-565.

Green, Paul E and Vithala R Rao (1970), "Rating scales and information recovery. How many scales and response categories to use?," The Journal of Marketing, 33-39.

Greenfield, Patricia M (1997), "You can't take it with you: Why ability assessments don't cross cultures," American Psychologist, 52 (10), 1115.

Greenleaf, Eric A. (1992), "Measuring extreme response style," Public Opinion Quarterly, 56 (3), 328-350.

Harkness, Janet A., Beth-Ellen Pennell and Alisú Schoua-Glusberg (2004), "Survey questionnaire translation and assessment," Methods for testing and evaluating survey questionnaires, 546, 453-473.

Harkness, Janet A., Michael Braun, Brad Edwards, Timothy P Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell and Tom W Smith (2010), "Comparative survey methodology," Survey Methods in Multinational, Multiregional, and Multicultural Contexts, 1-16.

Harris, Catherine L, Ayse Aycicegi and Jean Berko Gleason (2003), "Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language," Applied Psycholinguistics, 24 (04), 561-579.

Harris, Catherine L, Jean Berko Gleason and Ayse Aycicegi (2006), "When is a first language more emotional? Psychophysiological evidence from bilingual speakers," Bilingual Education and Bilingualism, 56, 257.

Harzing, Anne-Wil (2006), "Response Styles in Cross-national Survey Research A 26-country Study," International Journal of Cross Cultural Management, 6 (2), 243-266.

He, Yi, Michael A Merz and Dana L Alden (2008), "Diffusion of measurement invariance assessment in cross-national empirical marketing research: perspectives from the literature and a survey of researchers," Journal of International Marketing, 16 (2), 64-83.

Holmqvist, Jonas and Yves Van Vaerenbergh (2013), "Perceived importance of native language use in service encounters," The Service Industries Journal, 33 (15-16), 1659-1671.

Hornikx, Jos, Frank van Meurs and Marianne Starren (2007), "An Empirical Study of Readers' Associations with Multilingual Advertising: The Case of French, German and Spanish in Dutch Advertising," Journal of multilingual and multicultural development, 28 (3), 204-219.

Hult, G Tomas M, David J Ketchen, David A Griffith, Carol A Finnegan, Tracy Gonzalez-Padron, Nukhet Harmancioglu, Ying Huang, M Berk Talay and S Tamer Cavusgil (2008), "Data equivalence in cross-cultural international business research: assessment and guidelines," Journal of International Business Studies, 39 (6), 1027-1044.

Hupka, Ralph B, Zbigniew Zaleski, Jurgen Otto, Lucy Reidl, and Nadia V Tarabrina (1997), "The Colors of Anger, Envy, Fear, and Jealousy a Cross-Cultural Study," Journal of Cross-Cultural Psychology, 28 (2), 156-71.

Javier, Rafael A (1989), "Linguistic considerations in the treatment of bilinguals," Psychoanalytic Psychology, 6 (1), 87.

Kubat, Umut and Vanitha Swaminathan (2015), "Crossing the cultural divide through bilingual advertising: The moderating role of brand cultural symbolism," International Journal of Research in Marketing, 32 (4), 354-362.

Laczniak, Russell N (2015), "The Journal of Advertising and the Development of Advertising Theory: Reflections and Directions for Future Research," Journal of Advertising, 44 (4), 429-433.

Leclerc, France, Bernd H Schmitt and Laurette Dubé (1994), "Foreign branding and its effects on product perceptions and attitudes," Journal of Marketing Research, 263-270.

Lerman, Dawn and Michael Callow (2004), "Content analysis in cross-cultural advertising research: insightful or superficial?," International Journal of Advertising, 23 (4), 507-521.

Levitt, T. (1983), "The Globalization of Markets," Harvard Business Review, 61 (3), 92-102.

Little, Todd D. (1997), "Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues," Multivariate Behavioral Research, 32 (1), 53-76.

------ (2000), "On the Comparability of Constructs in Cross-Cultural Research A Critique of Cheung and Rensvold," Journal of Cross-Cultural Psychology, 31 (2), 213-219.

Luna, David (2011), "Advertising to the buy-lingual consumer," in Language and bilingual cognition, Vivian Cook and Benedetta Bassetti eds.: Psychology Press.

Luna, David and Laura A Peracchio (2001), "Moderators of language effects in advertising to bilinguals: A psycholinguistic approach," Journal of Consumer Research, 28 (2), 284-295.

------ (2005), "Sociolinguistic effects on code-switched ads targeting bilingual consumers," Journal of Advertising, 34 (2), 43-56.

MacCallum, Robert C, Michael W Browne and Hazuki M Sugawara (1996), "Power analysis and determination of sample size for covariance structure modeling," Psychological Methods, 1 (2), 130.

Marsh, Herbert W, Oliver Lüdtke, Benjamin Nagengast, Alexandre JS Morin and Matthias Von Davier (2013), "Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models," Psychological Methods, 18 (3), 257.

Mick, David Glen (1992), "Levels of subjective comprehension in advertising processing and their relations to ad perceptions, attitudes, and memory," Journal of Consumer Research, 411-424.

Minton, Elizabeth, Christopher Lee, Ulrich Orth, Chung-Hyun Kim and Lynn Kahle (2012), "Sustainable Marketing and Social Media: A Cross-Country Analysis of Motives for Sustainable Behaviors," Journal of Advertising, 41 (4), 69-84.

Moors, Guy, Natalia D Kieruj and Jeroen K Vermunt (2014), "The effect of labeling and numbering of response scales on the likelihood of response bias," Sociological Methodology, 44 (1), 369-399.

Muthén, Bengt and Tihomir Asparouhov (2012), "Bayesian structural equation modeling: a more flexible representation of substantive theory," Psychological Methods, 17 (3), 313.

------ (2013), "BSEM measurement invariance analysis," Mplus Web Notes, 17, 1-48.

Nickerson, Raymond S (2000), "Null hypothesis significance testing: a review of an old and continuing controversy," Psychological Methods, 5 (2), 241.

Noriega, Jaime and Edward Blair (2008), "Advertising to bilinguals: Does the language of advertising influence the nature of thoughts?," Journal of Marketing, 72 (5), 69-83.

Okazaki, Shintaro and Barbara Mueller (2007), "Cross-cultural advertising research: where we have been and where we need to go," International Marketing Review, 24 (5), 499-518.

Okazaki, Shintaro, Barbara Mueller and Sandra Diehl (2013), "A Multi-Country Examination of Hard-Sell and Soft-Sell Advertising: Comparing Global Consumer Positioning in Holistic-and Analytic-Thinking Cultures," Journal of Advertising Research, 53 (3), 258-272.

Pavlenko, Aneta (2007), Emotions and multilingualism: Cambridge University Press.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee and Nathan P. Podsakoff (2003), "Common method biases in behavioral research: A critical review of the literature and recommended remedies," Journal of Applied Psychology, 88 (5), 879-903.

Preston, Carolyn C and Andrew M Colman (2000), "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences," Acta psychologica, 104 (1), 1-15.

Puntoni, Stefano, Bart De Langhe and Stijn MJ Van Osselaer (2009), "Bilingualism and the emotional intensity of advertising language," Journal of Consumer Research, 35 (6), 1012-1025.

Rindfleisch, Aric, Alan J Malter, Shankar Ganesan and Christine Moorman (2008), "Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines," Journal of Marketing Research, 45 (3), 261-279.

Rose, G. M., V. D. Bush and L. Kahle (1998), "The influence of family communication patterns on parental reactions toward advertising: A cross-national examination," Journal of Advertising, 27 (4), 71-85.

Smith, Tom William (2003), "Developing comparable questions in cross-national surveys," Cross-cultural survey methods, 69-92.

------ (2004), "Developing and evaluating cross-national survey instruments," in Methods for testing and evaluating survey questionnaires, Robert M. Groves, Graham Kalton, J. N. K. Rao, Norbert Schwarz and Christopher Skinner eds., 431-452.

Steenkamp, Jan-Benedict E.M. and Hans Baumgartner (1998), "Assessing Measurement Invariance in Cross-National Consumer Research," Journal of Consumer Research, 25 (June), 78-90.

Tavassoli, Nader T and Jin K Han (2002), "Auditory and visual brand identifiers in Chinese and English," Journal of International Marketing, 10 (2), 13-28.

Taylor, Charles R. and Shintaro Okazaki (2015), "Do Global Brands Use Similar Executional Styles Across Cultures? A Comparison of US and Japanese Television Advertising," Journal of Advertising, 44 (3), 276-288.

Van de Vijver, Fons JR and Kwok Leung (1997), Methods and data analysis for cross-cultural research: Sage.

Van Vaerenbergh, Yves and Jonas Holmqvist (2013), "Speak my language if you want my money: Service language's influence on consumer tipping behavior," European Journal of Marketing, 47 (8), 1276-1292.

Weijters, Bert, Niels Schillewaert and Maggie Geuens (2008), "Assessing response styles across modes of data collection," Journal of the Academy of Marketing Science, 36 (3), 409–422.

Weijters, Bert, Maggie Geuens and Hans Baumgartner (2011), "Lost in translation: The effect of Language on response distributions in Likert data," in Society for Consumer Psychology Annual Winter Conference (SCP), 2011, February 24-26, , N. Mandel and D. Silvera eds., Atlanta, Georgia.

------ (2013), "The Effect of Familiarity with the Response Category Labels on Item Response to Likert Scales," Journal of Consumer Research, 40 (2), 368-381.

Weijters, Bert, Hans Baumgartner and Maggie Geuens (2016), "The Calibrated Sigma Method: An Efficient Remedy for Between-Group Differences in Response Category Use on Likert Scales," working paper.

Williams, Larry J and Patricia J Holahan (1994), "Parsimony-based fit indices for multiple-indicator models: Do they work?," Structural Equation Modeling: A Multidisciplinary Journal, 1 (2), 161-189.

Wyer, Robert S (2002), "Language and advertising effectiveness: Mediating influences of comprehension and cognitive elaboration," Psychology & Marketing, 19 (7-8), 693-712.

TABLE 1

A Typology of Linguistic Research Contexts

		Respondents’ language	
		A. All L1	B. L1&L2
Language used for data collection (stimuli and instruments)	1. Single language	No linguistic issues	Context II Multilingual
	2. Multiple languages	Context I Cross-linguistic	Context III

Note: L1 and L2 refer to a multilingual respondent’s native or non-native language, respectively.

TABLE 2

Recommendations for Making Items More Translatable

1. Use short simple sentences of less than 16 words.

2. Employ active rather than passive voice.

3. Repeat nouns instead of using pronouns.

4. Avoid metaphors and colloquialisms.

5. Avoid the subjunctive.

6. Add sentences to provide context to key items. Reword key phrases to provide redundancy.

7. Avoid adverbs and prepositions telling “where” or “when.”

8. Avoid possessive forms where possible.

9. Use specific rather than general terms.

10. Avoid words indicating vagueness (e.g., “probably,” “maybe,” “perhaps”).

11. Use wording familiar to the translators.

12. Avoid sentences with two different verbs if the verbs suggest different actions.

Note: Taken from Brislin (1986)

TABLE 3

Example of a Label Familiarity Check

		# Google hits	LN(Google hits "... agree" /Google hits "agree")
English	Strongly agree	2640000	-5.86
	Completely agree	6520000	-4.96
	Totally agree	13600000	-4.22
	Agree	927000000	
French	Fortement d’accord	46700	-6.88
	Complètement d’accord	380000	-4.79
	Tout à fait d’accord	11300000	-1.39
	D'accord	45500000	

Note: Based on Weijters, Geuens and Baumgartner (2013). In the current example, we did not specify a geographic location in the search; doing so (e.g. limiting the search to the UK) yields different results. Moreover, results can fluctuate over time, location and user. However, the general pattern is generally consistent. In this example, ‘completely agree’ and ‘Complètement d’accord’ are roughly equally familiar in English and French.

TABLE 4

Illustrative Example of the Calibrated Sigma Method

			Response category				
Group	Step	Operation	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Language A	1	Mean frequency (16 control items)	0.800	3.200	6.400	3.200	2.400
	2	Mean proportion	0.050	0.200	0.400	0.200	0.150
	3	Cumulative proportion (Pk,g)	0.050	0.250	0.650	0.850	1.000
	4	[½ * (Pk,g + Pk-1,g)]	0.025	0.150	0.450	0.750	0.925
	5	Sigma value	-1.960	-1.036	-0.126	0.674	1.440
Language B	1	Mean frequency (16 control items)	1.600	4.800	4.800	2.400	2.400
	2	Mean proportion	0.100	0.300	0.300	0.150	0.150
	3	Cumulative proportion (Pk,g)	0.100	0.400	0.700	0.850	1.000
	4	[½ * (Pk,g + Pk-1,g)]	0.050	0.250	0.550	0.775	0.925
	5	Sigma value	-1.645	-0.674	0.126	0.755	1.440

Note: k (1 to K) indexes response categories; g (1 to G) indexes language groups.