2.2 Developing and assessing comparable questions in cross-cultural survey research on tobacco

Introduction

The WHO FCTC aims to address the global tobacco epidemic by coordinating national policies to combat tobacco use. This volume illustrates possible conceptual frameworks, methods, and data sets that will be useful for conducting comparative, international research to better understand which policies work and why. This section aims to provide researchers with a basic overview of measurement issues involved in the design and analysis of cross-cultural comparative research, as well as some of the methods currently recommended for attempting to resolve these issues. When possible, we illustrate our points with examples from cross-cultural tobacco research. The organisation of the section follows the general stages of research design, illustrating the corresponding methods used to assess and to avoid introducing systematic measurement error due to cultural differences across the populations in which the research is carried out. The growing literature that we discuss generally reflects concerns related to conducting comparative research across nations and linguistic groups. In most cases, however, the implications and methods we describe extend to intranational studies involving different ethnic groups or even single ethnic groups that speak the same language (e.g. Spanish-speaking Latinos in the USA; people from different socioeconomic groups). In this regard, our general approach may be useful to researchers interested in ensuring the validity of comparative analyses across cultural subgroups within increasingly multi-cultural, intranational settings.

Cross-cultural and cross-national research is often done under the unexamined assumption that question meaning, comprehension and measurement properties are equivalent across cultural groups (Bollen *et al.*, 1993; Smith, 2004a). However, cross-cultural differences in language, social conventions, cognitive abilities and response styles may cause systematic measurement error that biases results in unpredictable ways (Fiske *et al.,* 1998; Harkness *et al.*, 2003a). Apparent differences found across socio-cultural groups may be merely due to measurement artefacts, such as systematic group differences in the meanings ascribed to the same question, whether phrased in the same or different languages. Conversely, true differences may be obscured by such factors as the differential influence of social desirability or the exclusion of items that are important indicators of study constructs in one cultural context but not in another. Whereas the implications of these issues appear most obvious for international comparative research, if left unaddressed, they may also impede our understanding of why certain tobacco policies work better among some socio-cultural groups than among others. In the end, valid cross-cultural comparison demands that measurement error be minimised across the settings and groups of interest (Bollen *et al.,* 1993; Smith, 2004a).

Equivalence of conceptual frameworks

Cross-cultural survey research should begin by assessing whether the conceptual definitions and theoretical frameworks that orient the study reasonably apply across the contexts in which the survey data will be collected. Consideration

of the universal applicability or culturally-specific nature of study concepts is important because their definitions should inform subsequent stages of question selection, development, adaptation and assessment. For example, some concepts may have single or multiple dimensions, each of which should be reflected in its conceptual definition. In some populations the social acceptability of smoking can be characterised by at least two dimensions, one that references close social network members and another that concerns perceptions of a more distal, abstract socio-cultural milieu (Thrasher *et al*., 2006a). These referents may be further subdivided by perceptions of the actual behaviour (i.e. descriptive norms) and desired behaviour (i.e. injunctive or prescriptive norms) (Cialdini, 2003). Hence, at least four dimensions could be delineated within a conceptual definition of the social acceptability of smoking. Nevertheless, the number of dimensions may vary between or within any particular population. Cross-cultural studies should consider construct dimensionality and whether it might differ across cultural groups.

Ensuring the equivalence of concepts across cultural contexts or groups should begin with literature reviews on the topic and concepts of interest. Pertinent literature may nevertheless escape the reach of search engines or the linguistic capabilities of those conducting the reviews, or this literature may simply not exist. This problem may be addressed by establishing collaborative research groups that involve at least one representative from each country or cultural group in which surveys will be conducted (Kuechler, 1987). Ideally, each representative should have native language proficiency and be knowledgeable of both the study topic and the particular contexts in which data collection will take place. Formulating the study's conceptual framework in dialogue among a team of such researchers can help anticipate incongruities in the conceptual framework across survey contexts, and thereby avoid any ethnocentric or universalist tendencies in measurement that might result (Van de Vijver & Hambleton, 1996). Furthermore, this dialogue may help identify cultural or contextual factors that may be important modifiers of tobacco policy effects. Such potential modifiers may otherwise escape consideration because researchers in one context either take them for granted because of their ubiquity or have never considered them because of their absence. For example, strong religious beliefs in some countries may play such a role.

The collaborative process of defining the concepts and framework that orient questionnaire design goes some way toward ensuring that the survey instrument will be meaningful for study participants. There are a number of tensions and difficulties with the collaborative approach, however. As the number of nations or cultural groups involved in the study increases, so do the amount of difficulty and time spent to coordinate efforts and reach consensus (Kuechler, 1987). Granting agencies often demand clearly defined conceptual frameworks before they will fund a project, and without funding to develop this framework, it may be difficult to engage collaborators. The "local" representatives with whom collaboration occurs may actually be quite cosmopolitan, perhaps directly or indirectly socialised into the Western scientific enterprise. Hence, the "cultural" perspective any particular representative provides may be a hybrid form that is at once transnational yet circumscribed by particular social class, gender, and cultural divisions within the country of interest. In this regard, people who have direct knowledge of the local realities of target populations in which survey research will take place may make more substantial contributions toward the development of culturally applicable concepts. Even so, status asymmetries among group members may ultimately overwhelm more local (and perhaps more locally relevant), epistemologies, theories and concepts, particularly if they are incongruent with Western scientific principles (Johnson, 1998). These challenges should be recognised and, to the extent possible, overcome. Collaboration with representatives from each cultural setting nevertheless

forces at least some consideration of cultural particularities and concerns. The resulting conceptual framework should be more likely to "fit" the contexts studied than a framework constructed in the absence of input and involvement of representatives from these different settings.

Question selection and development: equivalence of indicators

The practice of selecting or developing questionnaire items in one language and translating them into other languages is common in cross-cultural survey research. The use of established items saves time, is inexpensive, and allows for ready comparison with other studies that have used the same measures. Ideally, these items will have been pre-tested and found to have suitable measurement properties across subgroups who speak the source language, as well as among those from the linguistic and cultural groups in which the research will be conducted. Such analyses have been done only for a few tobacco survey questions, including those related to dependence (see Section 3.3). If sound measurement properties have been found for the item in one linguistic or cultural context, these properties do not necessarily carry over to the translated version of the item, no matter how good the translation (Harkness *et al*., 2003b). To help ensure equivalence of question comprehension and meaning, pretesting is needed in each major cultural context or major sociocultural group under consideration (see page 68).

One reason why item selection matters is that wording that appears neutral may actually contain phrases or terms with culturally idiosyncratic connotations, making translation difficult (Harkness, 2003). Attempts to capture the meaning of culturally anchored wording—no matter how unambiguous in the original language—may produce awkward translations that violate question design principles and thereby introduce systematic error. One clear example comes from the German General Social Survey item "Das leben en vollen zügen genieβen," which literally translates to English as the nonsensical "Enjoy life in full trains." For American English, a more appropriate translation is the adapted, non-literal phrase "Live life to the fullest" (Harkness, 2003). The often unconscious embedding of cultural anchors in questions may lead to their discovery only through the translation process itself. Similarly, question meanings may not be shared across contexts, and different items will need to be developed in order to adequately reflect study concepts. For these reasons, cross-cultural survey methodologists increasingly argue for methods that open up the translation process to greater scrutiny and more conscious group decision-making (Harkness & Schoua-Glusberg, 1998; Hambleton *et al*., 2005). When cultural anchoring is discovered, unambiguous phrasing in the translated version of the question may necessitate changing the wording of the original language item in order to maintain equivalence (see page 68). Literal question translation may nevertheless result in equivalent meanings across languages. However, it is crucial to consider whether the resulting question adequately captures the concept of interest and whether a non-literal adaptation of the question is necessary to do so (Van de Vijver & Leung, 1997; Van de Vijver, 2004).

Cross-cultural survey research generally involves translating items that are established measures for particular constructs in one language group. For this reason, our next sub-section focuses more intensively on translation approaches. However, researchers may nevertheless consider developing a core set of indicators for use across all sites, supplemented by culture-specific indicators of the same constructs. The selection of culturally-specific indicators should consider measurement research on the same or related concepts conducted within the culture. However, such research may not exist or may involve items that researchers believe are inadequate to capture the meaning of the concept of interest. Item development can follow any of a variety of methods that are standard practice in measurement development, including expert-

driven techniques (DeVellis, 1991) or those that involve eliciting meanings from the target group of interest, as with focus groups (Stewart & Shamdasani, 1998), structured interviews (Spradley, 1979), free-listing, pile sorts and other qualitative techniques (Bernard, 1994; Berkowitz, 2001). Rapid anthropological assessment techniques have also been developed to reduce the time and effort required for more traditional ethnographic methods, with one such effort having already developed a framework for tobacco-related research among youth (Mehl *et al*., 2002). These and other methods could also be used for developing equivalent concept definitions across contexts.

One rarely used approach to item selection and development involves simultaneous, yet independent work by each group responsible for a particular linguistic or cultural subgroup involved in the study (Harkness *et al*., 2003b). This strategy is likely to work best when teams use conceptual definitions that adequately apply across contexts, thereby removing the likelihood that the concepts under consideration are too culturally-specific and, hence, idiosyncratic. Each team would assemble and/or develop items that they believe best reflect the study concepts. In the end, however, incommensurability of items across contexts presents analytic difficulties, as few statistical techniques allow direct comparison of dissimilar stimuli. Furthermore, cross-cultural comparison of only those items with similar content may exclude culturally specific items that are the best and most meaningful indicators of the concept of interest. Overall, this approach involves relatively high development costs, openness to making changes to the source instrument, and complex organisational structure to adequately coordinate teams (Harkness *et al*., 2003b).

Example of focus groups for item development:

Before fielding an international survey of adult smokers in Mexico, in-depth interviews and focus groups were conducted with adult smokers, with discussions oriented by the conceptual domains included in the survey (Thrasher & Bentley, 2006; Thrasher *et al*., 2006a). One concept of interest involved perceived voluntary control over smoking behaviour. This attribution to tobacco consumption behaviour may not only be relevant to self-efficacy regarding quit attempts, but also to perceptions of tobacco products as deviant when compared to other products that people freely decide to consume. When prompted, most all Mexican smokers agreed that tobacco was addictive; however, they found it difficult to explain what "addiction" meant. It became clear that the more common manner of talking about and understanding tobacco's hold over their behaviour was through the term *vicio* or "vice", which connotes a guilty pleasure that is difficult to control, potentially dangerous, and often looked down upon socially. Participants generally agreed that the term addiction, as well as the term *droga* or "drug" also had these connotations. Analyses of data from a subsequent pilot survey of items developed to capture these additional meanings (fumar es un *vicio* ['smoking is a vice"]; *el cigarro es una droga* ["a cigarette is a drug"]) found that these items loaded onto the same dimension as the primary indicator of perceived behavioural control (*tabaco es adictivo* ["tobacco is addictive']), improving the measurement properties of the construct (Thrasher *et al*., 2006a). While the meaning of "a cigarette is a drug" would likely translate back to English, the use of an equivalent English language item that included the term "vice" may be meaningful only within certain subcultural religious groups. As such, this example helps illustrate the development of a culturally-specific item that complements a core item shared across surveys. Cognitive testing of the original item in English and Spanish (see sub-section on Questionnaire Pre-Testing) could complement further statistical analyses (see sub-section on Quantitative assessment) in order to determine whether the single item on vice in the Mexico sample might be used as equivalent to the single item on addiction in samples from other countries.

Approaches to survey translation

Translation of surveys in cross-cultural research is often an afterthought, with little attention paid to the design issues involved in the complex task of producing instruments with comparable measurement properties across languages and contexts (Harkness & Schoua-Glusberg, 1998; Harkness, 2003). Steps described above to ensure the applicability and relevance of construct definitions across diverse contexts provide a foundation for sound translation practices (Harkness *et al*., 2003b). Yet, even with such a framework in place, any of a variety of translation methods could be followed, each with its own advantages and disadvantages. Generally, survey research follows the "Ask-the-Same-Question" model, in which a questionnaire is developed in the "source" language and translated to other "target" languages. Because of its widespread use, we describe methods based on this model, including the "decentering" approach, whose iterative process of translation demands at least some flexibility in the wording of the source language questionnaire.

Ideally, people who translate a questionnaire should be skilled, professional translators who are bilingual in the source and target languages, while having at least some basic training in general principles for developing questions with good measurement properties (for some basic recommendations regarding instrument design, see: Dillman (2007), Bradburn and coworkers (2004) and/or Willis (2005)). If this is not possible, then translation should be conducted by people who are fluent in both languages and practiced in the translation between them. At first glance, a single-person translation appears time- and cost-effective. However, relying on a single person to make all translation decisions may introduce comprehension problems due to regional variance in linguistic expression and meaning, as well as the translator's own idiosyncratic interpretations and inevitable oversights (Harkness *et al*., 2004). Since these issues may result in non-equivalent stimuli and, hence, invalid comparison, the efficacy of single-translator methods increasingly has been called into question (Harkness & Schoua-Glusberg, 1998; Hambleton *et al*., 2005).

A team approach to translation, which involves more than one person who is fluent in the source and target languages, appears to help overcome some biases that result from single-person translations. Team approaches open up to examination and discussion the complex decision-making that occurs in translation, providing a greater range and more balanced critiques of translation options (Guillemin *et al*., 1993; McKay *et al*., 1996; Harkness & Schoua-Glusberg, 1998). Aside from skilled, professional translators (of which there may be more than one), Harkness (2003) suggests that two additional roles be filled in the team approach. *Reviewers* should have language abilities that are as strong as the translators', supplemented with knowledge of questionnaire design principles, study design and the topic of interest. *Adjudicators* should at least share this methodological and topical knowledge, as they will make the final decisions about which translation to adopt, preferably in cooperation with the reviewers and translators who have been more intimately involved in the details of translation and evaluation. When an adjudicator does not understand the source or target language well, Harkness suggests that consultants should be hired to provide this skill. Team approaches involve greater expense, time and coordination than single-person translations; however, this approach is recommended and used by numerous ongoing survey operations, including the Survey of Health Ageing and Retirement in Europe (Börsch-Supan *et al*., 2005), the US Consumer Assessment of Health Care Providers and Systems (Weidmer *et al*., 2006), the US Census Bureau (Pan & de la Puente, 2005) and the European Social Survey (Harkness & Blom, 2006).

The "committee approach" to translation is increasingly viewed as the gold standard in cross-cultural survey research (Harkness & Schoua-Glusberg, 1998; Harkness *et al*., 2004). Generally two to four translators are used, with each additional translator providing more material for critical

discussion of translation possibilities. The *parallel translation* method involves each translator independently translating the same source questionnaire in its entirety. Some of the costs associated with parallel translations can be cut by employing *split translations*, in which each translator is assigned different parts of the source questionnaire. In either case, translators bring their independent translations to a reconciliation meeting where at least one reviewer and perhaps the adjudicator work with the translators to reach agreement on the best translation. The chosen wording could be taken directly from one translation, a mixture of the different phrasings offered, or a previously unconsidered wording that emerges from discussion of the independent translations. Because each question is translated independently by at least two people, parallel translations are likely to offer a greater range of translation possibilities than either split translations or a single translator would produce. The final versions can be adjudicated at the reconciliation meeting or, perhaps provided to the adjudicator for later consideration.

The team approaches to translation may seem extravagant in the context of many low-resource environments. However, the relatively low additional cost of hiring a second translator is likely to offset subsequent costs and data quality issues that might result from an unscrutinised translation. Indeed, this process may anticipate and address questionnaire problems that otherwise only come to light in pre-testing or data analysis. This is not to suggest, however, that this strategy should replace questionnaire pre-testing. Both researchers and translators are likely to come from social strata that differ from the majority of research participants. Hence, translation assessment procedures described below are critical to ensuring sound comprehension and equality of measurement.

Researchers may want to consider allowing for minor changes to the source language questionnaire due to issues that emerge through translation. As described earlier, cultural anchoring of words and phrases may result in translated items that shift original meaning or that violate good question design principles. Either way, systematic measurement error may result. One possible approach to equalising question meaning involves an iterative translation process called "decentering" (Werner & Campbell, 1970). In this method, a source questionnaire provides the starting point for translation to target languages, which could be done using any of the aforementioned methods. However, translators and reviewers signal which items appear to introduce non-equivalence of meaning. Those in charge of each language version of the questionnaire then work in iterative fashion, changing items by tacking back and forth across the translations until all versions appear harmonised. For example, one project using this method translated an English language item that included the term "embarrassed," which existed in the target languages but had stronger connotations than in English. Researchers decided to substitute another term, "unhappy about," which was easier to harmonise across the target languages and did not compromise the measurement properties of the original language item (Eremenco *et al.*, 2005).

The iterative approach to translation is difficult, time-consuming and expensive, and each additional language included in the process will multiply these disadvantages (Harkness *et al.*, 2003b). Unlinking questions from their cultural connotations may result in unwanted ambiguity due to vague, unidiomatic phrasing. Furthermore, changes in source item wording may necessitate pre-testing in order to ensure that measurement properties have not suffered.

Whichever translation approach is taken, we strongly recommend that those involved in cross-cultural tobacco research document their decisions regarding item selection, development and translation. Study concepts should be clearly specified and linked to original, source language items. Translators should be encouraged to keep notes regarding their decision-making processes when translating the item to another language. Similarly, team approaches to translation review should involve further docu-

mentation about how final decisions were made. If the entire questionnaire is not subject to later pre-testing, these notes will help determine which subset of items should be scrutinised more closely. This documentation will also enable future researchers to adequately interpret the data associated with these questions, while providing critical information for further improvement of the measures in later studies.

Example of the committee approach:

One example of the committee approach using parallel translation involves translating an American English-language source survey of adult smokers to the Mexican variety of Spanish. Independent translations of the survey were provided by four bilingual professional translators, three of whom were Mexican nationals and the fourth an American who had been living in Mexico for 19 years and working as a professional translator for 24 years. Although all of them had at least some experience with survey translation, each was provided with summary materials on question design principles and asked to follow them. Two of the Mexican translators were recruited because they were regular smokers, as was a young adult, bilingual Mexican research assistant who had been involved in earlier stages of the project and who served as a reviewer at the reconciliation meeting. As members of the target population in which the survey would be administered, these three people helped ensure the use of natural terminology and comprehensibility among smokers. Because of logistical and cost constraints, representatives were not included from each of the different regions of Mexico where the survey would be administered. This was a potential limitation.

The reconciliation meeting involved a full day of work with three translators (one was unable to make the meeting but provided her independent translation), two bilingual reviewers, and a bilingual reviewer/adjudicator. After beginning the session with a further discussion of question design principles, we examined the original English version and all four translations, addressing one question at a time. As emphasised in the description of the methodology, this process produced a range of possible translations, even for questions that, on the surface, appeared straightforward. The beginning of the process was time-consuming and challenging. However, decision-making became easier as participants became comfortable with the process and as we reached agreement on terms, grammatical structure, and response options that were repeated throughout the questionnaire.

As an illustration of the decision-making processes involved in this method, the following describes how we translated the last phrase of the question “On average, how many cigarettes do you smoke each day, including both factory-made and roll-your-own cigarettes?” This clarification to this standard question had been included in the source language questionnaire in order to ensure that respondents considered “roll-your-own” cigarettes, particularly as switching to lower-cost tobacco is a common response to raising the price of cigarettes (Young *et al*., 2006).

One non-smoking translator deleted the last clause of the English version because she had never heard of people using such cigarettes in Mexico. However, we did not want to exclude mention of this practice since it occurs in Mexico, although at a low prevalence. Indeed, one aim of the survey was to estimate this prevalence, although it would be measured with more precision in a question that appeared later in the survey instrument. Two general options for describing factory-made cigarettes emerged: one was a more literal translation (*cigarros hechos en fábricas*, literally “cigarettes made in factories”) and the other turned the focus toward branded and marketed cigarettes (*cigarros de marcas comerciales*, literally, “commercial cigarette brands”). This second focus was discarded since rolling tobacco is also branded and marketed, even though unbranded, loose tobacco can be bought in some regions of Mexico. The more literal translation sounded awkward and seemed to divert attention from the main question content. In the end, we decided on a phrase that could be roughly translated as “cigarettes from the pack”

(*cigarros de cajetilla*), since the word for pack (*cajetilla*) connoted "factory-made" without sounding awkward, while setting up the contrast with the "roll-your-own" type cigarettes that would be mentioned thereafter.

For the final clause in the question, two options emerged from the three independent translations. One used a term for rolling that is also common for rolling marijuana cigarettes (*cigarros forjados a mano*) while the other introduced the participant as the one who "made" (hacer) the cigarettes (*cigarros hechos por* usted, literally "cigarettes made by you"). There was agreement that either option could confuse people who did not engage in rolling cigarettes — this would be the vast majority of study participants. However, reference to the participant making the cigarettes seemed on track, since not including the participant as agent could cause people to think of cigars, which are also hand rolled, but by someone else. We agreed on a longer version "cigarettes that you make by hand" (*cigarros que usted hace a mano*). Later cognitive interviews indicated that this phrase nevertheless connoted marijuana cigarettes for some participants, and so the final, pretested version clarified that these were cigarettes made with tobacco: En general, *¿cuántos cigarros al día fuma, incluyendo los cigarros de cajetilla y los cigarros de tabaco que usted hace a mano?* (Literally, "In general, how many cigarettes do you smoke each day, including cigarettes from the pack and tobacco cigarettes that you make by hand?"). Finally, interviewer training included a focus on the meaning of the question, so that interviewers could anticipate and respond to any comprehension difficulties that they sensed among participants.

This example illustrates a number of the advantages that accompany the committee approach to translation. Importantly, there were a variety of options to choose from. Consistency of terminology and phrasing across translation options would have provided support for selecting a particular translation. The example above indicated inconsistencies in the terms and wording, which led to group decision-making about the best way to resolve discrepancies. Moreover, resolutions to discrepancies did not appear in the originally translated versions. Finally, the version agreed upon in the reconciliation meeting still needed to be altered a little after cognitive testing indicated undesirable connotations for one part of the question.

Culturally moderated response styles

Comparisons across cultural groups may be biased by systematic differences in "response styles," such as social desirability, extreme responding, and acquiescence. Of particular concern are social desirability effects, which manifest when respondents misrepresent or edit their true responses to a question in order to project an image of themselves that accords with their perceptions of social norms and expectations (Marlow & Crowne, 1960). The phenomenon appears to be universal across societies, with stronger effects found when considering self-report of behaviours or beliefs that are socially sanctioned within a given cultural context (Johnson & Van de Vijver, 2004). Hence, the differential effects of social desirability on self-reported tobacco attitudes, beliefs, and behaviours should be proportional to the level of tobacco's social unacceptability across the socio-cultural groups under consideration. Because social desirability effects also appear stronger among minority or disenfranchised groups within a society (Ross & Mirowsky, 1984; Edwards & Riordan, 1994; Warnecke *et al.*, 1997), it may disproportionately influence national samples that contain more minority group participants.

Social desirability appears positively correlated with a number of macro-level societal characteristics, such as higher levels of "collectivism" and lower levels of "individualism." Higher levels of social desirability appear congruent with, and may stem from, collectivist codes of social interaction that emphasise courtesy, maintaining harmonious relations and saving face (Marín & VanOss Marín, 1991; Johnson & Van de Vijver, 2004). Smokers from collectivist societies that stigmatise tobacco use may view true representation of their thoughts and behaviours in an

interview context as threatening these more important elements of social interaction. On the other hand, people from individualist societies appear to have stronger prohibitions against providing misleading information (Triandis, 1995). Hence, smokers in these societies may be less likely to provide socially desirable responses independent of the extent of social sanctions against smoking. This suggests that individualism/collectivism and social sanctions against tobacco are likely to interact, producing differential social desirability effects on tobacco survey questions. The strongest effects of social desirability should occur under conditions of strong stigmatization of smoking behaviour in a collectivist society, whereas the weakest effects would occur in individualist societies with weak stigmatisation. Future research should empirically test this proposition.

Several other response styles have also been found to vary across cultures (Baumgartner & Steenkamp, 2001). Two that have perhaps received the most attention are extreme response styles (Smith, 2004b) and acquiescence (Knowles & Condon, 1999). Extreme response styles refer to the greater preference of respondents from some cultures to select the most extreme endpoints of response scales, whereas respondents from other cultures are more likely to make less extreme choices when answering. Moreover, some respondents exhibit a greater tendency to agree with questions read by interviewers, even when the questions are contradictory, a process referred to as acquiescent responding.

Although there is general agreement that social desirability, extreme responding and acquiescence are each moderated by culture, there is less consensus or available evidence regarding how to best account for these potential sources of measurement error when conducting cross-cultural research. Several researchers have attempted to neutralise social desirability effects by explicitly measuring these propensities and then statistically adjusting for them (Nederhof, 1985). Most reported attempts to introduce social desirability corrections, however, have been unsuccessful (Ones *et al.*,1996; Ellingson *et al.*, 1999; Fisher & Katz, 2000), suggesting that other approaches should be explored (for reviews of other methods of addressing social desirability in survey research, see Nederhof (1985) and Paulhus (1990)). Some researchers have also reported studies in which they assessed extreme responding and/or acquiescence via structural equation modelling (Mirowsky & Ross, 1991; Greenleaf, 1992; Watson, 1992; Billiet & McClendon, 2000; Cheung & Rensvold, 2000). In general, however, there is no consensus on how to best confront problems of systematic cross-cultural variability in survey response styles.

During data collection, efforts are also commonly made to minimise the social distance between respondents and interviewers by attempting to match them on ethnic background or demographic characteristics in hopes of minimising the social desirability pressures placed on respondents. For example, in contexts where deference to authority is a key cultural value, interviews conducted by older people of higher social status may induce strong social desirability effects. Numerous studies are available that demonstrate respondent deference to interviewers who represent differing cultural backgrounds (Cotter *et al.*, 1982; Anderson *et al.*, 1988; Finkel *et al.*, 1991; Davis, 1997; Johnson *et al.*, 2000), although it should be noted that none of these studies are based on experimental evidence. Under some circumstances, too little social distance between respondents and the person interviewing them may encourage socially desirable responding (Dohrenwend *et al.*, 1968). Concern with the effects of social distance can also be extended to interview mode, as the degree of privacy afforded by each mode of data collection may exert differential pressures on respondents to provide socially desirable information. Although little information is available with which to examine cultural variability in mode of interview effects (Marín & Marín, 1989), it would seem likely that the social sensitivity of the answers being requested and respondent culture might interact with survey mode in ways that either magnify or

minimise substantive differences across groups. These effects may be difficult to predict, particularly given the near absence of research on this topic. Researchers should thus carefully consider how the social sensitivity of the topics examined might vary across the groups studied, the types of questions asked, and how the mode of data collection might influence participants' responses.

Questionnaire pre-testing and translation assessment

We focus on two approaches to questionnaire pre-testing and translation assessment. First, we discuss back-translation, which has been used frequently and even viewed as a gold standard for translation assessment; however, we describe a number of pitfalls that recommend against its use as a sole assessment method. Second, cognitive interviewing is described, since it is increasingly recognized as a crucial pre-testing stage before surveys go into the field within particular socio-cultural settings. We suggest that the rationale in favour of this approach be extended to support the use of cognitive interviewing to assess translated questionnaires. Another method for determining comprehension and meaning attributed to items involves focus group evaluation with members of the target population. This assessment approach is likely to be better than no pre-testing of the survey instrument; however, the information from cognitive interviews may be of higher quality because it better approximates the dyadic interplay of survey administration than do focus group dynamics. Finally, another promising tool for assessing respondent cognitions related to translated questions is behavioural coding, a technique which codes respondent and/or interviewer reactions to questions in recorded interviews to identify problematic survey questions (Fowler, 1995; Van der Zouwen & Smit, 2004; Johnson *et al.*, 2006). Overall, we emphasise the importance of translation assessment and pre-testing as a means of ensuring sound measurement properties of the target language survey instruments.

Back-translation:

Back-translation is often mistaken as a method of translation, but it is actually a method for assessing the quality of a translation that has already been made into a target language (Harkness, 2003). It involves independent translation of the target language questionnaire back into the source language and comparing the result with the original source language questionnaire. Back-translation presumes that the greater the similarity between the results, the more acceptable the translation (Brislin, 1970). However, languages are not isomorphic, and an unnatural sounding or even incomprehensible target language translation may produce, or even be necessary for, a "good" back-translation. Although back- translation may reveal some problems with target translations, it does not adequately assess the translated questions' comprehensibility within the target population (Harkness & Schoua-Glusberg, 1998; Harkness, 2003). Furthermore, the methodology provides no guidance about what qualifies as an acceptable level of similarity across the source and back-translated versions. Finally, when a back-translated questionnaire depends on a single translator for the "forward" translation into the target language—as it often does—it neither opens up the translation process to critical scrutiny nor does it produce the range of translation options that are found in team approaches. These factors recommend against the use of back-translation as the only method of translation assessment. Translation quality also needs to be evaluated in a more direct fashion.

An example provided earlier helps illustrate these concerns. The German General Social Survey item "Das leben en vollen zügen genieβen" literally translates to English as "Enjoy life in full trains." This translation is readily back-translated to and reproduces with fidelity the original German source language phrase. However, the nonsensical nature of the English translation could go undetected without further review. Moreover, an appropriate British adaptation of this phrase ("Live life to the full") would sound awkward in American English, for which different wording would be necessary (i.e. "Live life to the

fullest."). Such nuances would be missed, and in fact be discouraged, with back-translation that did not entail further review by bilinguals (Harkness, 2003).

Cognitive interviewing:

Cognitive interviewing is increasingly used to pre-test and thereby improve comprehension and related measurement properties of questionnaires within particular societies (Willis, 2005). The rationale for and principles that orient this practice should extend to assessment of translated questionnaires. In the absence of such pre-testing, there is no guarantee that the target language instrument will have sound measurement properties, even when the instrument has been pre-tested in the source language and best practices have been followed when translating it (Harkness *et al.*, 2003b). We describe a few basic principles of cognitive interviewing, while referencing key works for readers who are interested in more detail.

Cognitive interviewing follows from research on the cognitive processes involved in responding to survey questions (Willis, 2005). The response process generally involves question comprehension (i.e. meaning of terms and perceived intent of question), retrieval from memory (i.e. availability of and strategies to access relevant information), judgment processes (i.e. motivation to respond and to respond truthfully) and mapping the internally generated response to the question onto the response categories provided. As each step along this pathway may introduce measurement error, cognitive interview techniques focus on these aspects of the recall process.

The "think aloud" and "verbal report" protocols generally involve asking participants to openly describe the stream of thought in which they engage as they answer a survey question (Ericsson & Simon, 1984; Conrad & Blair, 2004). Responses are usually audio-recorded and transcribed for analysis. Advantages of the method include the minimal training requirements for the interviewer, whose main task is simply to read the question and listen. This generally passive interviewer stance may result in lesser bias than more pro-active methods. However, although the open-ended format of this approach may allow unanticipated response issues to emerge, subjects may need to be trained to think aloud, with some people unable to develop the skills necessary to provide useful feedback. Even "good" participants wander off track, thinking in ways that may only vaguely correspond with the mental processes required to respond to the question under normal circumstances (Willis, 2005).

Verbal probing techniques are increasingly favoured over think-aloud strategies in cognitive interviews (Willis, 2004, 2005). Probes have been developed in accordance with principles of sound question design, with specific probes used to uncover specific processing issues (see Table 2.3). An interview protocol is generally developed to anticipate which kinds of probes, if any, will be necessary for each question. However, the interviewer may also freely employ probes to address issues that unexpectedly emerge during the course of an interview. As such, the use of verbal probes demands the active involvement and training of the interviewer. However, training is less of an issue for the survey respondent than in the think-aloud. Probes may nevertheless influence respondents in ways that do not adequately reflect cognitive processes under "real" survey conditions. In particular, care must be taken to develop unbiased, neutral probes that do not lead participants to respond in particular ways.

When addressing survey instruments within particular socio-cultural settings, Willis (2005) recommends that each round of cognitive interviews involve survey administration among 8 to 12 people from the target population. At least two testing rounds are necessary to assess the adequacy of the original questionnaire as well as changes that result from the first round. Although the number of testing rounds will depend on the quality of the original instrument and the proposed revisions, Willis suggests that there are likely to be diminishing returns after three rounds of testing. This may or may not be the case in dealing with more complicated cross-cultural issues that involve translated questionnaires, where each round

READING: Is it difficult for interviewers to read the question in the same way to all respondents?
- What to read: interviewer may have difficulty determining what parts of the question to read
- Missing information: information that the interviewer needs to administer the question is not provided
- How to read: question is not fully scripted and therefore difficult to understand

INSTRUCTIONS: Look for problems with any introductions, instructions or explanations from the respondents' point of view
- Conflicting or inaccurate instructions, introductions or explanations
- Complicated instructions, introductions or explanations

CLARITY: Identify problems with communicating question intent or meaning to the respondent
- Wording: question is lengthy, awkward, ungrammatical or contains complicated syntax
- Technical terms: terms undefined, unclear or complex
- Vague: multiple ways to interpret the question or to decide what is to be included or excluded
- Reference periods: missing, not well specified, or in conflict

ASSUMPTIONS: Determine problems with the assumptions made or underlying logic
- Inappropriate assumptions are made about the respondent or about his/her living situation
- Assumes constant behaviour or experience for situations that vary
- Double-barrelled: contains more than one implicit question

KNOWLEDGE/MEMORY: Check whether respondents are likely to or not know or have trouble remembering information
- Knowledge may not exist: respondent is unlikely to know the answer to a factual question
- Attitude may not exist: respondent is unlikely to have formed an attitude about the argument being asked about
- Recall failure: respondent may not remember the information asked for
- Computation problem: the question requires a difficult mental calculation

SENSITIVITY/BIAS: Assess questions for sensitive nature or wording and for bias
- Sensitive content (general): the question asks about a topic that is embarrassing, very private, or that involves illegal behaviour
- Sensitive wording (specific): given that the general topic is sensitive, the wording should be improved to minimize sensitivity
- Socially acceptable: a socially desirable response is implied by the question

RESPONSE CATEGORIES: Assess the adequacy of the range of options
- Open-ended question: is inappropriate or difficult to answer without categories to guide
- Mismatch: question does not match response categories
- Technical terms: are undefined, unclear or complex
- Vague: responses categories are subject to multiple interpretations
- Overlapping: categories are not mutually exclusive
- Missing: some eligible responses are not included
- Illogical order: order not intuitive

ORDERING OR CONTEXT problems across questions

Table 2.3 Questionnaire Design Issues, from Willis (2005)

Adapted from Willis & Lessler (1999) and Willis (2005)

would be followed by efforts to coordinate and translate questionnaire changes until any cross-group discrepancies in question interpretation and comprehension appear to be resolved.

Where equivalence of meaning cannot be achieved, researchers should document why, and make sure this documentation is accessible to those who will ultimately analyse the data. Researchers who use the data at a later date may otherwise believe that the questions are equivalent and make invalid comparisons across cultural groups. Drawing from the previous example regarding the "vice" connotation of "addiction" in Mexico (see page 62), it may be inappropriate to compare Mexican smokers' and smokers from other countries on the item "tobacco is addictive" if the dominant meaning of addiction is compulsive behaviour in other countries. This situation could be documented by describing how "addiction" in Mexico appears to more strongly connote vice and less strongly denote compulsion than in other countries.

Cognitive interviewing example:

One recent example of cognitive interviewing to pre-test translated items involved the Spanish version of the Adult Tobacco Survey (ATS) for the United States' National Center for Health Statistics and the Office on Smoking and Health at the Centers for Disease Control and Prevention. The goal was to produce a Spanish-language version of the ATS questionnaire that was equally comprehensible and that shared the same meaning among Latinos in the US who speak different national varieties or dialects of Spanish. In the first step, a committee approach was used involving independent, parallel translations by bilingual translators of Mexican, Puerto Rican and South American heritage. This was followed by two rounds of cognitive interviews with Latinos from nine countries and Puerto Rico. The first round involved 40 participants using "think-alouds" after every question. In the second round, the resulting survey was administered in normal fashion to 28 participants, followed by a debriefing that targeted particular comprehension issues.

One of the many issues that came up concerned the translation of the often-asked English-language question, "Have you smoked 100 or more cigarettes in your life." Participants repeatedly thought that this question referred to daily smoking, even after the word "entire" was inserted to read "in your entire life" (*en toda su vida*) and the phrase was printed in boldface type to ensure its emphasis by survey administrators. This underscores the point that modification of a question may not resolve the problem, hence modified versions should also be pre-tested (Forsyth *et al.*, 2004). To resolve the issue, an introductory phrase was added to both the English and Spanish language questions: "For this question, we want you to think of all the cigarettes you ever smoked in your whole life, not on a single day." In this case, changes made to the Spanish-language items meant re-evaluating and changing the wording of the original, English-language version in order to reinforce equivalence. Anecdotal evidence suggests that similar comprehension problems characterised the original English-language version, so the addition of this introductory phrase may have improved comprehension across languages.

Quantitative assessment of measurement properties and systematic measurement error

Despite all precautions to ensure item equivalence across social-cultural groups and linguistic variants of a questionnaire, some unaccounted-for factor may nonetheless systematically and differentially influence responses provided by the groups under consideration. The strategies described here are best employed after collecting pilot data, but before implementing the full survey. Results can be used to eliminate, change or replace items that appear to be biased. However, these methods can also be used to assess measurement equivalence after survey data are collected, with the drawback that it is too late to change items with poor measurement qualities. As has been

emphasised when addressing other measurement equivalence issues described in this section, it is recommended that such issues be documented so that others who use the data at a later date will be aware of these issues.

Three approaches are briefly described here: single indicators, "alternative indicators" and latent variable Structural Equation Modeling (SEM). When multiple indicators of a construct are used, more statistical means are available to try to rule out systematic measurement error across groups. However, some approaches demand that single constructs be measured with a large number of items, which makes them less applicable to survey research. These methods, such as multi-trait multi-method (Saris, 2003a), multi-dimensional scaling (Fonatine, 2003), and item response theory approaches (Saris, 2003b) are detailed elsewhere.

Single-item measures of constructs:

When a single item is used to measure a construct, it may be difficult to assess whether observed similarities or differences in the measure are valid or whether these observations result from some other nuisance factor. Differential patterns of item non-response or "do not know" may indicate non-equivalence. Indeed, these non-random patterns violate assumptions that are necessary when dealing with this issue through pairwise or listwise deletion, as well as when using multiple imputation techniques (Groves, 2001). Nevertheless, theory and previous empirical findings can be drawn upon in order to predict how the indicator should correlate with other variables. In other words, expected correlations with other particular variables provide evidence of convergent validity. The absence of such correlations does not necessarily disprove the validity of the measure, however. Rather than disconfirming the validity of the measure, this lack of correlation may instead merely indicate the inadequacy or general inapplicability of the theory. Indeed, even when the measure under consideration is correlated with a set of theoretically related variables, this merely provides evidence — not confirmation—of the measure's convergent validity; systematic measurement error across the theoretical set of variables may still bias group comparisons.

Alternative measures of the same construct:

When there are multiple indicators of a particular construct, differential item functioning across cultural groups can be assessed by alternatively considering each indicator (Bollen *et al*., 1993; Smith, 2004a). With two items, a relatively clear indication involves consistent results for group differences in means (e.g. both higher in one group versus another) and in correlations with other constructs (e.g. number of days and number of cigarettes per day correlated with addiction). If the two indicators show inconsistent results, then strong claims about either result will depend on one's ability to convincingly argue for the use of one indicator over another. Although such post-hoc argumentation may be suspect, it can also establish the focus for subsequent research to clarify measurement and the interpretations that result. With three alternative indicators of the same construct, results from the third indicator can tip the balance in favour of the "preponderance of evidence." Consistency across all three indicators provides relatively strong confirmation of the validity of the results. Smith suggests that the most robust evidence will come from consistent results across alternative indicators that not only contain linguistically different stimuli, but that also have different response formats (Smith, 2004a).

Simultaneous assessment of multiple indicators:

Data collection on multiple indicators of the same construct also allows for statistical assessment of all indicators simultaneously, instead of the sequential format outlined above. Simultaneous consideration of multiple indicators lessens the impact of idiosyncratic, and therefore problematic, indicators (Bollen, 1989; Bollen *et al*., 1993). It also allows for the application of more formal statistical procedures to test, improve and attempt to equalise construct measurement properties across groups.

Exploratory factor analysis (EFA) techniques can provide evidence for the equivalence of construct dimensionality and discrimination across groups, although special techniques are often necessary to ensure adequate comparison (Van de Vijver & Leung, 1997). Items may be considered for elimination if substantial group differences are found for factor loading values on the same dimension or for the extent of cross-loading across dimensions. Cronbach's alpha may also be used to determine group differences in inter-item reliability. Although some statistics are available for evaluating factorial agreement across groups, the sampling distributions for these statistics are unknown, hence there are no statistical means of testing for what counts as an unacceptable difference (Van de Vijver, 2003). Moreover, these techniques generally assume normally distributed, continuous variables, and survey indicators often violate these assumptions.

Latent variable structural equation modelling (SEM) offers a more direct means of testing invariance of construct parameters and measurement properties across groups (Bollen, 1989, 2002; Joreskog & Sorbom, 1996). As with EFA, the dimensionality of different concepts can be examined. However, a key advantage of SEM concerns the ability to use statistical tests of construct parameter equivalence across groups. Moreover, whereas factor analysis parameter estimates assume continuous, normally distributed indicators, SEM allows estimation using non-normally distributed categorical and ordinal indicators (Joreskog & Sorbom, 1996; Muthen & Muthen, 2004). SEM techniques estimate items' unique weighted contributions toward the measurement of latent variables. EFA, on the other hand, involves summing or averaging variables that comprise a particular dimension, treating each indicator as equally weighted. Finally, several SEM packages now adjust for study design effects and sampling weights—adjustments that are often important in generating reliable, unbiased estimates in cross-cultural survey research. Taken as a whole, these key advantages recommend SEM methods over standard EFA techniques. Cepeda-Benito and colleagues (Cepeda-Benito *et al.*, 2004) provide a recent example of the use of these models to compare the structure of the Questionnaire of Smoking Urges survey instrument across samples of American and Spanish smokers.

Summary and Recommendations

Evaluation of tobacco control policies and other population-level interventions often involves data collection efforts across diverse national, cultural, linguistic and social groups. Comparison across such groups is often necessary to clarify policy effects, how these effects happen, and how effects might differ across populations. The literature discussed in this section suggests that these comparative studies should consider measurement equivalence issues in the following ways:

- Research teams should include collaborators from the socio-cultural groups in which the study is being conducted in order to help anticipate issues regarding the comparability of the theoretical framework, constructs and the measurement of these constructs across groups. When research involves participants from distinct language groups, at least one, and preferably more, team members should be fluent in the source language and the target language in which the survey will be administered.
- Whenever possible, it is recommended to use measures that have been appropriately validated for the populations in which the questionnaire will be administered. Even when a measure has been validated within one population group, its validity may not extend to other groups, and additional steps may be necessary to increase validity and improve the value of comparisons across groups.
- Translation of questionnaire items from one language to another should involve experienced translators. Review and adjudication of multiple, independent translations of the same items is currently considered the gold standard. If only one person translates

the questionnaire, then translation review should involve a group of bilingual people who are knowledgeable of questionnaire design principles and of key study concepts. Translation assessment should not merely consist of backtranslation.

- Researchers should carefully select and translate items with the goal of achieving equivalence of construct meaning across study populations. In some cases, literal translation of a questionnaire item across linguistic variants of the survey will not adequately capture the construct of interest, and more flexible translation and adaptation of the question will be necessary.
- All surveys, not just those that are translated, should be pretested to assess comprehension issues among the populations in which the survey will be administered. Ideally, pre-testing would involve cognitive interviewing before a survey is fielded. Cognitive interviewing or other pre-testing methods may also be used post-hoc to increase the validity of comparisons or to determine whether inconsistent results may be due to differential question comprehension.

  Researchers should consider and seek solutions to minimise the ways in which culturally moderated response factors (e.g. social desirability, acquiescence, extreme responding) may influence responses.

  Researchers should document decisions related to measurement development and item wording, especially where conceptual equivalence is suspect, translation is difficult, or where cognitive interviewing or other pre-testing methods reveal systematic differences in meaning. Researchers should also document issues around survey administration.