1-1-2009

Agreement answer scale design for multilingual surveys: Effects of translation-related changes in verbal labels on response styles and response distributions

Ana Villar
Stanford University, anavillarc@gmail.com

AGREEMENT ANSWER SCALE DESIGN FOR MULTILINGUAL SURVEYS: EFFECTS OF TRANSLATION-RELATED CHANGES IN VERBAL LABELS ON RESPONSE STYLES AND RESPONSE DISTRIBUTIONS

by

Ana Villar

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For The degree of Doctor of Philosophy

Major: Survey Research and Methodology

Under the Supervision of Professor Janet A. Harkness

Lincoln, Nebraska

December, 2009

AGREEMENT ANSWER SCALE DESIGN FOR MULTILINGUAL SURVEYS: EFFECTS OF TRANSLATION-RELATED CHANGES IN VERBAL LABELS ON RESPONSE STYLES AND RESPONSE DISTRIBUTIONS

Ana Villar, Ph. D.

University of Nebraska, 2009

Adviser: Janet A. Harkness

Answer scales in survey instruments are widely used, but little is known about how to choose verbal descriptors as labels. In multilingual research, this matter is further complicated because answer scales must be appropriate for all languages and function comparatively. Comparing source questionnaires to translations of multinational projects (e.g., the World Values Survey, the European Social Survey), it was observed that certain verbal features differed across languages, countries, and modules. This dissertation empirically investigates the effect of such changes on response distributions. The verbal feature examined is the presence or absence of an intensity modifier in the second and fourth labels of a 5-point agreement scale: *strongly agree, (somewhat) agree, neither/nor, (somewhat) disagree, strongly disagree*.

Two studies are conducted analyzing data from more than 40 countries of the International Social Survey Programme. The first investigates whether two methodological features (scale translation and administration mode) affect cross-cultural differences in three response styles. It was expected that adding an intensity modifier would increase extreme response style, decrease middle response style, and not affect acquiescence. Acquiescence was expected to be higher in interviewer-administered than in self-administered surveys. Using multilevel models in five ISSP modules, the analyses show that, as predicted, adding an intensity modifier results in higher extreme response style, and does not affect acquiescence. No support was found for the hypothesized effect on middle response style. Partial support was found for the hypothesized effect of data collection mode.

The second study investigates the effect of adding the intensity modifier on the use of each response category. Using an 8-item attitudinal scale, Graded Response Models for each of

the answer scale versions were compared. For the answer scale under study, adding an intensity modifier made the scale points less useful to measure the underlying attitude than were scale points without modifier.

These findings suggest that modifications made to answer scale language versions are a critical source of variability in response patterns and distributions.

Item bias statistics should be routinely reported in cross-cultural studies. There is no rationale for the common practice of taking observed cross-cultural differences at face value without a check of item accuracy.

Van de Vijver and Leung, 1997, p. 84

Acknowledgements

First and foremost, I would like to thank my supervisory committee for generously giving their time to guide me and accompany me in this process. Both the dissertation and I greatly benefited from their input. The dissertation would not have happened without the guidance of Janet Harkness, my committee chair, who mentored me, supported me, and encouraged me in particularly difficult moments. She devoted long hours to discussing and challenging the findings, and also to go through loops that were new for us both. Words cannot describe my gratitude for her dedication.

Two professors stand out among the great ensemble of professors whose courses I had the pleasure to take: Lesa Hoffman's class at UNL, together with all the tools she makes publicly available, were an *essential* resource in the process of analyzing the data. I deeply admire her dedication to teaching. Kristen Olson's course was incredibly inspiring and provided me with very useful research tools.

I have been lucky to be able to share an interest for the various methodological aspects discussed in this dissertation with fabulous researchers. I would like to thank Annelies Blom, Michael Braun, Tzu-Yun Chin, Rory Fitzgerald, Andrés González, Peter Mohler, Alisú Schoua-Glusberg, Fons van de Vijver, Yongwei Yang, and many other researchers at the Comparative Survey Design and Implementation international workshop for their helpful input.

Having great colleagues at the Gallup Research Center helped in so many ways. Their insightful comments were inspiring, their hard work contagious, and their support invaluable. Special thanks to Debbie, Emilio, Ipek, Jennie, Kathleen, René, and Yelena.

Thanks to all the Lincoln family for the fantastic moments of fun and friendship. They are, alphabetically: Anne, Emanuele, Erika, Eva, Guillermo, José, Kristi, Manolo, Matías, Mónica, and Trisha. Special thanks to Kristi for hosting me so many times!

I would have failed to get through the most stressful moments without the emotional support of Ana Cris Eiriz. She knows how deeply thankful I am of having her as my friend.

I thank Mario for his confidence in my research, his infinite patience, and how he took care of *everything* while I was writing day and night. Needless to say that his help was crucial from the beginning of the process, back in 2005, until the very last day.

And thanks to the best family one could hope for, from the youngest toddler to those who passed away while I was in the midst of preparing my comps and to whom I could not say good bye… their support and faith have taken me further and higher than I ever dreamt of.

Table of Contents

List of Tables

List of Figures

Introduction

Cross-cultural research is becoming increasingly important and prevalent. In the past few decades, new cross-national efforts have been emerging in the social sciences with the aim of comparing nations and their societies, and of monitoring these differences. The International Social Survey Programme—initiated in 1984, the European Social Survey—first fielded in 2002, and the World Mental Health Survey Initiative—officially established in 1998, are only a few recent examples of projects that involve more than twenty countries. In addition, populations of countries of a traditionally homogeneous composition are becoming more diverse. The United States is a clear example of a country with increasing cultural diversity. Data from the 2008 Current Population Survey show that in the U.S. context almost 20 percent of the population select a race other than white. The Hispanic population has increased from 12.5 percent to 15 percent in the past decade, and the Asian population has reached 13 million[1]. This trend is more pronounced in states such as California, Florida and Texas. However, large populations of Hispanics are also found in states traditionally considered to be monocultural. For example, Nebraska has seen a steady increase in the Hispanic population, from 2.3 percent in 1990 (U.S. Census Bureau, 1991) to 5.5 percent in 2000 (U.S. Census Bureau, 2001), and 7.3 percent in 2007 (U.S. Census Bureau, 2009).

Together with this shift in cultural composition, there has been an increase in the proportion of households where English is not the first language. Of particular concern

[1] http://www.census.gov/cps/

for household surveys is the increase of linguistically isolated households, where none of the household members 14 or older speaks English "very well" (Siegel, Martin, & Bruno, 2001). If not approached in a language they speak, this household could be a sample unit from which no information is gathered. Language coverage has thus become an issue in surveying the American population, and its impact on response rates and response bias is yet understudied. Multilingual surveys are, as a consequence, increasing in number and importance in within-country research and cross-national research (Harkness, 2008). In the United States, major survey efforts such as the California Health Interview Survey, the National Health Interview Survey, the Decennial Census and the General Social Survey reflect a growing tendency to attempt to interview respondents who primarily speak languages other than English. The 2000 Census form, for example, was available in six different languages (English, Spanish, Chinese, Vietnamese, Korean and Russian) (Siegel, et al., 2001). Assistance was provided in 44 other languages (Whitworth, 2001). In the United States, survey interviews in Spanish are becoming more common in federal and state commissioned work, and private companies are following suit (see, for example, Wivagg & Santos, 2006). Understandably, in a survey context of declining response rates, including non-English speakers helps reduce survey nonresponse due to language barriers[2].

Irrespective of whether the motivation is to increase representativeness or to compare populations, numerous problems arise in surveying people from different cultures, whether they are sampled from one country or from several countries. Surveying diverse populations has implications for all stages of the survey cycle, and involves all

[2] According to AAPOR Standard Definitions, survey cases that are not interviewed because of "language problems" receive the disposition code "eligible, no interview", and therefore constitute nonresponse (American Association for Public Opinion Research, 2008).

sources of survey error. First, high survey quality needs to be attained for all collected data (Jowell, 1998), but expertise and research capacity may differ greatly across locations (CSDI, 2009). Second, comparison of survey estimates from different populations presumes equivalence of those estimates. This involves not only equivalence regarding how the constructs are measured in each group, but also the impact that essential survey conditions have on all sources of error for each group. The strategy that most cross-cultural survey research follows in striving for equivalence is to replicate features of survey design in the best possible way across countries and cultures (Harkness, Mohler, & van de Vijver, 2003, p. 8), from sampling design to data processing. This means that all countries involved are expected to implement the same procedures and follow the same methods. Deliberate adaptation, as opposed to unintended variation, of the procedures and strategies to particular cultures is minimal and usually introduced *ad hoc* (Harkness, et al., 2003). However, it has long been realized that equivalence is not guaranteed just by using identical procedures and methods across groups (Braun, 2003; Scheuch, 1968; Verba, 1971). Keeping the same mode of data collection, for example, could be detrimental to data quality and comparability. In regions where telephone penetration is low, coverage error will be more prominent than in other regions where penetration is higher. Couper and de Leeuw (2003), moreover, discuss how nonresponse mechanisms differ across countries. In countries where nonresponse is mostly due to noncontacts, for example, offering an incentive would not be as effective as increasing the number of interview attempts. Therefore, response enhancement strategies need to be tailored to the particular nonresponse makeup of each country.

This dissertation and the research it describes focus on measurement error in multilingual contexts. In particular, it concentrates on errors arising from two aspects present in multilingual surveys. The first is translation and adaptation of answer scale labels. When surveys involve more than one language group, researchers must decide what approach to follow in designing questionnaires. Measurement error related to linguistic issues needs to be minimized and comparability maximized in spite of the issues. The available body of research to guide such decisions is scant, both in multilingual (Harkness, 2003) and in monolingual questionnaire design. This dissertation intends to contribute to this body of research. In particular, I hope that the reader will feel compelled to pay close attention to how answer scales are/have been developed both in monolingual and multilingual settings, as well as to assess the potential impact of seemingly small changes in how these are labeled. The second aspect relevant to multilingual surveys refers to cross-cultural differences in response styles, that is, the possible influence of culture specific tendencies unrelated to content on how respondents select a response category from a rating scale.

Structure of the Dissertation

The dissertation is organized in six chapters. Chapter 1 presents a discussion of the literature on response styles and the methods that have been used to study this phenomenon. Chapter 2 reviews the available literature on answer scale verbal labels. Chapter 3 presents a brief description of the methodological aspects of the data used in this dissertation. Chapter 4 describes the findings regarding country differences in response styles, and Chapter 5 deals with the effect of changes due to translation in

answer scale verbal labels on response distributions. Chapter 6 summarizes and discusses the main findings in this dissertation.

Chapter 1. Response Styles in the Context of Cross-Cultural Research

Researchers conducting studies that involve different cultures or societies meet numerous challenges. Some of these challenges are related to whether the concepts of interest exist in the cultures under study, and whether the concepts are similarly described in each culture. If the concepts are fundamentally different across cultures it may not be advisable to compare the cultures at all. Yet other challenges are related to the methods and procedures used to collect survey data from each culture. Even when trying to keep all methodological aspects the same across all populations, adjustments are likely to be necessary. For example, instruments need to be rendered in languages understood by most of the individuals in the target population; interviewers need to be selected and trained in accordance to cultural norms and laws that govern conduct in the target cultures and societies; consent from different parties may need to be obtained before an individual can be interviewed.

The use of questionnaires is one of the aspects that may in itself be a source of inequivalence. Cultures differ in how familiar they are with survey instruments and with an interview style using close-ended questions. Cultures may thus differ in how they behave in such communication context. Indeed, differences in response patterns across cultures have been documented that raise concerns for data comparability. Therefore, this dissertation looked at response styles as manifestations of cultural differences in the survey interview.

This chapter reviews the concept of response styles as reflected in the literature, including the most frequently discussed forms of response style and the explanatory variables that have been proposed for these. The chapter will also review the literature reporting cross-cultural differences in response styles and it aims to identify the role of the methodological factors of surveys (e.g., answer scale design features) in studying response styles.

Response Styles

Researchers who set out to measure attitudes by means of self-reports embrace this technique with the hope that the resulting measures will be reliable and valid. At the same time, random and systematic sources of error may taint the statistics of interest. Response styles are one of the forms of bias that have been frequently documented for studies using self-reported measures (Baumgartner & Steenkamp, 2006; van de Vijver & Leung, 1997). The definition of response styles (or response sets) is to a large extent widely accepted as the systematic tendency to choose certain portions of a rating scale irrespective of the content of the item (Cronbach, 1946, 1950). However, there is no clear-cut distinction between the terms *response style* and *response set*. Baumgartner and Steenkamp (2001) noted that some authors use response styles to refer to stable tendencies inherent to the respondent, and response sets to define effects moderated by the context or the instrument itself. There is neither a generally accepted terminology nor agreement on a distinction between response styles and response sets. Throughout this dissertation no distinction is made between response styles and response sets. Yang,

Harkness, Chun, and Villar (in press) used the term "displayed response" to refer to patterns of response behavior actually chosen by respondents in answering items with a similar answer scale format. This term allows them to refer to observed preferences in the use of rating scale positions without deciding *a priori* whether the pattern is driven by an individual pervasive preference, or is an artifact of the measurement instrument, or the result of some other factor. Even though this is the preferred term here, in this text, response styles and response sets are used also, because that is the prevalent terminology in the literature.

In addition, the existing body of literature is sometimes contradictory in terms of explanations offered for responding behavior observed across a number of items or questions. Explanations vary with regard to the importance given to respondent, context or instrument. The psychological and psychometric traditions have focused on personality and other individual level variables (e.g., anxiety, cognitive development, gender, or age). Survey research has also focused on individual variables (e.g., education, social status) as well as question wording features (e.g., presence of a middle point or whether the answer scale is construct-specific or an agree/disagree scale). Cross-cultural research has focused on "cultural explanations" of response styles (e.g., child-rearing practices, or modesty norms).

There are numerous challenges in the study of response styles. Firstly, the ambiguity involved in the expression "irrespective of the content of the item" has not been discussed in the literature but deserves some attention. Different authors endorsing definitions that include this phrase seem to have somewhat different understandings of what it entails. Some publications seem to imply that an individual more likely to

acquiesce than another will do so in any given context, with any given measurement instrument. Less absolute positions are generally sustained, acknowledging the role of moderating variables on estimates of response style (e.g., Cronbach, 1946; Culpepper, Zhao, & Lowery, 2002).

What the expression “irrespective of content” generally seems to be taken to mean is that, regardless of whether two respondents have the same underlying hypothetical “true” value, they will choose different answers to the same question because of their inner tendencies to prefer one scale point to another as descriptor. The researcher, of course, would want respondents with the same “true” value to choose the same response option; therefore, this kind of mismatch is usually seen as measurement error.

Given the terminology used to describe some of the response styles (“extreme response style”, “middle response style”) the implicit assumption seems to be that the individual preference is driven by the location or position of the category relative to the full answer scale (extremes vs. non extreme, middle vs. non middle). In turn, this suggests that “irrespective of the meaning of the item” extends also to the content of the answer scales provided. However, there is little evidence of stability of response styles outside personality research. Furthermore, research has shown that various features of answer scales have an impact on response patterns that are regarded as response styles. If patterns regarded as response styles depend on methodological features of the instrument, this can be taken as evidence that response styles are less pervasive than assumed.

Early research on response styles emerged from studies in educational psychology (e.g., Cronbach, 1946, 1950) and personality psychology (e.g., Couch & Keniston, 1960; Jackson & Messick, 1958; Lentz, 1938) in the United States. Lentz (1938) expressed

concern about the possibility of acquiescence bias distorting personality measures such as the Minnesota Multiphasic Personality Inventory (MMPI) or Adorno's authoritarianism F scale. Cronbach (1946, 1950) published two reviews that summarize evidence for a large number of response sets that were observed among individuals answering, variously, achievement tests, personality measures, attitudinal items, and psychophysical tasks. This type of response styles have also been discussed, to a lesser degree, in social survey research (e.g., Landsberger & Saavedra, 1967; O'Neill, 1967), business research (e.g., Harzing, 2006), and market research (e.g., Baumgartner & Steenkamp, 2006). They have also been discussed in cross-cultural research (e.g., Chun, Campbell, & Yoo, 1974; Johnson, Kulesa, Cho, & Shavitt, 2005; Triandis & Triandis, 1962; van Herk, Poortinga, & Verhallen, 2004).

Various kinds of response styles are proposed in the literature. Cronbach (1942, 1946, 1950) observed that some students were more likely than others a) to answer items they were unsure about, b) to choose "true" over "false" when they did not know the answer to the item, and c) to choose speed over accuracy (or vice versa), for example. Broen and Wirt (1958) and Messick (1968) both identified eleven kinds of response style, with considerable agreement in their classifications. However, the focus on survey research and attitudinal research has remained almost exclusively on acquiescence, extremity, and to a lesser extent on middle response style. Therefore, these are the response styles that will be discussed in more detail here.

Acquiescence

Acquiescent response style, or yea-saying as it is sometimes called, refers to the tendency to agree with statements irrespective of the content of the items, or alternatively

to prefer “true” or “yes” over negative responses. Some authors have referred to this response style as directional bias (Hui & Triandis, 1985) or positivity bias (Baumgartner & Steenkamp, 2001). The discussion of directional bias includes a wider range of bipolar rating scales, beyond that of agree/disagree, yes/no or true/false scales. It is not clear whether theoretical explanations applicable to the agree/disagree format may be generalized to other scales (e.g. satisfied/dissatisfied, very likely/very unlikely). The focus here will be on the agree/disagree answer scales.

Several explanations have been proposed for acquiescent responding. A considerable body of research describes acquiescence as a reflection of a personality trait (Cronbach, 1942; Hamilton, 1968). The argument behind this conception is that the tendency, for example to agree with statements, is stable and manifested in an individual’s behavior when answering a questionnaire. Evidence in favor of this argument comes from shown stability of response sets across time in test-retest studies (Cronbach, 1946; Messick, 1968) and from correlations with other factors of personality inventories (Couch & Keniston, 1960; Messick & Jackson, 1961). They all seem to agree that acquiescence may be a stable tendency and may correlate with substantive constructs. At the same time, the items used in personality research are selected because they show stability across time. Hui and Triandis (1985), using attitudinal data instead, found evidence for a lack of stability; they observed that the further apart in the questionnaire item batteries are, the lower the correlations among response sets computed from those batteries.

In survey research acquiescence is seen as a form of response bias. Two explanations are proposed. The first one is that acquiescence reflects deference that low-

status respondents show toward middle-class interviewers by not disagreeing with him/her (Carr, 1971; Lenski & Leggett, 1960). The second hypothesis sees low cognitive ability as the root of acquiescence (Campbell, Converse, Miller, & Strokes, 1960; Cronbach, 1942). The assumption behind the cognitive hypothesis is that respondents with lower education accept statements in a less critical way.

Early tests of the deference hypothesis show that respondents of low social status try to avoid disagreement with someone with higher social status. Lenski and Leggett (1960) compared the percentage of respondents that agreed with two items of pairs designed to hold logical contradictions across levels of social status. After controlling for education, they find that lower status respondents were nonetheless more likely to agree with both statements. However, Fisher (1974) found mixed results related to interviewer-respondent status differences using race/ethnicity as a proxy. When comparing agree responses given to Black or White interviewers, the deference pattern is replicated for some items, but not for others.

The evidence regarding cognitive ability is also inconclusive. Using educational level as a proxy for cognitive ability, Bachman and O'Malley (1984b) and Moors (2003, 2004, 2008) found no differences in acquiescence for different groups. However, Landsberger and Saavedra (1967), Light, Zax and Gardiner (1965), Marín, Gamba and Marín (1992) and Johnson et al. (1997) reported significant differences as a function of education in the proposed direction—namely, more acquiescence among those with lower education.

Krosnick (1991) and Narayan and Krosnick (1996) found higher acquiescence levels among the lower educated as well, and they provide a theoretical framework for

understanding these findings. Krosnick (1991) argued that the "reasonable" appearance of survey statements discourages respondents to actively search for reasons why the statements are *not* true. The commonly accepted models of survey response propose four groups of cognitive processes in survey response—comprehension, retrieval, judgment and reporting (Cannell, Miller, & Oksenberg, 1981; Sudman, Bradburn, & Schwarz, 1996; Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000). Optimal answering should undergo all four stages comprehensively and consequently calls for a great deal of cognitive effort (Krosnick, 1999). The high cognitive demands of this task may lower motivation, and respondents who were previously responding optimally might shift to "satisficing", that is, to providing a satisfactory, adequate, response rather than an optimal one (Krosnick, 1991). How satisficing is manifested depends on the decision heuristics a given respondent uses. Satisficing theory may therefore help explain why response patterns unrelated to content are formed.

Using more sophisticated proxies for cognitive ability and deference than previous research, Knowles and Nathan (1997) simultaneously tested both the educational and the deference hypotheses. The traditional proxy for cognitive ability and social status of the respondent with respect to the interviewer is self-reported level of education. Knowles and Nathan (1997) collected measures of personality that relate to "social concerns" (such as conformity or social participation) and "cognitive style" (e.g., complexity, breath of interest). They found that acquiescence was related to cognitive simplicity, rigid mental organization and intolerance of alternatives, but not to social desirability or other social concerns. This led them to conclude that acquiescence response style is more likely to be due to limitations in cognitive processing than to

concerns of self-presentation or impression management. Zhou and McClendon (1999) found a similar result using yet a different proxy for cognitive sophistication, i.e., the number of correct answers in a verbal test. The effect of education and race on acquiescence became nonsignificant after controlling for cognitive ability.

A different test of these hypotheses comes from research comparing modes of data collection. If the deference hypothesis is true, interviewer administered methods should lead to higher levels of acquiescence than self-administered methods (Dillman, Phelps, et al., 2009; Schuman & Presser, 1981). However, if the cognitive ability hypothesis holds, a different relationship should be found. Weijters, Schillewaert and Geuens (2008) found slightly higher levels of acquiescence in telephone interviews than self-administered questionnaires (both paper-and-pencil and online), supporting the deference hypothesis.

There is, however, an additional explanation that would predict higher acquiescence response style in interviewer administered modes than in self-administered. Drawing on findings from linguistic anthropology, Schaeffer (1991) postulated that habitual conversational norms of talking may contribute to the survey interaction. In ordinary conversations, questions and statements would be designed "for agreement" (Pomerantz, 1984; Sacks, 1987). In addition, the recipient of a statement can use various strategies to disagree without direct disagreement, such as using silences as well as agreement preceding disagreement: "Yes, I agree (pause) however…" Respondents may expect the same kind of interaction to take place in the survey interview context, but in standardized interviews interviewers are instructed to accept responses and not to negotiate reformulations towards "agreement". Such discourse analytical interpretation

could explain lower levels of disagreeing responses. However, without empirical evidence no generalizations can be made.

Although research on acquiescence bias is quite extensive, it is still inconclusive about what the causes of acquiescence are. There is conflicting evidence for the two competing hypotheses described, although the two studies that test them simultaneously Knowles and Nathan (1997) and Zhou and McClendon (1999) presented support for the cognitive ability hypothesis and not for the deference hypothesis.

Extreme response style

Extreme response style, or extreme responding, is understood as the tendency to choose the end points of rating scales. The first attempts to systematically study extreme response style explored group differences. Berg and Collier (1953) found higher levels of extreme response style for White females than for White males, and for Black males than for White males. Light et al. (1965) reported lower use of extremes for tenth-grade children than in younger ones, as well as for children with higher IQ scores; their findings do not substantiate the gender differences found by other authors. There is some evidence that extreme response style is negatively correlated with educational level (Marín, et al., 1992; Stening & Everett, 1984), and anxiety (Berg & Collier, 1953; Lewis & Taylor, 1955). However, other studies found no relationships, so that no clear picture is available to date.

There are also two main ways of interpreting the mechanisms proposed for an extreme response style: it is seen as a consequence of the application of social norms to

the “survey game” or alternatively as the result of psychological and cognitive aspects of the survey response process (Watkins, 1992).

A social norm or “cultural” explanation has been used to account for differences observed across different ethnic groups and different nationalities. Norms that apply to social interaction in a particular culture may apply to survey response as well. Zax and Takahashi (1967) argued that Japanese are less drawn by impulses to extreme positions than Americans and are more “capable of moderation” because of child rearing practices that emphasize restrain. Other cultural dimensions related to response styles are individualism/collectivism and power distance (Harzing, 2006; Johnson, et al., 2005). Modesty norms rooted in Confucian philosophy were also used to explain moderacy response style in Chinese managers (Culpepper, et al., 2002). The section on cross-cultural differences in response styles considers these explanations in further detail.

Cognitive variables have been studied in relation to extreme response styles; ambiguity tolerance and its counterpart need for certainty, for example, have been considered to affect extreme response style (Hamilton, 1968). The rationale is that respondents high in need for certainty or rigidity will tend to choose the extremes more often as a means to achieve greater degree of structure. Brengelman (1959, 1960) found the correlation of ambiguity tolerance and extreme response style to be between .28 and .45.

Other cognitive variables have been used to explain extreme response style. The concept of “meaningfulness” has also been used. In this line of research, meaningfulness seems to be understood as the extent to which an item is important to the respondent. Gibbons, Zellner and Rudek (1999) asked respondents “From your own personal

perspective, how meaningful is this statement to you?" and show positive correlations between meaningfulness scores and extremity. It is unclear what the above question might be actually measuring, but this finding may suggest that traditional measures of response styles are not as independent of the content of questions as definitions assume.

Extreme responding has also been related to the need to evaluate. People high in need to evaluate tend to evaluate the positive and negative aspects of objects and experiences they encounter (Jarvis & Petty, 1996). As a consequence, they are more likely to form attitudes about many objects, and they have shown to hold more extreme attitudes (Federico, 2004). Holbrook, Johnson and Cho (2006) found higher levels of extreme response style in respondents that score high in need to evaluate.

Much of the literature on extreme response style is focused on explaining cultural differences in extreme response style; therefore it will be explored in more detail in the section on cross-cultural differences in response styles.

Middle response style

Middle response style has many synonyms: central tendency, middling response bias, moderacy response style, or midpoint responding. Research on middle response style is considerably less extensive, possibly because it is sometimes understood as complementary to extreme responding. However, there is evidence that these concepts are not in fact complementary. Correlations between extreme response style and middle response style are only moderately negative. Baumgartner and Steenkamp (2001), for example, reported a correlation of -.55. Stening and Everett (1984) found that some populations may show high levels of both response styles. They find that Hong Kong

respondents showed high use of both the midpoint and endpoints of the scale. Previous analyses of the International Social Survey Programme (ISSP) data (Villar, 2006) also showed that even though Japanese tend to choose the middle point more often than most countries, they also select the endpoints more as exemplified in figures 1.1 and 1.2.

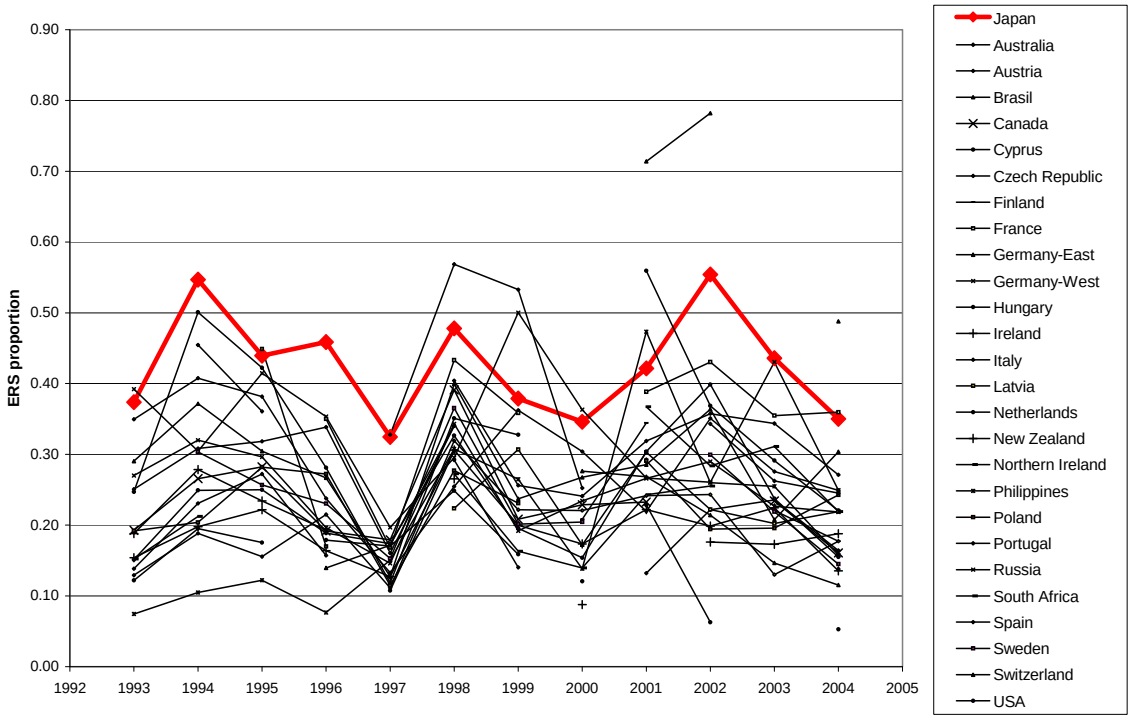

Figure 1.1. Proportion of endpoints from 1994 until 2004 for Japan

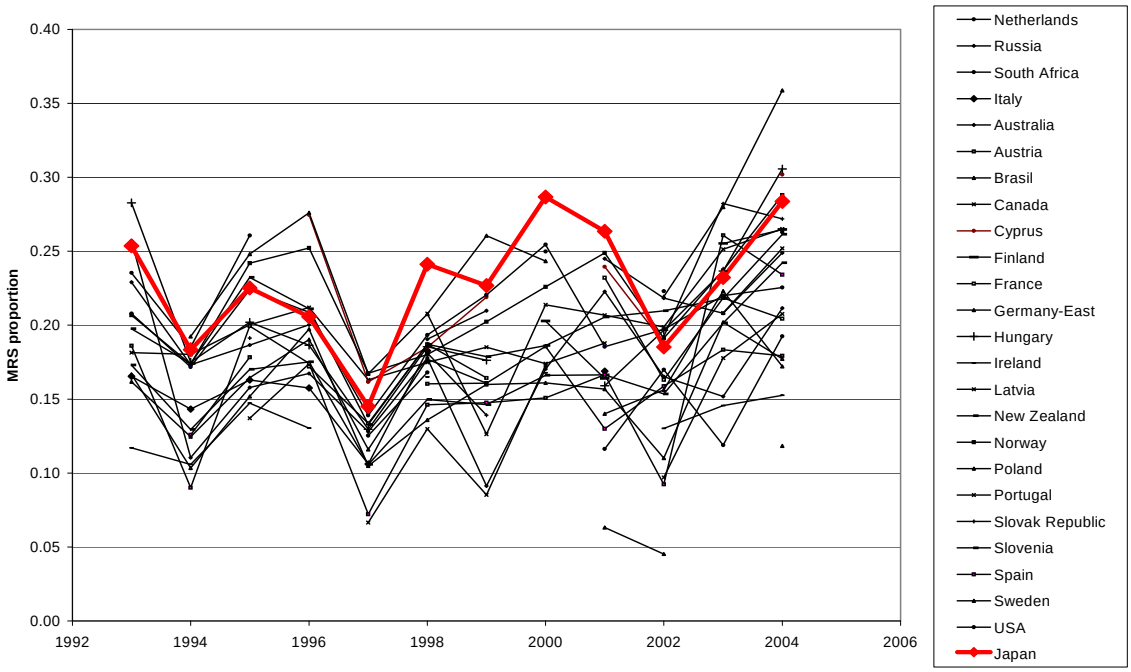

Figure 1.2. Proportion of middle points from 1994 until 2004 for Japan

Finally, if the two response styles were in fact complementary, correlates of one could be expected to be inversely related to the other. For example, given that education is negatively correlated with extreme response style, it could be expected to correlate positively with middle response style. However, Baumgartner and Steenkamp (2006) noted that Narayan and Krosnick (1996) found higher use of the middle alternative in respondents with lower education respondents than in highly educated ones.

Explanations for middle response style have emerged in trying to account for differences between respondents from Asia or with Asian heritage and respondents from Western countries (Chen, Lee, & Stevenson, 1995; Hamamura, Heine, & Paulhus, 2008; Stening & Everett, 1984). The main focus has been given to ambiguity tolerance (Brengelmann, 1959, 1960), uncertainty avoidance (Harzing, 2006) and the Confucian philosophy, that values moderation, modesty, and cautiousness (Chen, et al., 1995).

Response Styles Research

Main findings and covariates of response styles

In trying to explain how cultural differences in extreme response style emerge, Hui and Triandis (1989) presented a description of the cognitive mechanisms through which the different response categories are selected. The model is relevant because it provides a tentative description of the mechanisms behind the hypothesized response styles, and particularly of cross-cultural differences in response styles.

Hui and Triandis (1989) followed Wyer and Carlston's (1979) distinction between subjective categories of judgment—those that exist in the mind of the respondent, or that are generated in the process of coming up with an answer, and response categories—those that are offered by the researcher through the questionnaire.

The model resembles somewhat that presented by Parducci (1965) for judgment of stimuli. Parducci's range-frequency model is largely focused on judgment of physical properties, such as the length of an object. However, the principles described in the model have been applied to cognitive models of survey response (e.g., Tourangeau, et al., 2000). The mechanism Parducci describes involves using the most extreme stimuli as anchors that define the "psychological representation of the stimulus range" (p. 418)—the *range principle*. The remaining stimuli are then compared to those anchors and the distance between them is translated onto a distance in the response categories. In addition, the model predicts that respondents will tend to use the answer scale points with similar frequencies—the *frequency principle*. The body of response styles literature, however, presents evidence against the frequency principle as it would apply in

attitudinal research. Respondents do not seem to use response categories with similar frequencies when answering attitudinal questions. The model proposed by Hui and Triandis (1989) tries to accommodate these findings from the response styles literature.

In the process of translating a judgment into an offered response option, respondents must find a way to solve a mismatch when the number of subjective categories and response categories is not equal. Hui and Triandis (1989) argued that group differences in response styles stem from the way respondents map their subjective categories onto the response categories. When the number of response categories is lower than the number of subjective categories, the strategy the respondent follows determines the quality of the resulting data. Figure 1.3 exemplifies how respondents may match their subjective categories to the response options provided by the researcher (figures 1.3A to 1.3C). Hui and Triandis (1989) hypothesized that the observed differences between Hispanics and non Hispanics stem from differences in how they establish the connection between their subjective categories and the categories offered by the researcher. In figure 1.3A, the respondent is assigning equal number of subjective categories to the response options; comparatively, a respondent answering in a scheme like presented in figure 1.3C, would choose endpoints more often than a respondent using the scheme in figure 1.3A. Hui and Triandis offered no account of what would happen if the subjective categories were fewer than the ones offered by the researcher.

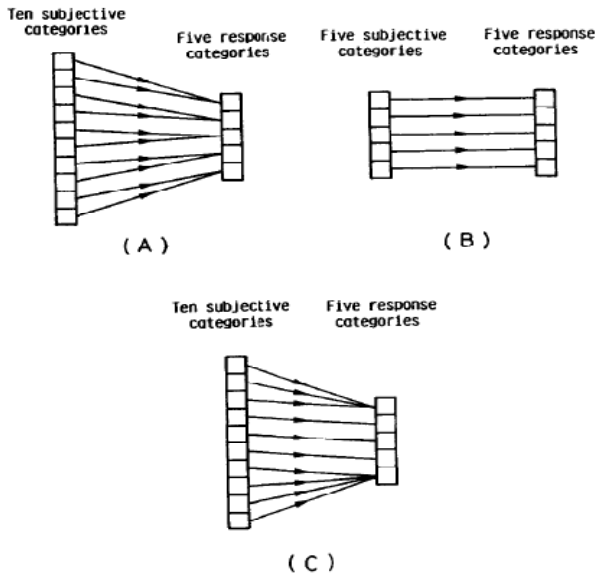

Figure 1.3. Mapping of Subjective Categories on Response Categories

(Hui & Triandis, 1989, p. 299)

Many studies repeatedly provide evidence for various response patterns, but so far there is little agreement on what the underlying mechanisms are. Empirical research directly investigating such mechanisms is scarce; the main focus of response styles research falls on three aspects: a) how response styles affect data quality, b) what factors predict and explain response styles, and c) how to deal with the biasing effect on the data.

a) During the 1960's and 1970's one focus of debate was whether response styles were indeed a threat to validity. Within psychological research, it was argued that response styles were not part of systematic measurement error but rather reflected personality characteristics such as agreeableness (Block, 1965; Couch & Keniston, 1960; Messick, 1968; Rorer, 1965). Acquiescence and extreme response style were indicators of personality traits and not errors in the measures. However, if people high

in agreeableness tend to choose the positive side of a scale regardless of item content, their measures will be to some extent contaminated with measurement error (Messick, 1991). The second focus of debate raised the question of how much the variables of interest were affected by response styles. Even though Block (1965) and Rorer (1965) argued that the actual impact on data quality was minimal, the great majority of publications seems to suggest that it is an important threat to measurement. Baumgartner and Steenkamp (2001) demonstrated that although the systematic effect of response styles on validity was not large, it was statistically significant. Moreover, differences in response styles across countries and cultures raised concerns about the interpretation of differences in comparative research (Chun, et al., 1974; Triandis & Triandis, 1962). Numerous studies have been devoted to exploring these differences. The section on cross-cultural differences in response styles will review the findings of these studies.

b) Different predictors have been studied depending on the response style being described. The most frequently used are sociodemographic variables and cultural explanations. Sociodemographic variables have been almost universally used to compare response patterns, with a general tendency showing that respondents with lower education (Marín, et al., 1992; Schuman & Presser, 1981), lower social status (Carr, 1971; Lenski & Leggett, 1960) and living in a rural environment (Arce-Ferrer, 2006) are more likely to exhibit response patterns that are believed to distort survey estimates. As described above, numerous cognitive and cultural variables have been

proposed as predictors of response styles, and they will be addressed in further detail in coming sections.

c) As with many other sources of error, two main approaches can be adopted in order to deal with these response styles: statistical control and instrument design. Statistical control of extreme response styles has been pursued by standardizing scores within subjects (van de Vijver & Leung, 1997), or by using the response style measure as controls in analyses in various ways, such as partial correlation, or the use of stylistic factors in structural equation models (Baumgartner & Steenkamp, 2006; Cheung & Rensvold, 2000; Welkenhuysen-Gybels, Billiet, & Cambre, 2003). A general criticism of these methods is that eliminating differences across respondents can reduce the impact of the error in the statistics of interest, but does not necessarily make the survey estimates for each group more valid.

Some of the methods for statistical control call for specific requirements for measurement of the response styles themselves. This typically affects the design of the questionnaire. In the case of acquiescence, for example, “balanced scales” are recommended (Lentz, 1938). These scales are sets of items with a similar number of “positive” and “negative” items. Positive items are those expected to correlate positively with the construct of interest. Systematically agreeing with two opposing views is seen as an indicator of acquiescence. The more a respondent agrees with pairs of items that are supposed to present contradicting beliefs, the higher their score on acquiescence. These scores are then used to adjust estimates (e.g., Cloud & Vaughan, 1970). Greenleaf (1992) emphasized the need of using uncorrelated items

to create response style indicators so that response patterns are not confounded with the true values of the constructs measured by the items.

In order to prevent the emergence of response sets some authors have suggested modifications to the questionnaire. Cronbach (1946), for example, recommended adding special instructions to encourage respondents to answer all questions of an exam. The goal was to persuade respondents who avoid giving an answer when unsure to provide an answer, thus reducing differences in item nonresponse due to personal tendencies to answer when uncertain. Other authors have recommended the use of "forced-choice" questions (Cronbach, 1946; Schuman & Presser, 1981) to avoid acquiescent response. There is little empirical evidence about when and how to use these strategies.

Limitations of the response styles literature

Researchers starting to investigate response styles are likely to come across a large amount of work from psychology and a somewhat narrower set from social survey research. Even though the terms of acquiescence, extreme response style and middle response style are used in both fields, some of the publications in survey research do not use the terms "set" or "style" to refer to the same patterns. Even though survey research explores similar concepts, it does so under different names. Therefore, findings that could contribute to better understanding the mechanisms underlying response styles are not considered in the response styles literature framework. Survey literature can point to different variables that have not been considered as predictors of response styles. In connection to this, a number of characteristics of survey research designs that make it

different from response styles studies may help advance response styles research: the sampling design, the types of items used, and the mode of data collection.

One distinction between the body of research from survey methodology and that of psychology is sample composition. In survey research the use of probabilistic samples is much more common than in other kinds of social sciences research. The population of inference tends to be wider than university students, allowing certain variables such as educational level, age or language to vary more widely across respondents. Hence, probabilistic samples make generalizability of survey estimates to a wider population possible.

Another feature that differs between survey research and psychological research examining response styles is the number of items that are used to measure one construct. Research in most surveys from representative samples tends to use fewer items to measure a single construct than psychological research using highly motivated university students. As a consequence, the instruments differ greatly, and so does what can be done with them in order to study response styles. Having fewer items per construct in survey research makes it more difficult to create variables that reflect *patterns*—giving the same answer to a couple items would not constitute a pattern. On the positive side, they may encompass different types of answer scales, question formats and modes of presentation. If patterns were found to be consistent in spite of all these methodological changes, this would suggest that response styles are a tendency inherent to the respondent. In contrast, personality research is characterized for long batteries of items, similar in length and content, with the same answer scale. The large number of items is optimal to establish

claims of stability, but the similarity of the stimuli may have increased the likelihood of choosing the same response options repeatedly.

The literature has failed to clearly link the findings on response styles with other studies from survey research that focused on the same patterns (e.g. tendency to agree, preferred selection of the middle point) without hypothesizing an inherent idiosyncratic habit on the respondent. In other words, the literature that understands response styles as measurement error that stems from the respondent and the literature that considers the source of error to be methodological aspects of the survey have been predominantly independent from each other.

Survey researchers have studied the impact of manipulating certain survey design features (e.g. data collection mode, question format) on these measures. The intention is to demonstrate how manipulations of the material presented to the individual can vary the response distributions that are obtained. The "response styles" literature has much less work of experimental nature and could clearly benefit from this kind of studies.

Emphasis in survey research is placed on what can be done on the side of the researcher to avoid responses that are not content related. Velez and Ashworth (2007), for example, studied the impact of perceived item clarity on endorsement of midpoints. They were trying to reduce measurement error through careful survey design. Saying that the literature on response styles has ignored these aspects would be misleading, yet considerable work needs to be done. The present study intends to relate findings from survey research that have not been yet associated to response styles, namely, the wording of verbal labels of answer scales.

Cross-cultural Differences in Response Styles

One consistent finding in the literature is that cultures seem to differ in the way they use rating scales (Johnson, et al., 2005). Results have repeatedly shown differences in the use of answer scales, whether the comparison is made on the basis of race, ethnicity, or country origin.

Research has shown that European American respondents exhibit lower extreme response style (I. Clarke, 2000a; Hui & Triandis, 1989; Johnson, et al., 1997; Marín, et al., 1992) and acquiescence than Hispanic respondents (I. Clarke, 2000b; Hui & Triandis, 1989; Johnson, et al., 1997; Marín, et al., 1992; Ross & Mirowsky, 1984) or African-American respondents (Bachman & O'Malley, 1984a; Berg & Collier, 1953; I. Clarke, 2000a; Holbrook, et al., 2006; Johnson, et al., 1997).

Outside the North American context, Watkins (1992) reported that Blacks in South Africa select endpoints more often than the other ethnic groups under study. Shapiro, Rosenblood, Berlyne and Finberg (1976) found that Bedouin adolescents select extremes more than Moroccans adolescents. Javeline (1999) showed that Kazakhstanis are more likely to acquiesce than Russians.

Across countries, it has been found that Mediterranean (van Herk, et al., 2004) and Latin American countries (I. Clarke, 2001) show more extreme and acquiescent response style than Northwestern European countries or the United States. Chen et al. (1995), Chun et al. (1974), Hamamura et al. (2008), Harzing (2006) and Zax and Takahashi (1967), found that Asian respondents tend to choose midpoints of a scale more often than American respondents. Clarke (2000a) found that French select endpoints more often than Australians.

The main concern arising from these findings is that the validity of the differences across cultures or countries in the variables of interest can be compromised. Systematic tendencies unrelated to content may bias survey estimates. If response styles differ across cultures, all other things being equal, survey estimates in different cultures will experience different biases. As a consequence, observed differences between two cultures could be attributed either to differences in true values or differences in measurement error. Thus, cross-cultural researchers need strategies to reduce or control the effect of response styles so that inferences can be made regarding the variables of interest. Understanding why cultural differences in response styles occur is crucial in developing techniques to reduce its effect on survey estimates. This section reviews the main factors that have been hypothesized to help explain response styles. I focus here on those for which there is empirical evidence, even though more factors have been brought to discussion *ad hoc* as differences were found.

Research exploring the causes of these differences has focused on three types of factors: psychological aspects of the individuals, cultural norms, and measurement procedures (Arce-Ferrer, 2006).

Psychological aspects of individuals

Cultural differences in response styles could be the manifestation of differences in thinking styles, language use, and other cognitive factors that are shaped by culture.

- *Meaningfulness.* As mentioned before, Gibbons et al. (1999) hypothesized that meaningfulness of the questions will drive respondents towards the endpoints. One of their concerns was that questionnaires developed for one culture may be less

meaningful when applied in a new culture. They found a positive correlation between self-reported meaningfulness of the items and extreme response style.

- *Dialectical thinking* describes the tolerance for agreeing with apparently opposing ideas (Peng & Nisbett, 1999). Peng and Nisbett found that Chinese often endorsed two arguments that Americans viewed as incompatible. Hamamura et al. (2008) hypothesized that the greater accessibility of contradictory beliefs would induce a tendency to select the midpoint of an answer scale. Using a dialectical thinking scale they found that the effect of culture on moderacy response style disappeared after controlling for dialectalism.
- *Need to evaluate.* Holbrook et al. (2006) investigated the effect of controlling for need to evaluate on the relationship between race and response styles. They showed that Black respondents chose endpoints more often than other respondents, and this effect was *not* mitigated when controlling for need to evaluate. A similar finding was presented for Hispanics.
- *Language.* Language has been studied from two perspectives. The first is concerned with the effect of regional variations in language use on response styles. Bachman and O'Malley (1984) compared levels of extreme response style for different regions of the United States, as a proxy for language use. Even though they found differences across regions, these differences did not account for the differences between ethnic groups.

  The second approach compares questionnaires answered in the respondents' native language versus a second language (usually English). When answering in their native language, respondents are expected to use moderate positions less often.

Gibbons et al. (1999) and Harzing (2006) find that respondents used more extreme categories when answering in their first language. They interpret this difference as the result of lower level of competence in the second language. Harzing (2006) confirms this hypothesis finding that English proficiency correlates positively with extreme response style, and negatively with middle response style.

Cultural factors

Ever since differences in levels of response styles across cultures were documented (Brengelmann, 1959; Triandis & Triandis, 1962; Zax & Takahashi, 1967), explanations using cultural factors have been proposed. However, the number of empirical studies that relate response styles to measures of cultural variables is small. The cultural explanation provided frequently relies on assumptions and *a priori* conceptions of the culture, without providing empirical evidence of the relationship. For example, Marín et al. (1992) found differences in levels of acquiescence and extreme response style between Mexicans, Mexican Americans and White nonhispanic Americans, and use the correlation between acculturation and level of response style as evidence that levels of response style are related to culture. Even if culture is identified as a source of variance, the explanations adventured by these authors are not supported by any empirical evidence.

Much of the literature that does use measures of cultural norms and factors focuses on Hofstede's measures (2001) of the cultural dimensions of power distance, uncertainty avoidance, and individualism/collectivism. One of the advantages of Hofstede's variables is that scores for over 60 countries are easily available to researchers

(for example through Hofstede's own website[3]). In an attempt to validate the observed relationships between Hofstede's dimensions and response styles, some authors have used similar measures of these dimensions, such as the Global Leadership and Organizational Behavior Effectiveness (GLOBE, (House, Hanges, Javidan, Dorfman, & Gupta, 2004)) project (Harzing, 2006; P. B. Smith, 2004). Using country-level variables of cultural dimensions, Peter B. Smith (2004) found that variables from both projects are correlated with global measures of acquiescence. Harzing (2006) also found that cultural dimensions are related to country-level measures of response styles.

There is a lack of research with variables other than Hofstede's cultural dimensions. Van de Vijver, Ploubidis, and van Hemert (2004) are one exception, including in their analysis such diverse variables as educational system, length of democracy history in a country, and aggregate measures of satisfaction, religious denomination, and social desirability.

Variables regarding discourse tendencies in certain languages could be included if measures existed that described them quantitatively. Other areas that could be explored include cultural norms, or behavior patterns specific of a particular group (Triandis, Marín, Lisansky, & Betancourt, 1984).

Marín, et al. (1992) found that acculturated Mexicans were less prone to select endpoints than those who are not acculturated. Acculturation is not a cultural norm, but it is used by the authors as an indicator of differences in cultural norms. Under this assumption, they take their findings as evidence that response style is a "culturally

[3] http://www.geert-hofstede.com/

driven" phenomenon (Gibbons et al., 1999). Furthermore, Johnson et al. (1997) did not find a significant effect of acculturation neither on acquiescence nor on extreme response style. Provided here is a description of how the most commonly used cultural predictors relate to response styles:

- *Modesty/moderation.* Social norms that call for modesty have been used to explain the preference of Asians for the middle point as compared to Americans, Canadians, British and Australians. Zax and Takahashi (1967) are among the first to use a cultural explanation for the cross-cultural differences in response styles. They hypothesized that Japanese are less prone to exhibit extreme response style because their child rearing practices emphasize restrain. Their results showed that Japanese college students were less likely to select the endpoints than American students, and more likely to select the middle point. However, no direct evidence was provided that the differences in response styles were indeed due to child-rearing differences.
- *Hofstede's cultural dimensions.* Power distance, uncertainty avoidance, and individualism/collectivism have been used as explanatory variables for cultural differences in response styles. A number of studies have used Hofstede's scores of cultural dimensions. As described in detail below, these studies found mixed results, and no clear tendencies have been identified of how these variables relate to response styles.
    a) Power distance is related to how inequality is seen and dealt with within a culture (Hofstede, 1980; 1991; 2001; 2006). This dimension measures "interpersonal influence" as perceived by those in the less powerful position, where influence is understood as the extent to which one person can determine the behavior of

another (Hofstede, 2001). Deference towards higher status figures and submissiveness are characteristic of high power distance countries. Therefore, countries high in power distance are expected to exhibit more acquiescence, in deference to the interviewer as a figure of higher hierarchical status. Findings in van Hemert, van de Vijver, Poortinga and Georgas (2002) supported this hypothesis. Harzing (2006) and P.B. Smith (2004), however, showed only partial support for this. Both studies found significant relationships in the expected direction when using Hofstede's country level indicators, but not when using the GLOBE study scores. Evidence from other studies, at the same time, showed either no relationship (van de Vijver, et al., 2004) or a significant relationship in the opposite direction (Johnson, et al., 2005). The latter also found a significant positive correlation with extreme response style.

b) Uncertainty avoidance is the society level equivalent for ambiguity tolerance. Johnson et al. (2005) found uncertainty avoidance to be negatively related to acquiescence (the more they avoid uncertainty, the less they agree with the statements). Although somewhat counterintuitive, this finding is consistent with the description that Hofstede made of countries low in uncertainty avoidance: "(…) uncertainty accepting cultures, are more tolerant of opinions different from what they are used to" (2006).

   Van de Vijver et al. (2004), however, found no significant effect of this variable, and Harzing (2006) found mixed results.

c) Individualism/collectivism. In collectivistic, interdependent cultures, fitting in and maintaining harmony with the ingroup are important aspects, whereas in

individualistic, independent ones members want to stand out, and they see group membership as something they can freely choose (Oyserman, Coon, & Kemmelmeier, 2002). On account of the stress on harmony, and the lesser emphasis on individual opinions, Harzing (2006) and Johnson et al. (2005) hypothesized more middle response style and more acquiescence in collectivistic cultures. Chen et al. (1995) found that individualism is negatively related to the use of the midpoint and positively to using the endpoints. This finding was observed for students in the United States, Japan, and Taiwan. In addition, the two traditionally collectivistic countries (Taiwan and Japan) showed preference for the middle point. Johnson et al. (2005) also found support for the hypothesized relationship with acquiescence. Harzing (2006), however, reported correlations in the opposite direction.

The mechanisms proposed as explanations of the (sometimes observed) relationship between response styles and the cultural dimensions of power distance, uncertainty avoidance, and masculinity vs. femininity are based on a number of assumptions.

However, no evidence is provided in support of those assumptions, which raises concerns about whether the proposed links are more than just speculation. Schwarz, Oyserman and Peytcheva (in press) argue that the cognitive consequences of cultural variations other than individualism vs. collectivism are, indeed, “not yet sufficiently understood to lend themselves to fruitful discussion”. Consistent with this view, the body of research reviewed here that empirically tests these relationships shows contradictory results.

Measurement procedures

A number of aspects related to research design, instrument design, and the data collection process have been mentioned when discussing differences in response styles across cultures. At the same time, few have explicitly focused on empirically testing their effect on response styles.

- *Familiarity with materials.* Shapiro et al., (1976) found that Bedouins endorse endpoints more often than Moroccans. They hypothesized that it is due to the lack of familiarity of the former with Western forms of testing. Arce-Ferrer (2006) observes higher extreme response style in Mexican respondents that live in rural areas than those from urban locations. He argued that this was due to differential familiarity with rating scales. Neither provided a rationale for why lower familiarity would necessarily induce higher use of the endpoints of a scale, and not some other pattern.
- *Mode of data collection.* In research involving more than one country, it is common to find that different countries field the questionnaire using different modes. But even the same mode of data collection may have a different impact on response patterns for different countries, depending on how familiar respondents are with the mode—compare, for example, the effect of a computer assisted interview among a population where computers are not easily accessible to the use of such devices in mainstream individuals from Western countries.

  Involving a different source of error, interviewer-administered surveys may have a stronger impact in countries where democracy is young and people may still fear consequences of sharing their opinions with strangers (van de Vijver, et al., 2004).

Another example is the potential impact of visual design in self-administered surveys. The impact of seemingly small changes in visual aspects can trigger important shifts in response distributions (Smith, 1995; Smyth, Dillman, & Christian, 2007; Tourangeau, Couper, & Conrad, 2007). Therefore, having some countries use modes with visual input and others without it can have an important effect. Research on mixed mode effects can point to important aspects of cultural differences in response styles. In response styles research, it has already been shown that aspects related to the mode of presentation of the instrument can have an impact on preferences for certain points of rating scales in Belgium (Weijters et al., 2008) and in the United States (Dillman, Phelps, et al., 2009). In both studies interviewer-administered interviews were shown to lead to higher use of the endpoints. To my knowledge there is no research that links data collection mode and cultural differences in response styles.

- *Types of answer scales*. There are two main aspects of the answer scale format that have been studied in relation with cross-cultural differences in response styles: the use of unfolding questions and the number of scale points.

  Albaum and colleagues (1988; 2007) manipulated the format of the answer scale in order to study the effect it has on extreme response style. They compared single-stage attitudinal questions (also known as branching or unfolding questions) to two-stage questions, finding that the latter yields more extreme response style. Arce-Ferrer (2006) failed to replicate these findings with respondents in Mexico.

  Hui and Triandis (1989) found that differences in extremity between Hispanics and Non-Hispanics are moderated by the number of points of the answer scale.

Hispanics tended to use more extremes of the endpoint labeled scales than Non-Hispanics when the answer scale had 5 points, but not when it had 10 points. Watkins (1992) replicated this finding with South African respondents and extended their research to gender and age. They found a significant effect of race, where Blacks were more extreme than Whites, Indians and Colored when using the 5-point scale but not when using the 10-point scale. Even though there were no main effects of age and gender, they found an interaction effect (race x gender x age) with 5-point scales that was not observed with the 10-point scales.

I. Clarke (2000a, 2000b) experimentally manipulated the number of scale points, ranging from 3 to 10. Using labels in the endpoints, he found an effect on extreme response style across ethnic groups and countries. The overall finding as that extreme response style was reduced as the number of response options increased for almost all countries and ethnic groups. This difference, however, was generally nonsignificant when moving from a 5-point scale towards higher numbers. More importantly, differences across groups varied depending on the number of scale points used, confirming previous findings by Hui and Triandis (1989) and Watkins (1992). Cultural differences were maximal using the 5-point scale when comparing Hispanics and Non-Hispanics, and Black to Non-Black. When comparing Australian and French respondents, however, extreme response style differences were found when using 7- and 9-point scales.

Summary

The existing literature on cross-cultural differences in response styles offers some guidance on the kind of variables that can help explain them. However, the mechanisms through which these patterns emerge are still not fully understood. More empirical evidence is needed to help describe a complex picture.

There are two main gaps in the current literature on cross-cultural differences in response styles. First, few studies have evaluated the simultaneous impact of individual, cultural, and methodological factors on response styles. Harzing (2006), Van de Vijver et al. (2004) and Johnson et al. (2005) focus on individual and cultural aspects simultaneously. I. Clarke (2000a; 2000b; 2001) and Arce-Ferrer (2006) study methodological and individual aspects. More attention needs to be given to the impact of methodological aspects while including individual and cultural factors in a comprehensive model.

Second, there is too little known about the effect of answer scale features on response styles. Previous findings strongly suggest that changes on the answer scale format can affect conclusions about cross-cultural differences in response styles (I. Clarke, 2000a; 2000b; Hui and Triandis, 1989; Watkins, 1992). This is particularly relevant when conducting multilingual research because more than one version of each answer scale is needed.

Chapter 2. Answer Scales Development in Multilingual Surveys

In closed-ended questions, researchers offer a list of response options along with the question; respondents are expected to select one of these options as their answer to the question. Answer scales are understood here as a specific form of response option list in which a given order underlies the options (although nonsubstantive options such as "don't know" or "not applicable" may be presented along with the scale). In this definition, therefore, purely categorical, unordered lists of answers are excluded.

Answer scales are, thus, one form of response alternatives to closed-ended questions. The use of closed-ended questions is motivated by practical concerns of dealing with answers to open-ended questions (Bradburn, Sudman, & Wansink, 2004). Open-ended questions need to be coded in order to be susceptible of statistical treatment, raising concerns about variance stemming from the interviewer (who records the verbatim response) and/or the coder (who assigns a given category to each answer). The coding task involves the interpretation of the respondent's answer by a third party—sometimes the interviewer—who matches it to one category from a set of agreed categories. In addition to the time consuming nature of the task, selecting coding categories for analysis involves challenges similar to those involved in the process of writing categories for closed-ended questions. In reality, researchers use try to avoid open-ended questions to measure attitudes, although they have proven valuable, particularly in qualitative research—for example, in exploratory research to gather

information for the question design (Bradburn, et al., 2004). Researchers expect to offer response options that best represent how most people express their "value" in the variable of interest, and aim for categories that make the task of matching the answer to the precoded categories easy and the outcome accurate.

In the literature of survey research and other social sciences, authors refer to answer scales also as rating scales (e.g., Bradburn, et al., 2004; Sudman, et al., 1996; Tourangeau, et al., 2000), response scales (e.g., Smith, 2003), and ordinal scales (e.g., Dillman, Smyth, & Christian, 2009). The response options can be referred to as categories, alternatives, or scale points. The words used to define the response categories can be referred to as verbal labels, or as scale anchors when talking about the endpoints of a scale. Throughout this dissertation, the term "answer scale" refers to any form of ordered list that is provided to the respondent in a questionnaire. In an answer scale, an object is presented to the respondent, who must evaluate it along a dimension, that is, along a scale that varies according to an underlying continuum between two extremes (Oldendick, 2008). A respondent can be asked to evaluate, for example, a same-sex marriage law (the object). The respondent can then be asked to use the dimension "consequences for society" to evaluate the object. The underlying continuum hypothesized can be delimited by the extremes: "Very bad for society" and "Very good for society", with a middle option describing lack of effect on society.

This chapter aims to identify the main findings on verbal labels used in answer scales and to indicate some areas in need of further research. Much of the literature of labels in answer scales refers to or was based on monolingual research; only a few efforts have been made to investigate how verbally labeled answer scales perform in

multilingual settings. Exceptions are the work of Harkness and colleagues (see, for example, Harkness, 2005; Harkness, Mohler, Smith, & Davis, 1997; Mohler, Smith, & Harkness, 1998; and Smith, Mohler, Harkness, & Onodera, 2009), and efforts in Quality of Life (QoL) research such as those described in Szabo, Orley, and Saxena (1997). These studies are described in the section discussing considerations of answer scales use in multilingual contexts. This dissertation intends to contribute to this line of research, and join them in stressing the particular importance that the effects of verbal labels have in multilingual surveys.

Answer Scale Design Features

The design of survey answer scales involves a large number of decisions. Researchers must decide how many scale points to use, whether to use an odd or even number, how many of those points to label, what specific words to use as labels, and which end of the scale should be presented first. The scale points may or may not be accompanied by numerical labels; if numerals are present, a range for the numerical labels must be selected. When answer scales are presented visually, decisions must be made regarding whether the scale should be presented horizontally or vertically, the spacing between categories, whether visual guides are provided (e.g., ladders or other iconic representations) and for computer assisted surveys the kind of response formats that will be used, such as drop-down menu or radio-buttons. These decisions are especially important because answer scales are often presented as response alternatives for various questions within an instrument, whether these questions belong to an item battery or not. This means that problems in the design of answer scales will potentially

affect a larger number of variables of interest than problems in the design of the question stem wording.

In 1997, Krosnick and Fabrigar noted that questionnaire design manuals provide little guidance regarding how to develop answer scales. The situation has not changed greatly since then, although exceptions are recommendations found in Dillman, Smyth and Christian (2009), Bradburn et al., (2004) or, with a cross-cultural perspective, in Smith (2003; 2004), Harkness (2008), and Harkness, Edwards, Hansen, Miller, & Villar (in press).

Research and Recommendations for Choosing the Verbal Labels: The Case of Multilingual Surveys

In multilingual settings, decisions about the answer scale design are made with respect to the language in which the questionnaire is developed: the *source* language. Different questionnaire development strategies lead to different levels of cross-cultural input when developing the source answer scale (Harkness, 2008; Harkness, et al., in press; Harkness, et al., 2003). Ideally, if the questionnaire is to be used across a number of cultures, input from all those cultures would be brought to the questionnaire developing stage. This way, potential problems in the use of answer scales can be identified and tackled early in the process. Whether cross-cultural input has been present early in the process or not, the source answer scale version is usually finalized before translation starts (Harkness, 2003; Harkness, Pennell, & Schoua-Glusberg, 2004). Scales are then translated as needed into languages that are expected to be found in the population(s): the *target* language(s). The most common model involves trying to

produce target language versions of source scales on the basis of translation, assumed to be a way to achieve comparability (Harkness, 2003). However, answer scales have been shown not to translate easily or consistently in written translation (Harkness, 2003, 2005), in oral translation (Harkness, Schoebi, Joye, Mohler, & Behr, 2008), or in interpreted telephone interviews (Harkness, et al., 2009). More sophisticated models are possible and will be later described (e.g., Mohler, et al., 1998; Skevington & Tucker, 1999; Szabo, et al., 1997), but close translation is by far the most common.

Harkness (2003) mentions three reasons why differences across translated response scales may occur: a) structural and lexical differences across languages, b) preference for certain formulations/formats from a fielding institute (or survey tradition), or c) inadvertent change. In addition, countries may decide to alter the scale in order to adjust to known aspects of measurement error within the culture. For example, if the individuals in charge of producing the target questionnaire are familiar with the literature in response styles, they may want to adjust the verbal labels in order to make certain scale points more appealing to a particular culture.

Whether in monolingual or in multilingual contexts, two main components need to be considered when creating verbal labels for answer scales: the dimension in which the object will be evaluated (e.g., importance, satisfaction, or agreement) and the intensity associated with each scale point.

Scale dimension

Choosing the verbal labels to measure an attitude of interest involves having a clear idea of the dimension the researcher intends to investigate (Bradburn, et al., 2004;

Dillman, Smyth, et al., 2009), of how the dimension is cognitively represented in the respondent's mind, and of how it is verbally expressed in daily life. Recommendations regarding how to define the answer scale dimension are related to general question design strategies, and involve clearly stating the research goals, selecting the constructs, indicators, and variables that are needed, drafting questions and pretesting those drafted questions until a satisfactory version is produced (Bradburn, et al., 2004; Fowler, 1995; Fowler & Cosenza, 2008; Harkness, et al., in press).

When choosing a dimension to evaluate the object of an attitude, a very common practice in survey research is to formulate "Likert-type" statements and ask respondents whether they agree or disagree with each of those statements. This approach to attitude measure presumably has advantage for implementation in that the same answer scale can be used for a large number of statements, which reduces the number of answer scales that need to be tailored to a specific item and the space needed in a visually presented questionnaire. However, numerous problems arise from it and are described in the remainder of this section.

Using a question from the ISSP 1999 questionnaire, Table 2.1 shows an example of a construct-specific answer scale and its "equivalent" generic agreement answer scale.

Note that any of the five response options of the construct-specific answer scale could be used to create the stem for the statement on the right. One could choose "I earn much less than I deserve" or "I earn what I deserve" and then present the agreement options to respondents. It is not clear which one will yield the most accurate results. In addition, how respondents interpret of the response options to mean is open to forms of variability that are not met when using construct-specific answer scales. In the generic

portion of the example, conscientious respondents who consider they earn *much* more than they deserve may feel indecisive as to whether the researcher expects from them. They may feel inclined to choose *strongly agree*, using the word "strongly" to indicate

Table 2.1.
Example of construct-specific answer scales vs. agree/disagree statement

	Construct-specific answer scale	***Generic agree/disagree scale***
Question stem	**Would you say that you earn...**	**I earn more than I deserve**
	Much less than I deserve	Strongly agree
	Less than I deserve	Agree
	What I deserve	Neither agree nor disagree
	More than I deserve	Disagree
	Much more than I deserve	Strongly disagree

the amount they earn relative to the statement, or *disagree*, because the statement does not quite represent their perception. Furthermore, respondents who perceive themselves as earning what they deserve could strongly disagree with the statement "I earn more than I deserve". Saris, Krosnick, and Shaeffer (2005) refer to this phenomenon as a violation of the "presumed monotonic relation between answers and respondent placement on the underlying dimension of interest" (p. 6).

Another criticism to a Likert-type approach comes from the lack of fit between what the question asks and what the response scale offers (Dillman, Smyth, et al., 2009). These authors hypothesize that this lack of fit may increase the respondent's burden due to a postulated extra step in the cognitive response process.

Saris et al. (2005) give a more detailed account of the misfit, proposing a theory of the cognitive processes involved in the response to agreement questions. The example in Table 2.1 is used here to illustrate the cognitive steps Saris et al hypothesize:

1. First, the literal meaning of the statement must be understood.
2. Then, respondents need to "discern the underlying dimension of interest to the researcher (…) by identifying the variable quantity in the question stem" (Saris, et al., 2005, p. 5). In the example above, variability is identified by the relative quantity "more than"—how much one earns as compared to a hypothetical value.
3. The respondent then needs to place him or herself on that dimension of interest, that is, respondents must decide how much they earn relative to what they deserve.
4. In a last step, respondents must convert their position on the dimension identified in step 3 into an agree/disagree response option. They need to find the agreement response option that corresponds with their value on the dimension and, as mentioned before, this process may be less straightforward than assumed by question designers. This last step would not be necessary with a construct-specific answer scale, which would reduce the burden involved of the response process.

An additional criticism of using agree/disagree to measure other dimensions is based on the body of research showing the presence of acquiescence when using agreement answer scales (Saris, et al., 2005). Saris et al found empirical evidence supporting these theoretical considerations. In two different studies, one using a cross-sectional split-ballot study and one using a multitrait-multimethod approach, they found

that questions with construct-specific answer scales yielded more reliable and valid answers than the same questions measured by means of agree/disagree answer scales.

In multilingual scales, the actual dimension may vary across languages due to translation-induced mistakes or structural and lexical differences. Add example

Close translation of answer scales does not guarantee that the resulting answer scale is comparable to the source answer scale. Even an adequate rendition from a semantic point of view can result in something that functions very differently than what was intended. Adaptation will often be needed in order for a scale to work in a given context. Struwig and Roberts (2006) have reported that the use of 5-points bipolar answer scales proves really challenging in South Africa.

A dimensional change sometimes observed in translation of agreement answer scales is the switch from a bipolar to a unipolar answer scale. Unipolar answer scales typically range from the absence of the measured dimension (e.g., *not at all satisfied*) to a large value of it (e.g., *extremely satisfied*). Bipolar answer scales, instead, range from a (strong) negative end of the dimension (e.g., *extremely dissatisfied*) to a (strong) positive one (again, *extremely satisfied*). Everything else held constant, the bipolar scale is expected to yield a negatively skewed distribution—that is, a distribution where the negative side of the answer scale is less frequently selected by respondents. If a unipolar scale is the result of a translation of a bipolar scale, comparison of questions associated with that scale could be problematic.

It is important to note that the range of numerical labels has been shown to influence whether an answer scale is understood as a bipolar or a unipolar scale. Schwarz, Knauper, Hippler, and Noelle-Neumann (1991) manipulated the numeric labels

associated with the endpoint labels of an 11-point scale ranging from *not at all successful* and *extremely successful.* They found that when the scale was accompanied with a 0 to 10 scale, *not at all successful* was understood to mean absence of success. However, when the numeric labels ranged from -5 to +5, they interpreted it as the presence of failure.

The ISSP agreement answer scale, for example, was translated into French as an approval scale (Harkness, 2003). Such differences could pose a threat to comparability, introducing, for instance, variation in the burden they experience or the interpretation of the item. Because of the scant attention answer scales have received in multilingual survey research, little empirical evidence of the consequences of such changes is available. Researchers make predictions of problems generated by these changes at face value, on the basis of semantic considerations and common sense. However, there are numerous sources of information that are hardly ever mentioned, and could be of invaluable help in survey translation, and in particular in translation of answer scales. Harkness and McKinney (2009) review concepts of communication and discourse across different disciplines that could help creating benchmarks and improving best practices in survey translation.

Two important lines of research that have addressed this type of issue during the past decade are research in multilingual calibration studies and research in culture, cognition and communication. These will be discussed in the next section.

Intensity of the scale points

When designing answer scales, researchers aim to obtain scale points that have a monotonic relationship with the variable of interest (Saris, et al., 2005). That is,

researchers expect that the higher the true value of the respondent on that variable, the more likely the respondent will be to select a response option from the higher end of the answer scale (or the lower end in the case of "reversed" items). Achieving this goal ensures that the variable can be treated in analysis as an ordinal variable (Stevens, 1946). In addition, researchers often try to obtain answer scales with equally spaced intervals between adjacent scale points (Dillman, Smyth, et al., 2009; Friedman & Amoo, 1999; Krosnick & Fabrigar, 1997). If scale points are indeed "equidistant", additional mathematical operations can be performed with them, and analyses requiring "interval scale" measures would be justifiable.

Verbal labels have been recommended in the literature because they have shown to improve reliability and validity of the measures (Krosnick & Berent, 1990; Krosnick & Fabrigar, 1997). The choice of verbal labels is likely to affect the perceived distance between scale points. For example, the distance between "completely satisfied" and "satisfied" seems, at face value, smaller than the distance between "completely satisfied" and "slightly satisfied". However, little guidance is provided in the literature for how to find labels that represent equally distant scale points, partly because evidence on how verbal labels affect response distributions is limited.

Smith et al. (2009), in a review of methods for assessing and calibrating response scales, describe three approaches for evaluating the intensity of verbal labels of response categories: a) rating the strength of terms defining each point of the scale; b) measuring the distributions generated by using different response scales; and c) using anchoring vignettes to establish comparability across measures. Each has advantages and disadvantages, but using any of these methods seems to improve upon answer scale

version production in multilingual settings in which a close translation of the source response scale is the aim—and the source response scale would also need to be developed using the procedure of choice so that across-language comparability could be established.

Of the three approaches for evaluating the intensity of response categories, the rating of response categories has been used in multilingual applications (Harkness, et al., 1997; Mohler, et al., 1998; Smith, et al., 2009; Szabo, et al., 1997). In this method, respondents are usually asked to assign a number (for example, between 0 and 20) to various different labels (Mohler, et al., 1998; Smith, et al., 2009). Averages and variances are computed for each label. Averages indicate the intensity of the label, and variances indicate the extent to which meaning is shared across participants. A label with low variance suggests that there is high level of agreement in the intensity respondents assign to it; equal meanings are sought-after in any form of verbal stimulus that researchers presented to respondents in surveys, therefore lower variances are a desirable feature of scale anchors. This calibration method has the potential to a) better guide the selection of evenly distant wording and b) serve as guides to choose comparable wordings across languages in multilingual surveys (Mohler, et al., 1998; Smith, et al., 2009).

Multilingual calibration studies show that respondents[4] are able to perform these tasks across all languages used in the studies—including in languages as different from English as Croatian, Hebrew, Cantonese, Thai, Hindi and Shona (Szabo, et al., 1997). Szabo et al., however, report that some languages produced fewer "descriptors" to be rated for each answer scale than initially envisioned. The authors attributed it to "differences in the character of languages in which the scales were developed" (p. 273).

[4] Szabo et al. (1997), however, report that some respondents found the task too abstract, particularly low-educated respondents.

Nevertheless, the outcome of the calibration task was satisfactory for the researchers, and the answer scales of the World Health Organization Quality of Life Assessment Instrument (WHOQOL) were labeled worldwide using this procedure.

Even though the calibration procedure appears more rigorous than using the researcher's intuition of particular verbal labels work or just replicating previously used answer scale labels, it remains to be proven that the verbal labels obtained through this procedure indeed lead to more accurate measurement. The researchers in the WHOQOL project successfully validated the ordinal nature of the scale (Skevington, Sartorius, Amir, & The WHOQOL group, 2004; Szabo, et al., 1997), but they did not attempt to test the distance between intervals empirically. Smith et al (2009) use the calibration scores to adjust data from the ISSP where some of the labels they studied had been used. These authors compare the correlations as obtained with the raw data to the correlations obtained when using the calibration scores, and find that the adjustment attenuates correlations. They hypothesize that when presented as a response scale, people may assign equal distances to the words and phrases that describe the scale points. That is, respondents would "shift from scale-independent evaluations of the response terms to more ordered, scale-dependent assessments".

Culture, cognition and communication

As a form of communication, surveys need to abide to some extent by the rules of conversation, discourse, and human interaction. Thus, experience answering questions in other situations in life carries over to the survey situation, affecting how respondents make sense of the survey question (Groves, et al., 2009). One of the aspects of

communication that has been identified in the survey context is that respondents assume that the researcher is following standard conversational norms when creating the questions (Schwarz, 1996; Sudman, et al., 1996). Schwarz and his colleagues show that the way respondents interpret answer scales is affected by an expectation that the researcher is complying with these conversational norms.

Complicating matters, cross-cultural survey research needs to take into consideration that discourse norms, interview experiences, question and answer experiences, or examination experiences are uniform across cultures. Different cultures observe different rules of interaction, a phenomenon that affects how respondents, interviewers, and researchers approach the interview situation. This has long been acknowledged (e.g., Jones, 1963) but the consequences have sometimes been ignored in the survey literature, and rarely been empirically studied.

Largely unexplored are the consequences that differences in cognitive processing across cultures may have on how respondents use answer scales. Innovative work on the effect of cultural syndromes on cognition (Haberstroh, Oyserman, Schwarz, Kuhnen, & Ji, 2002; Ji, Schwarz, & Nisbett, 2000; Oyserman & Lee, 2007; Schwarz, 2003; Schwarz, et al., in press) expands the study of context effects previously investigated in German or American samples to cross-cultural survey research. They combine samples with individuals of different cultural backgrounds (e.g., Asian American students and European American students) with cultural priming techniques (Oyserman & Lee, 2007). Cultures differ in which mindsets are chronically accessible. Cultural priming techniques are meant to activate a given “mindset” or cultural orientation in respondents, whether that is the chronically accessible mindset for their culture or not. That way, causal

connections between “culture” and other variables (for example, measurement error indicators) can be tested.

Incorporating this methodological approach to the study of response styles research and research on measurement error associated with answer scale verbal labels would be greatly beneficial. Another methodological advance already being applied to the study of answer scales in multilingual contexts is the cross-cultural implementation of calibration studies described above (Harkness, et al., 1997; Mohler, et al., 1998; Skevington & Tucker, 1999; Smith, et al., 2009; Szabo, et al., 1997). However, these studies are limited in that the resources needed to implement such research designs across a large number of cultures or countries are difficult to gather. Therefore, it is important to conduct research that identifies relevant variables and possible measurement error mechanisms across large groups of cultures, and secondary data analysis provides an affordable opportunity for it. This type of secondary research can also serve as basis for securing future funding of research targeted to investigate phenomena described in it.

Answer Scale Variations in Cross-National Project: Evidence From Five Cross-National Surveys

This section intends to show the reader the extent to which variation in crucial answer scales exists across a number of important cross-national survey projects. It is important for cross-cultural research to evaluate whether these occurrences exist in isolation or whether, on the contrary, closer attention needs to be paid to how answer scales have been translated. The examples presented from major multilingual and multinational projects should suffice to document that considerable variation occurs in

each project and to indicate the relevance of the exploration undertaken here to multinational and multicultural research in general.

Indeed, looking at available documentation from the European Social Survey (ESS), the International Social Survey Programme (ISSP), the European Values Survey, the World Values Survey, the World Mental Health Survey Initiative and the Eurobarometer, translations show variation in numerous ways and instances. Answer scales vary across and within countries, across and within years, and across and within languages. Sometimes a country uses a close translation of the source answer scale for language A but adds some modification when translating into language B. Countries sometimes use different answer scales in different years or modules of the same survey project.

The ISSP Swiss agreement answer scales provide examples of several types of variation across response scales. Reproduced here (Table 2.2) are the translations for the verbal labels of the *agree* scale point for Switzerland.

Table 2.2.
Swiss Translations of the ISSP *agree* scale point in the 2000, 2002, 2003, and 2004

Year	**Language fielded**	**Translation**	**English rendition of translated label**
2000/2002	Italian	D'accordo	In agreement
	French	Plutôt d'accord	Rather in agreement
	German	Stimme eher zu	Tend to agree
2003	Italian	Sono d'accordo	I am in agreement
	French	D'accord	In agreement
	German	Einverstanden	In agreement
2004	Italian	D'accordo	In agreement
	French	Plutôt d'accord	Rather in agreement
	German	Stimme zu	Agree

In the ESS, the Belgium questionnaires for rounds 1 through 3 provide another relevant example. Table 2.3 shows how the questionnaire in French presented two different translations of the label *agree* in different parts of the same questionnaire in Round 1 and in Round 2. In Round 3 only one version was used throughout. In contrast, the Dutch version remained the same throughout the entire questionnaire across all years.

Table 2.3.
Belgian Translations of the ESS *agree* scale point in Rounds 1, 2, and 3

Round	**Language fielded**	**Translation**	**English rendition of translated label**
Round 1	French (version A)	Plutôt oui	Rather yes
	French (version B)	Plutôt d'accord	Rather in agreement
	Dutch	Eens	Agree
Round 2	French (version A)	Plutôt d’accord	I am in agreement
	French (version B)	D'accord	In agreement
	Dutch	Eens	Agree
Round 3	French (version A)	Plutôt d’accord	In agreement
	No version B		
	Dutch	Eens	Agree

The examples presented in tables 2.2 and 2.3 as well as in following tables are based on the agreement answer scale. The translation of the agreement answer scales deserves special attention because they are used across all the international surveys that were reviewed in this study, and because they are one of the most common types of answer scales in surveys in general.

The two clear types of language versions of the answer scales verbal labels were the “translated” and the “modified intensity” versions. Other types of modifications were

found and will be further described below. Without additional information of the rationale behind the choice of verbal labels, it is

a) In a "translated" version, the scale is linguistically and structurally faithful to the source text. The source answer scale is a 5-point bipolar rating scale with the following labels: *Strongly agree, Agree, Neither agree not disagree, Disagree, Strongly disagree*. These faithful versions are referred to here as "translated versions". Table 2.4 shows an example of a "translated version" based on the answer scale used in the Portuguese ISSP for the years 2000, 2002, 2003 and 2004.

Table 2.4.
Portuguese translations of the ISSP agreement scale, 2000-2004

	Fielded answer scale	**English rendition of Portuguese scale**
Portuguese, Portugal 2000-2004	Concorda totalmente	Totally agree
	Concorda	Agree
	Nem concorda nem discorda	Neither agree nor disagree
	Discorda	Disagree
	Discorda totalmente	Totally disagree

b) In a "modified intensity" version, a modifier of intensity is added to or removed from one of the scale points relative to the source answer scale (Harkness, 2003, 2005). Several countries in several modules of the ISSP added a modifier to the second and fourth labels (agree and disagree in the source). The addition of a modifier (e.g., adding *somewhat* to *agree*) can attenuate the perceived intensity of the scale points. Therefore, this type of answer scale version will be referred to here as "modified intensity version". The example of an added intensifier in Table 2.5 was

fielded for the ISSP in Brazil in 2004 and in Portugal from 1997 to 1999 and from 2005 to 2006.

Table 2.5.
Portuguese and Brazilian translations of the ISSP agreement Scale,

	Fielded answer scale	**English rendition of Portuguese scale**
Portuguese, Brazil, 2004 and Portuguese, Portugal 1997-1999, 2005-2006	Concorda totalmente	Totally agree
	Concorda *em parte*	*Somewhat* agree
	Não concorda nem discorda	Neither agree nor disagree
	Discorda *em parte*	*Somewhat* disagree
	Discorda totalmente	Totally disagree

Sometimes, a reversed modification happens, and an intensity modifier is "dropped" in translation. In 2006, the General Social Survey fielded in two languages (English and Spanish) for the first time. Smith (2009) reports that the happiness answer scale that in English read *very happy, pretty happy, not too happy,* but was translated into Spanish as *muy feliz, feliz, no muy feliz* (very happy, happy, not very happy). The modifier "pretty" was dropped in translation. An experiment was conducted in 2008 to investigate the impact of the change. Smith (2009) finds that respondents select *very happy* more often when presented with the modified intensity version.

c) Other types of modification. Countries differ in the extent to which the translation closely reflects the source labels. Whereas most countries seem to aim for the linguistically "closest" term to *agree*, some choose other words. For example, the Japanese ISSP uses *favorable* instead for some years, and the Australian ISSP offers

the scale: *Yes!!*, *Yes*, *??*, *No*, *No!!*. Other changes are present that make the country version of an answer scale look different from the source version. An interesting case is the U.S. agreement answer scale in 2002, where the *disagree* scale point was not presented to respondents for some items, even though no translation is involved. No information is provided in the ISSP website regarding this modification.

Summary and dissertation objectives

A general lack of detailed documentation of certain research design features in literature on cross-cultural response styles leaves some issues unclear. The actual question wording of the fielded languages is seldom reported, at least the answer scales. One is left to assume that answer scales presented to respondents were in fact comparable. However, as mentioned previously, an inspection of several international surveys suggests this may not be the case. Less than a handful of conference papers have empirically studied the impact of changes in verbal labels on response distributions of multilingual studies (Sapin, Pollien, Joye, Leuenberger-Zanetta, & Schoebi, 2008; Smith, 2009), and none has examined such changes across a large number of countries.

As survey methodologists, we need empirical evidence of the consequences due of language version production (Harkness, 2003), as well as theoretical frameworks that allow us to understand and predict when such changes will indeed pose a hazard to comparability. This study attempts to provide empirical evidence of the consequences of changes observed in the context of verbal label translation of answer scales. In the dissertation, I argue that modifications made to answer scales are a critical source of variability in response patterns and distributions and could help explain observed

differences in response patterns (“styles”) across cultures, and empirically investigate answer scale modifications using a large cross-national survey program.

Chapter 3. ISSP Data Collection Procedures

Dataset Selection Criteria

To study the effect of answer scale verbal label modifications on the estimates of interest and on the hypothesized response styles, the ideal dataset would contain the "true" values of all the variables of interest and multiple responses for survey questions on those constructs for all possible measurement protocols. Because this study is set out to investigate cross-cultural differences in measurement error, a dataset with such characteristics would be needed for each of the cultures of interest. That way one could perfectly disentangle differences due to the variables of interest from those due to the differences in answer scales and to the cultural differences in response bias. Because such a dataset is not possible to obtain, the intention is to analyze data from cross-national surveys that meet a number of quality criteria and analytical requirements:

a) use of probabilistic samples;
b) use of agreement answer scales with at least 5 points where answer scale labels were translated following different strategies and thus yielded different verbal labels;
c) availability of a sufficient number of items that are constant in everything but content and answer scale labels;
d) availability of individual level data and country level data;

e) adequate sample size for each group of the answer scale version predictor—a minimum of five units per cell is required.

f) presence of cultural groups that encompass a large range of regional variation.

Available documentation from several cross-national studies was considered: the Eurobarometer, the World Values Survey, the European Values Survey, the ESS, and the ISSP. The source questionnaires for all rounds of these studies were reviewed first, looking for answer scales that were used for at least eight items or questions. After these were identified, the documentation available for other languages was reviewed, searching for translations of the scale point *agree* that differed from a close translation. The Eurobarometer, the World Values Survey and the European Values Survey were discarded because none of the bipolar answer scales with cross-language verbal label changes was used in a large enough number of items. In the ESS, enough agreement items were used, but only three countries in Round 3, and four countries in rounds 1 and 2 had modified verbal labels. In addition, the ESS, being a European project, involves a more reduced range of regional variation than the ISSP or the World Values Survey. The ISSP met all of the criteria and was therefore the preferred option for analysis.

The ISSP: Background

Describing cross-nationally collected data methods is cumbersome if the goal is to achieve the same level of detail observed in reports of one-country data collections. Even if a projects seeks maximum harmonization of procedures, deviations and variations are bound to happen, and reporting them is an arduous enterprise, but recommended practice

(Mohler, et al., in press). Given that this dissertation explores several years of cross-national research, the task is even harder. This chapter sets out to familiarize the reader with the ISSP basic data collection procedures, its strategies in striving for comparability, and deviations from the basic research design that are relevant for the analyses presented here.

The ISSP is an ongoing multinational effort that started in the mid 1980s and currently involves more than 40 nations (http://www.issp.org/). As a cross-national research project, the ISSP faces numerous methodological challenges, in addition to the common issues of survey research in traditionally mono-cultural settings. Each methodological decision involves considering the extent to which harmonization of methods will be pursued, and how this will be implemented (Lynn, Lyberg, & Japec, 2006). In the ISSP, a number of methodological aspects are agreed democratically among its members, and every country is responsible for complying with those decisions. Not all countries manage to comply with all these aspects; inevitably, variation takes place, creating concerns regarding the effect of such methodological differences on comparability. Several relevant methodological aspects are therefore registered in the merged datasets, so that they can be added as controls when analyzing substantive or other methodological variables. Other variations are recorded in the annual "Study Monitoring Report". From the information contained in this report, more control variables can be created. This dissertation combines both sources of documentation—dataset variables and information contained in the annual report—in creating what is called here "methodological factors".

Available Documentation

The ISSP was a pioneer in documentation for comparative survey research (Jowell, 1998; Mohler, et al., in press). They provide documentation including questionnaires for most languages, individual study descriptions that summarize methodological aspects of the data collection, and study monitoring annual surveys that are then summarized and reported together with the other documents. Like in any dataset, the ISSP data documentation has missing pieces of data. Sometimes countries fail to deliver the study description, the questionnaire or miss the deadline to submit the data so that the main dataset does not contain individuals for that country. In order to overcome such limitations, help provided from the Zentral Archive in Germany was of great value[5].

For the purposes of this dissertation, there were additional data problems that required contacting the National Coordinators for some countries. The ISSP datasets considered in this dissertation do not provide information regarding the language in which the interview was conducted. Several countries field the ISSP survey in more than one language. Knowing the interview language for each respondent is crucial for the study of translation effects on survey outcomes. Using contact information provided in the ISSP website, I was able to obtain most of the missing questionnaires and missing data files, as well as an additional variable describing the language of the interview in Switzerland for years 2000 through 2007.

---

[5] I would like to acknowledge the invaluable help from the Zentral Archive and prompt answer to my queries; special thanks to Dr. Evi Scholz, who helped me gain access to a number of questionnaires and provided very useful information to get other missing data.

Sampling design

In the ISSP, each country creates their sampling procedure, which is meant to be probabilistic (Braun & Uher, 2003). The target populations are non-institutionalized adults. The definition of adult and the laws regarding the age at which it is appropriate to interview an individual without parental consent differ across countries(Jowell, 1998). As a consequence, age ranges vary somewhat, starting at 15 for some countries and at 18 for most. Table 3.1 presents the number of countries included for the datasets analyzed in this dissertation, as well as the sample sizes for each country.

Table 3.1.
Number of countries and sample sizes for 1999, 2000, 2002, 2003, and 2004 datasets

	1999	**2000**	**2002**	**2003**	**2004**
Number of countries	25	26	33	32	36
Australia	1672		1404	2717	1928
Austria	1016	1011	2047	1006	1006
Brazil			2000		2000
Bulgaria	1102	1013	1003	1069	1125
Canada	984	1127		1238	1238
Cyprus	1008		1004		1004
Czech Republic	1862	1244	1289	1276	1322
Denmark		1069	1379	1322	1186
Finland		1528	1353	1379	1354
Flanders			1360		1398
France	1889		1976	1724	1531
Germany	1432	1501	1318	1462	1484
Great Britain	804	1133	2312	873	984
Hungary	1208		1023	1021	1035
Ireland		1273	1256	1090	1090
Israel	1208	1205	1209	1067	1034
Japan	1325	1180	1132	1102	1343
Latvia	1100	1000	1000	1000	1002
Mexico		1262	1604		1201
Netherlands		1609	1102		1823
New Zealand	1108	1112	1025	1038	1370
Northern Ireland		1800	1800		
Norway	1268	1452	1475	1469	1404
Poland	1135		1252	1277	1277
Portugal	1144	1000	1094	1607	1607
R Chile	1362	1362	1274	1308	1242
R Philippines	1200	1200	1200	1200	1212
Russia	1719	1723	1827	2429	1853
Slovakia	1082		1133	1152	1082
Slovenia	2024	2174	1093	1093	1054
South Africa				2483	2784
South Korea				1315	1323
Spain	1211	958	2471	1212	2481
Sweden	1150	1067	1080	1186	1295
Switzerland		1006	1039	1037	
Taiwan			1983	2016	1781
Uruguay				1108	1108
United States	1398	1419	1171	1216	1472
Venezuela				1199	1199

Mode of data collection

Currently, the ISSP allows for countries to field either in self-administered mode (the preferred mode) or face-to-face. Online documentation describes the procedures and administration details for each country, and the datasets contain information regarding mode.

Response rates

Table 3.2 presents the response rates for each country and year considered in the analysis.

Not all countries succeed in carrying out probabilistic sampling designs. Some countries allow substitution of sample units, resulting in lack of information about the mechanisms and rates of nonresponse. As a consequence, some countries do not provide the necessary disposition codes, and response rates cannot be computed.

Table 3.2.
AAPOR Response Rate 6 for years 1999, 2000, 2002, 2003, and 2004.

	1999	**2000**	**2002**	**2003**	**2004**
Australia	60.1		59.5	43.9	38.6
Austria	66.4	66.2	63.9	60.3	60.3
Brazil			NC		NC
Bulgaria	94.1	87.6	87.0	90.4	71.9
Canada	27.6	41.6	0.0	43.1	43.1
Cyprus	74.8		71.7		77.2
Czech Republic	53.3	55.1	57.7	53.4	46.9
Denmark		54.7	69.6	66.4	60.3
Finland		61.2	90.9	55.4	54.4
Flanders			65.9		59.7
France	17.3		20.6	14.9	15.4
Germany	47.1	46.2	41.9	48.9	45.6
Great Britain	44.6	61.5	61.5	46.4	52.4
Hungary	66.1		63.0	69.0	46.4
Ireland		61.6	59.9	67.5	67.5
Israel	35.7	38.2	34.5	60.2	58.5
Japan	74.8	66.1	67.3	64.8	77.4
Latvia	58.0	60.7	58.5	58.5	56.1
Mexico		72.2	88.8		69.0
Netherlands		16.8	21.8		41.4
New Zealand	52.8	60.6	57.7	52.5	60.8
Northern Ireland		64.1	62.2		
Norway	51.4	58.1	60.2	60.0	58.5
Poland	66.5		67.3	67.1	67.1
Portugal	69.7	53.3	55.5	57.2	57.2
R Chile	90.5	90.5	85.2	87.2	82.5
R Philippines	NC	NC	46.3	42.2	55.1
Russia	42.1	50.4	33.3	42.5	30.5
Slovakia	NC		NC	79.0	77.6
Slovenia	71.0	69.0	72.4	72.4	69.3
South Africa				76.4	82.3
South Korea				66.8	67.2
Spain	98.5	68.7	99.2	98.5	97.5
Sweden	61.2	57.2	57.2	60.4	65.6
Switzerland		17.2	33.8	30.2	
Taiwan			54.2	47.8	46.2
Uruguay				79.8	79.8
United States	66.6	67.4	56.4	67.7	66.2
Venezuela				92.7	92.7

NC: The response rate could not be computed due to lack of the necessary minimum disposition codes

Questionnaire development: Drafting the source questionnaire and translation procedures

The design of questionnaires for the ISSP is led by the drafting group. Researchers from other countries have numerous opportunities along the developmental process to provide input on the dimensions, question selection, and specific wording. The ISSP assembly elects a different drafting group for each module. The group usually comprises researchers from six countries of different regions and cultures (Kalgraff Skjåk, in press) who develop the questionnaire using English as lingua franca. After piloting at least some of the questions, the drafting group presents a draft to the assembly, where each items is discussed, and decisions are made on the basis of majority votes (Kalgraff Skjåk, in press).

Substantive questions. Country versions of the ISSP substantive module are produced following an Ask-the-same-question (ASQ) approach (Harkness, 2003): questions, response options and question order are fixed (Scholz, 2005). After the source questionnaire has been agreed upon, countries are responsible to produce their own versions. This is typically achieved through translation by a professional and/or by members of the research team with some degree of knowledge of English. The translation is then assessed by an expert, a group of people that gather different areas of expertise, or, in a few cases, comparing the source questionnaire to a "back translation" (an English rendition of the translation to be assessed). There is no clear policy on harmonization of instruments for the wording of questions across countries sharing a language, therefore differences in question wording across these countries are common.

Background variables. The design of sociodemographic variables in the ISSP follows an *ex-ante output* harmonization strategy. In this procedure, multicultural input is pursued at the design stages, although each country is free to formulate the questions as they consider best (Braun & Uher, 2003; Scholz, 2005). The goals of the indicators as well as the final harmonized categories are shared at the annual meetings (where every country is supposed to be present). Input from participating cultures can be shared at that stage. Before the module is fielded, the Zentral Archive shares these background variables in a data file format. Countries must take these considerations into account when formulating the specific questions. After data has been collected, countries send their variables to the Zentral Archive, and data are harmonized and merged there.

Answer scales in the ISSP. There are a number of features that are common to the questionnaires used in the ISSP across years. First, for every year there are a number of items that use agreement response scales, ranging from 8 to 31 items. Second, for all these items, the introduction reads in the source text: "To what extent do you agree or disagree…?" Following, statements are presented, and participants are expected to answer by choosing between one of five response categories, that typically are visually presented to the respondent. The source answer scale is a 5-point bipolar rating scale with the following labels: *Strongly agree, agree, neither agree not disagree, disagree, strongly disagree*.

Translation of answer scales in the ISSP. Translation of answer scales in the ISSP is carried out together with the rest of the questionnaire translation. Problems with this approach have been identified and discussed in the context of the ISSP since the mid 1990s (Harkness, et al., 1997). Despite the considerations that the report and subsequent

research have brought up, translation of the agreement answer scale in the ISSP is remains inconsistent both across countries and across years in some countries. For example, Denmark is still using a modified answer scale version, and Portugal keeps switching back and forth between a modified version and a close translation. It is unclear why Portugal went from a modified version in 1997, 1998, and 1999 to a closely translated one in the 2000, 2002, 2003, and 2004; however, it seems logical to assume that they used the modified version again in the 2005 survey because it repeated the Work Orientations module and they those to replicate the questions used in the 1997 survey. Researchers conducting secondary data analysis of the ISSP data would benefit from documentation on the rationale for why these labels change over time.

Specific Datasets Used in This Dissertation

The data analyzed in this dissertation come from surveys conducted by the ISSP within 1999 and 2004. Although the initial intention was to analyze data from 1997 until 2006, a number of data considerations led to the analysis of five years instead. The main problem stemmed from lack of variability in the main predictor of interest (answer scale version). When the number of countries using the modified answer scale version was lower than five, the cell size for the main independent variable (answer scale version) was considered insufficient for analysis.

In 1997, five countries had verbal labels that included a modifier for *agree*. However, one of those countries was Switzerland, and the modification appeared only on the Italian questionnaire. This was problematic because no variable in the dataset provided information regarding language of the interview. Through contact with the

organization in charge of the Swiss ISSP data collection, a "language of interview" variable was provided for years 2000 through 2007. The additional materials made analysis of the Swiss data from 2000 onwards available. However, no access to a language of interview variable was provided for the 1997 dataset, and it was not possible to build a variable that represented the observed answer scale variation within the Swiss data.

In 1998, only two countries used the modified intensity version of the answer scale. In 2001, only seven items were available to most countries. Similarly, only eight agreement items were used in 2006. Even though a larger number of items was available for 2005, a considerable portion of them was not asked in all countries, leading to complex missing patterns in the data. Therefore, this dataset was also discarded.

For every year considered in the analysis, at least twenty countries participated in the survey. For most of the years and countries the sample size is usually at least of 1,000 respondents (see Table 3.1). The topic of the ISSP modules varies from year to year, with replication occurring in several occasions. During the years considered in this dissertation, five different topics were studied: Environment (2000), Family and Changing Gender Roles (2002), National Identity (2003), Social Inequality (1999), and Citizenship (2004).

Limitations of the Data for the Purposes of this Dissertation

One limitation of this data analysis concerning the investigation of the impact of answer scale verbal labels on response styles stems from the nonexperimental nature of the dataset. Randomly assigning groups of respondents to different answer scale versions

would ensure that the observed effect is not due to other variables. To the best of my knowledge, there is no publicly available dataset involving a large number of nations that experimentally manipulates the verbal labels of answer scales. In the ISSP datasets, subjects are not assigned randomly to an answer scale version, and there are two main potential confounds: time of the interview (year) and content of the items.

Chapter 4. Effects of Changes in Answer Scale Verbal Labels on Response Style Differences Across Countries

Research investigating individual differences in response styles has a longer tradition, and explanations tend to be more empirically based and less speculative than explanations for country-level differences. The purpose of this chapter is to investigate what factors help explain country differences in response styles, taking in consideration variables that describe individuals (e.g., educational level, age) as well as variables that describe countries (e.g., mode of data collection used in the country). As discussed in Chapter1, there have been three main types of variables hypothesized to explain response styles and response style differences across countries. These are psychological aspects (e.g., personality and cognitive ability), cultural factors (e.g., norms of modesty and individualistic tendencies), and methodological factors (e.g., number of answer scale points, mode of data collection, and order of presentation of response choices). The analytical approach I adopt here—multilevel analysis or hierarchical linear models—allows the simultaneous investigation of variation between individuals and variation between nations. This way, it is possible to take into account differences between respondents that are due to belonging to a certain group. For this dissertation, these models estimate to what extent the response style of a respondent is due to common

variables that affect all respondents of his or her country, and to what extent the response style reflects his or her individual dispositions. Moreover, this analytical approach allows to examine the effect of country-level variables (such as Hofstede's cultural dimensions) on individual responses while controlling for individual-level variables (such as education or age). In this dissertation I examine three different types of variables simultaneously—individual factors, cultural factors, and methodological factors—involving a data structure at two levels—individuals and countries. Because methodological factors have been largely ignored in the literature, special attention is paid to the effect they have on response styles.

Hypotheses

I expected different response styles to be affected differently by the predictors considered in this dissertation. Table 4.1 summarizes the predicted relationships between the individual- and country-level predictors and the three response styles investigated in this study: extreme response style, middle response style, and acquiescence. The table contains one column per response style. Each cell reflects the expected direction of the relationship of the predictors with the three response styles. A plus sign indicates an expected positive relationship, a minus sign indicates an expected negative relation, a plus and minus sign separated by a slash suggest that there are theoretical arguments for either direction, a question mark indicates that evidence has been contradictory and theoretical arguments unclear, and NS stands for predicted lack of relationship. Justifications for the predictions of each model are discussed below.

Table 4.1.
Predicted relationships for each response style with all predictors

	Extreme Response Style	**Middle Response Style**	**Acquiescence**
Predictors			
Demographic variables – Individual level			
Age	+	?	+
Education (years)	+	-	+
Gender (0 = male, 1 = female)	+/-	+/-	+/-
Cultural factors – Country level			
Power distance	+/-	+/-	+/-
Individualism (vs. collectivism)	+/-	-	-
Uncertainty avoidance	?	?	?
Masculinity/femininity	+	-	-
Methodological factors – Country level			
Answer scales version (close translation vs. modified intensity)	+	-	NS
Data collection mode (self-administered vs. interviewer administered)	NS	NS	+

Hypotheses regarding demographic variables. Respondents with lower cognitive ability were expected to acquiesce and use the endpoints of a scale more often than those with higher cognitive skills. Age and education were used as proxies for cognitive ability. In addition, past research has shown that education is negatively related to the use of the middle category of an answer scale (Narayan & Krosnick, 1996), but no clear mechanism has been proposed as to why this happens. A possible explanation is that respondents see themselves as well informed and thus feel more confident about their answers than respondents with less formal education.

Based on previous research on the effect of survey topic on gender effects (Kane & Macaulay, 1993), I expected that gender differences in acquiescence or extreme response style manifested in some topics and not in others. For example, females may put

more effort into answering items from the Family and Changing Gender Roles questionnaire that refer to women than into items about other topics.

A complicating matter is that all the effect of these demographic variables on response styles may be different depending on the country. These interactions are accounted for in the models presented below. The mean years of education in a country was thus entered as a predictor in addition to the individual-level education variable.

Hypotheses regarding cultural factors.

The power distance dimension is related to how a culture deals with inequality. In cultures high in power distance, those in low-power positions tend to be more submissive towards those in high-power positions. This led researchers to expect that in cultures high in power distance, more acquiescent responding would be observed than in cultures low on power distance. In addition, Johnson et al (2005) postulate that "decisiveness" and "definitiveness" in communication may be a characteristic of high-power distance cultures. In such case, they argue, high power distance could be associated with higher levels of extreme responding. Findings regarding the relationship between power distance and response styles are inconsistent. Studies using Hofstede's measures tend to find significant relationships, therefore the expectation was that significant relationships would be found in this study as well.

The relationship of individualism vs. collectivism with extreme response style was hypothesized to be moderated by other factors. Different collectivistic nations have shown to exhibit high levels of extreme response style (e.g., Mexico) as well as low levels (e.g., Taiwan). Individuals in collectivistic cultures monitor their behaviors to

make sure that they are conforming to the norm of their in-group. If the norm in such collectivistic culture favors modesty as a virtue, perhaps the respondent will honor the modesty norm by choosing less “extreme” scale options. However, a collectivistic culture favoring other communication styles may lead to respondents choosing the endpoints more often. Therefore, depending of what the norm of the in-group is in a given collectivistic culture, the expected effect could be different.

Findings regarding the relationship between acquiescence and individualism seem somewhat more conclusive. The underlying argument is that individuals will choose agreeing responses more often in search for in-group harmony. Therefore, the expectation in this dissertation was that collectivistic countries would show higher levels of acquiescence than individualistic countries. Similarly, individualistic nations were expected to choose the middle point less often.

Masculinity has been linked to assertiveness and competitiveness (Hofstede, 2001). This may encourage respondents in masculine cultures to endorse endpoints more often than respondents in less masculine cultures (Johnson et al, 2005). Similarly, assertiveness may lead respondents to search disconfirming arguments more often; therefore, masculinity is expected to be negatively related to acquiescence. This may be reinforced by the association between femininity and modesty. Therefore, the hypothesis in this dissertation was that masculinity would be positively related to extreme response style and negatively related to acquiescence and middle response style.

Uncertainty avoidance. As described by Hofstede (2001), cultures high in uncertainty avoidance tend to seek structure. Ways to create structured situations include the establishment of rules, laws, and careful planning of situations. Cultures high in

uncertainly avoidance may thus orient their members towards avoiding ambiguity in their opinions, which would lead to predictions of higher endorsement of endpoints. Such relies on the assumption that endorsing an endpoint such as strongly agree would be less ambiguous than endorsing a scale point with lower intensity. In line with this, it is more difficult to introduce new ideas and concepts, which may lead to higher levels of acquiescence among cultures high on uncertainty avoidance (Johnson et al, 2005).

Hypotheses regarding methodological factors. The main hypothesis of this dissertation is that translation- and adaptation-related changes in verbal labels of answer scales have an impact on response distributions. In this chapter, the hypothesis is tested by studying the impact that answer scale version (closely translated vs. modified intensity) has on response style indicators. In particular, I expect this factor to account for part of the country-level variability in some of the response styles under study—namely, extreme response style and middle response style.

The predictions regarding the direction of the effect were based on calibration studies. Calibration studies have shown that words presented with modifiers such as the ones observed in the ISSP translations (e.g., *somewhat agree, partially agree*) receive lower "intensity scores" than words presented without them (V. A. Clarke, Ruffin, Hill, & Beamen, 1992; Kuz'min, 1981; Mohler, et al., 1998). I hypothesized that these ratings provided in laboratory settings would translate into a respondent's perceived intensity of the scale point in a survey interview setting. Consequently, the choice of verbal label for an answer scale point would affect how respondents match their subjective category with those categories provided by the researcher. It was hypothesized that the addition of an

attenuating intensity modifier to the *agree* and *disagree* scale points would increase the likelihood that a respondent chooses the endpoints of the 5-point scale. Similarly, respondents may be more prone to venturing an opinion when the offered scale point reads *somewhat agree* than when *agree* is presented, therefore lower middle response style was expected for countries with a modified intensity label.

Countries that conducted self-administered surveys were hypothesized to exhibit lower acquiescence scores than countries where surveys were conducted by an interviewer, based on the deference hypothesis. Past research has documented that primacy effects are more likely in measures obtained from visually administered instruments and recency effects more likely in orally presented surveys (Alwin & Krosnick, 1991). Both effects would increase the likelihood that one of the endpoints of the answer scale is selected. In addition, in the literature review I reported findings pointing to a tendency to select the extreme positive scale point more often in aural than in visual modes. However, most countries in the ISSP used show cards when conducting interviewer-administered surveys, which may cancel the effect. For that reason, no mode effect was expected on extreme or middle response style.

Method

The data analyzed here were collected for five different ISSP modules by each country. Table 3.1 in the previous chapter presented the sample sizes at both levels of analysis: number of countries and number of individuals within each country. Specific sample sizes are presented when discussing the results for each model in tables 4.5

through 4.24. Before discussing the models, the next section describes the dependent and independent variables and how each variable was operationalized.

Dependent variables: Operationalization of response styles

Step 1. Selection of the items

Response style indicators are usually computed using items that were designed to measure other constructs, namely, the variables of interest the questionnaire intends to measure (Baumgartner & Steenkamp, 2001). This approach has received criticisms because items used to compute stylistic variables could be substantially correlated (Arce-Ferrer, 2006; Couch & Keniston, 1960; Greenleaf, 1992). As a consequence, response styles measures obtained from these items may be—ironically—contaminated by substantive differences. A significant body of literature recognizes the need to include an eclectic, uncorrelated set of items in order to better capture tendencies that are independent of substantive positions (Arce-Ferrer, 2006; I. Clarke, 2001; Greenleaf, 1992). These authors advocate the use of items created specifically for the measurement of response styles, a recommendation valid for the design of new data collections, but not for analysis of secondary data. Selecting the least correlated items of a dataset can be an approximation when secondary data analyses are conducted. This can be done by selecting the set of items with the lowest item-total correlation (I. Clarke, 2000b; Greenleaf, 1992).

The items used in this dissertation to create response style indicators were all those that offered an agreement answer scale as response options. Tables A.1 through A.5 in the appendix present the wording for all the items used to create the dependent

variables, in the order that they appear in the source questionnaire. The ISSP target questionnaires are meant to keep this same question order.

Following Greenleaf's (1992) and Clarke's (2000b) recommendation, the intention was to use uncorrelated sets of items to create the response style indicators. However, not all the datasets considered in this dissertation contain enough items to follow this approach. Two of the datasets—namely, 1999 and 2004—have only 9 and 11 items, respectively, so that no set of a sufficiently large number of uncorrelated items could be obtained for these years. For the three remaining years (2000, 2002, and 2003), models were estimated using the full set of items for each year, and additional models were tested with the ten lowest correlated items.

Low-correlated items were selected using reliability scores and item-total correlation coefficients "backwards"; that is, items with the highest inter-total correlation were excluded one by one. The reduction in reliability was sizeable and the final item-total correlations satisfactorily low (Cronbach's alpha was .30, .22, and .008 for years 2000, 2002 and 2003, respectively). Tables A.6, A.8, and A.10 in the appendix present the item-total correlation coefficients as well as the alpha values for the complete sets of items for 2000, 2002 and 2003. Tables A.7, A.9, and A.11 include the final, reduced set of items, their item-total correlations and the alpha reliability value.

Step 2. Computation of extreme response style and acquiescence indicators

The computation of response style indicators was based on items selected in step 1 for all the countries with five ISSP datasets: 1999, 2000, 2002, 2003, and 2004. This includes four indicators for years 1999 and 2004 (one for each type of response style

indicator as described below), and eight for years 2000, 2002 and 2003 (each indicator was computed first using the total item set, and then using the low-correlated item set).

The four types of indicators of response style were:

1) A traditional Extreme Response Style indicator (ERS).

   ERS is estimated by dividing the number of endpoints chosen by the number of responses given to agreement scales for each respondent. ERS ranges between 0 and 1, with higher values representing higher levels of extreme response style.

2) A refined Response Style Indicator (RSI) (Thomas, Bremer, Terhanian, & Smith, 2006). RSI is a measure similar to ERS, but it is computed by following four steps:

   a) range-adjust all variables to range from 0 to 1;

   b) take the absolute value of the deviation for each measure from .5 (the midpoint of the 0 to 1 scale);

   c) compute the average of the total absolute deviations;

   d) compute the average of the deviations across items selected in step 1.

   Again, higher values of RSI

3) A traditional Middle Response Style indicator (MRS).

   MRS was measured as the proportion of times a respondent chooses the middle point of an answer scale.

4) An Acquiescence Response Style indicator (ARS).

   Several acquiescence indicators for bipolar answer scales have been proposed and used in the literature. Some indicators are computer using all scale points of the positive side of an answer scale, some use only one. Certain indicators, in

addition, lower acquiescence scores when respondents use the negative side of the scale. I computed four different acquiescence indicators: a) the proportion of responses on the positive side (*agree* and *strongly agree*) of the scale minus negative responses (*disagree* and *strongly disagree*), b) the proportion of responses to the *agreement* side; c) the proportion of *agree* responses; and d) the proportion of *agree* minus the proportion of *disagree*. Because of the subtraction involved, this indicator, unlike the previous ones, can take negative values. Respondents that choose answers from the "negative" side of the scale more often than from the "positive" side have negative scores. Theoretically, this variable can take any value from -1 to +1, where a score of -1 would represent a respondent that always answered using the negative side of the scale, +1 a respondent that always used the positive side of the scale, and 0 a respondent that used both sides equally often (or always the middle point).

The final choice of indicator was based on various statistical criteria. Correlations among these indicators were considerable high (the lowest being .41 for indicators b) and c) in 2002, the highest being .90 for indicators a) and c) in 1999). Indicator c) showed the lowest correlation with the other response style measures (ERS, RSI and MRS), indicating less overlap in the type of phenomenon that each are measuring. Preliminary analyses with the 2002 ISSP dataset showed that the sample distribution of acquiescence indicator c) deviates less than the remaining indicators from the normal distribution, and some of the analyses envisioned assume normality of the data. Therefore, indicator c) was selected to perform the analyses.

Tables A.12 through A.15 in the appendix present the mean proportion for each response style in all years and all countries.

Predictors

Country level predictors

Two types of country-level variables were included: cultural factors (Hofstede's dimensions scores) and methodological factors (answer scale version and mode of data collection).

- In the 1970s, IBM conducted a number of surveys among its employees with the objective of measuring work culture. Hofstede observed that those values manifested in the work environment were related to other features of the national culture of the countries where the surveys took place. From those surveys he derived a set of country scores that have shown significant relationships with various theoretically chosen predictors. The scores used in this dissertation are obtained directly from his book (Hofstede, 2001) and are presented in the appendix on table A.16.

  An attractive feature of Hofstede's scores for the analyses conducted in this dissertation is that the dimensions are meant to describe *national cultures*, and approach that fits the type of data from the ISSP.

- Answer scale version (ASV). In order to create a variable that identified whether a country used an answer scale version close to the source or not, I examined the wording of the agreement scale in each country as available

from the ISSP website. Many of these scales had been rendered in English by the ISSP members themselves, by research assistants at the University of Nebraska-Lincoln (including myself), and by professional translators for an Answer Scales research project directed by Janet Harkness. From that project, 31 translations of answer scales in 14 different languages were available, and they were used as the basis for assigning a code on the ASV variable.

For the languages that were not yet rendered in English, online resources helped identify the agreement answer scales in each language and to determine whether an intensity modifier had or had not been added. In most cases, the *strongly agree* verbal label was composed by a modifier—intended to convey the intensity of "strong"— and the word(s) signifying agreement. The *agree* verbal label usually had the same agreement word(s), which was sometimes accompanied by an additional term. Dictionaries were used to find the possible meanings of each word in the verbal labels, whether it was just one word or more. Attention was paid to understanding whether any component of the word(s) conveyed intensity. There was considerable variability in the type of modifiers used in the ISSP: "rather in agreement", "more agree than disagree", "agree to some extent", "partially agree", but most seemed to make the scale point less strong than "agree".

In addition to the examples of versions already mentioned in Chapter 3 (Portugal, Brazil, Switzerland, Japan, and Australia), several other countries in the ISSP have either modified intensity or adapted versions. For example, Denmark consistently uses the same modified intensity version; Russia,

France and Spain are some of the countries that have used both versions across the years. The answer scale version variable was created from the inspection of these answer scales. Countries where a modifier of the second and fourth scale points was present were coded as “modified intensity version”. For the remaining countries, two codes were possible: 1) if the translation retained a general connotation of agreement and no modifier was present, it was coded as close translation; 2) if variations were found in how *agree* was conveyed, it was coded as missing. These variations were not included in the analyses because each specific variation was only present in one country. Thus, no statistical inference was possible. I should note that this type of variation did not only occur in translation, but also in English speaking countries. The answer scale of the United States in 2002, for example, had four scale points instead of five, and therefore was considered a missing value in the answer scale version variable. Table 4.2 shows the wording of the *agree* scale point for all years and countries analyzed. This includes translations for countries where languages other than English were spoken, as well as the wording intended for English-speaking respondents.

Table 4.2.
Rendition of scale point *agree* in each country for years 1999, 2000, 2002, 2003, 2004

	1999 - 11 agreement statements	**2000 - 28 agreement statements**	**2002 - 25 agreement statements**	**2003 - 29 agreement statements**	**2004 - 9 agreement statements**
Australia	Yes	=	Yes	Agree	Agree
Austria	Stimme zu	**Stimme eher zu**	Stimme zu	**Stimme eher zu**	**Stimme eher zu**
Brazil	=	=	**Concordo em parte**	=	**Concorda em parte**
Bulgaria	Съгласен	съгласен	съгласен	**донякъде съгласен**	CbrnaceH
Canada - English	Agree	Agree	=	Agree	Agree
Canada - French	D'accord	D'accord	=	D'accord	D'accord
Cyprus	συμφωνώ	=	συμφωνώ	=	Not available
Czech republic	**Spíše souhlasím**	**Spíše souhlasím**	**Spíše souhlasím**	**Spíše souhlasím**	**Spíše souhlasím**
Denmark	**Delvis enig**	**Delvis enig**	**Delvis enig**	**Delvis enig**	**Delvis enig**
Finland – Finnish	=	Samaa mielta	Samaa mielta	Samaa mielta	Samaa mielta
Finland – Swedish	=	Av samma asikt	Av samma asikt	Av samma asikt	Av samma asikt
Flanders	=	=	Mee eens		Mee eens
France	**Plutot d'accord**	=	**Plutot d'accord**	**Plutot d'accord**	**Plutot d'accord**
Germany	**Stimme eher zu**	**Stimme eher zu**	Stimme zu	Stimme zu	Stimme zu
Great Britain	Agree	Agree	Agree	Agree	Agree
Hungary	Egyetért	=	egyetért	egyetért	egyetért
Ireland	=	Not available	Agree	Agree	Agree
Israel-Arabic	موافق	=	=	موافق	موافق
Israel-Hebrew	מסכים	מסכים	מסכים	מסכים	מסכים
Italy	=	=	=	=	=
Japan	Dochiraka to ieba sou omou	Dochiraka to ieba sansei	Dochiraka to ieba sou omou	Dochiraka to ieba sou omou	Dochiraka to ieba sou omou
Latvia - Latvian	Piekritu	Piekritu	Piekritu	Piekritu	Piekritu
Latvia - Russian	Согласен	Согласен	Согласен	Согласен	Согласен
Mexico	=	De acuerdo	De acuerdo	=	De acuerdo

Note: Cells with bold fonts correspond to modified intensity verbal labels. Cells with a dash indicate countries that did not collect data for a given year. The table also indicate those countries that did participate but where no questionnaire was available for a given language.

Table 4.2. (continued)
Rendition of scale point *agree* in each country for years 1999, 2000, 2002, 2003, 2004

	1999 - 11 agreement statements	**2000 - 28 agreement statements**	**2002 - 25 agreement statements**	**2003 - 29 agreement statements**	**2004 - 9 agreement statements**
Netherlands	=	=	Mee eens	Mee eens	Mee eens
New Zealand	Agree	Agree	Agree	Agree	Agree
Northern Ireland	Agree	Agree	Agree	=	=
Norway	Enig	Enig	Enig	Enig	Enig
Poland	Zgadzam sie	=	Zgadzam sie	Zgadzam sie	Zgadzam sie
Portugal	**Concorda parcialmente**	Concordo	Concorda	Concorda	Concorda
R Chile	De acuerdo	De acuerdo	De acuerdo	De acuerdo	De acuerdo
R Philippines	Sumasang-ayon	Sumasang-ayon	Sang-ayon	Sang-ayon	Sumasang-ayon
Russia	**Скорее Согласен**	Согласен	Согласен	**Скорее Согласен**	**Скорее Согласен**
Slovakia	**Skôr súhlasím ako nesúhlasím**	=	**Skôr súhlasím ako nesúhlasím**	Súhlasím	Súhlasím
Slovenia	Soglasam	Soglasam	Soglasam	Se strinjam	**v celoti se strinjam**
South Africa	=	=	=	=	
South Korea	=	=	=	=	
Spain	De acuerdo	De acuerdo	De acuerdo	De acuerdo	De acuerdo
Sweden	Instammer	Instammer	Instammer	Instammer	Instammer
Switzerland French	=	**Plutôt en désaccord**	**Plutôt d'accord**	D'accord	**Plutôt d'accord**
Switzerland German	=	**Stimme eher zu**	Stimme zu	Einverstanden	Stimme zu
Switzerland Italian	=	D'accordo	D'accordo	Sono d'accordo	D'accordo
Taiwan	=	=	同意	同意	同意
United States	Agree	Agree	Agree	Agree	Agree
Uruguay	=	=	=	De acuerdo	De acuerdo
Venezuela	=	=	=	De acuerdo	De acuerdo

Note: Cells with bold fonts correspond to modified intensity verbal labels. Cells with a dash indicate countries that did not collect data for a given year. The table also indicate those countries that did participate but where no questionnaire was available for a given language.

Table 4.3 provides a list those countries and languages where the answer scale version had labels with modified intensity, for the years considered in the analysis.

Table 4.3.
List of countries where a modified intensity answer scale version was used

1999	2000	2002	2003	2004
Czech Republic	Austria	Brazil	Austria	Austria
Denmark	Czech Republic	Czech Republic	Bulgaria	Brazil
France	Denmark	Denmark	Czech Republic	Czech Republic
Germany	Germany	France	Denmark	Denmark
Portugal	Switzerland (German)	Slovakia	France	France
Russia	Switzerland (French)	Switzerland (French)	Russia	Russia
Slovakia				Slovenia
				Switzerland (French)
n = 7	n = 6	n = 6	n = 6	n = 8

- Data collection mode. Initially, the intention was to have two different data collection mode variables: one that differentiated whether the survey was self-administered or interviewer-administered, and a second one identifying whether the answer scales had been visually presented or not. However, the variable available in the ISSP datasets does not provide information of whether visual aid was offered to the respondent for every country in every survey. Therefore, only one variable was used, with two categories: self-administered surveys vs. interviewer-administered surveys.

Individual level variables

The independent variables were:

a) Gender was measured in all countries, and it was coded with a 0 for men and a 1 for women.

b) Age was measured in years.

c) Education. Two education variables are available in ISSP datasets that provide values harmonized across all countries: years of schooling and degree achieved. Due to differences in education systems across the globe, none of these variables are exempt from problems when it comes to comparing levels of education across countries. Advantages of one over the other are debatable. However, to my knowledge, no empirical research has shown which one of the two represents a better proxy for the real variable of interest, namely, cognitive ability or social status. "Degree achieved" may be a better proxy for comparisons of social status, whereas years of schooling may better represent cognitive ability. However, but the debate is still open. In practice, managing an 8-category ordinal variable can be cumbersome to interpret, and cut-off points often need to be selected for the variable to yield meaningful results. These cut-off points may be different for different countries. Therefore, I chose the quantitative variable "years of formal education" for these analyses. Nonparametric correlation between both variables was moderately high (usually higher than .75), and the models presented below did not yield different results when using either variable as predictor.

Primary Findings

Across-country and within country variation in response styles was examined using multilevel models estimated in SAS PROC MIXED (v9.1). Models differing in

fixed effects were compared using maximum likelihood (ML), and models that compared random effects alone were compared using restricted maximum likelihood (REML).

The Intraclass Correlation Coefficients (ICCs) show that countries differ in response styles. The ICCs were computed as the ratio of the country-level variance component to the total variance (pseudo-R^2). For all years, the ICCs show that a greater part of the total variance is due to individual differences in response styles than to country differences. This is typically the case when studying social phenomena. However, there was a considerable amount of variance at the country level.

Table 4.4 presents the intraclass correlation coefficients for each year and each response style indicator. The ICC values vary considerably from year to year. Given that there is no clear increasing or decreasing trend across time, it seems plausible that the variation is due to the topic of the questionnaire and to the specific countries that participated in each given year.

There is also variation across response styles. The relative magnitude of between-country variation as compared to within-country (between respondents) variation is larger for extreme response style than for acquiescence or middle response style. This means that differences in extreme response style are affected by the country of origin to a larger extent than acquiescence and middle response style. Each response style is described separately in the following sections.

Because a considerable portion of the variance was accounted for by country-level differences, a multilevel modeling approach was deemed appropriate.

Table 4.4.
Percentage of country-level variance for each response style and year

	1999	2000	2002	2003	2004
ERS	.24	.16	.29	.14	.12
RSI	.20	.11	.23	.12	.10
MRS	.09	.08	.11	.12	.08
ARS	.11	.20	.16	.14	.08

Models

Separate multilevel models were estimated for each response style; the effect of the predictors on response style indicators was evaluated using four main multilevel models across all years considered. There were, thus, four models, four response style indicators (for three response styles), and five years, resulting in 80 multilevel models reported in this document. An additional 48 models were estimated using the indicators obtained from the low correlated items. These models showed similar results to those obtained with the full set of items. All these models are described in tables 4.5 through 4.24. This section describes the estimated models and how to interpret the tables provided in the remainder of the chapter.

The first of the four models was the *baseline* model, also known as empty model or intercept only model. This model provides information about how much each dependent variable varies across countries, using the ICCs. With this information, one can assess to what extent differences in response styles across all respondents are due to individual factors, and to what extent they are due to differences across countries. In the baseline model, no predictors are included that explain why variance at either level occurs. In addition to giving an estimation of the proportion of variance that is accounted for by living in a certain country, the baseline model provides a reference for estimation

of the effect of different predictors on the outcome by comparison of overall level of fit between different models.

The second and third columns of each table provide the estimates, sample size, and fit statistics for the empty model. The information contained in the first two rows allows the estimation of the ICCs (the country level variance of the empty model divided by the total variance). The model fit statistics of the empty model establish the baseline for model comparison. ML deviance can be used to compare nested models, whereas AIC and BIC are information criteria to compare non-nested models. In the case of nested models, the difference of the deviances approximates a chi-square distribution. If the difference between two models is statistically significant,

Comparison of "total country level variance" as well as the "residual variance" across models serves as basis for estimation of the effect size of the predictors. That is, by monitoring whether the estimated variance at each level augments or diminishes when predictors are added to the model, one can estimate the magnitude of the effect of such predictors. For example, in table 4.5 the total country level variance is 0.0154 for the empty model and 0.0104 in the model with answer scale version as predictor. Therefore, the remaining variance across countries after accounting for answer scale version is 32% smaller than the original country-level variance. Comparing the same value against the remaining country-level variance in model 3, it can be seen that after taking into account data collection mode (in addition to answer scale version) only half of the country level variance remains to be explained. At the same time, the residual variance at the individual level has not decreased with the inclusion of answer scale version or mode as predictors. In the last model, after including age, gender, education, and the four cultural dimensions,

88% of the country variance has been explained, as well as 5% of the differences within countries. Another way to appraise model fit is by estimating R^2, calculated here as the correlation between the observed values and the values predicted by the model, and presented in the tables for the model containing all the predictors.

The previous paragraph has already advanced that the fourth and fifth columns present the estimates and model fit for a model including the methodological variable considered to be most important for each response style. This variable was answer scale version for extreme and middle response style, and mode of data collection for acquiescence.

The third model tested the effect of both methodological variables at once. The goal was to estimate what proportion of the country-level variance was accounted for by the methodological variables. A more practical reason for considering these variables alone was the loss of statistical power related to the fact that adding the cultural dimension variables resulted in considerable sample size reduction. The sixth and seventh columns show the estimates for these models.

The last model tested the simultaneous effect of all predictors considered theoretically relevant. The objective was to examine how much of the variance of each response style (including differences across respondents of the same country and differences across countries) was explained by the predictors, and which predictors were significantly related to the response styles. The eighth and ninth columns contain the estimates and fit statistics of these models.

Table 4.5 contains the ERS models and information for 1999, and tables 4.6, 4.7, 4.8, and 4.9 for years 2000, 2002, 2003, and 2004, respectively. Tables 4.10 through 4.14

present the estimates and fit statistics for the models where RSI is the dependent variable. Tables 4.15 through 4.19 describe the models for MRS, and tables 4.20 through 4.24 the models for ARS.

Additional models not presented here were tested when potential confounds were considered possible. For example, the effect of educational attainment at country level was tested for all response styles and years, but showed no significant effect.

Extreme response style

Methodological factors. As expected, countries that used a modified intensity label showed higher scores in ERS than countries where the response category was a close translation of the source text. This effect was significant across all years; the largest difference (about 30 percentage points) is observed in 2002, and the smallest (about 8 percentage points) in 2000. The "Pseudo-R^2 – Country variance" values show that answer scale version explained between a 28% and a 67% of the country-level variance of ERS, and between a 17% and a 46% of RSI. Even though the effect of answer scale version is more pronounced for ERS than for RSI, the effect on RSI is significant and in the same direction as the effect on ERS for four of the five years—and nonsignificant only for the 2000 module on environment. Furthermore, the inclusion of other predictors in models 2 and 3 did not affect the relationship between answer scale version and extreme response style.

The average extreme response values for countries with translated answer scales and countries with modified intensity answer scales are depicted in figure 4.1 for ERS and 4.2 for RSI. In all years and in both variables, respondents in countries with a

modified intensity verbal label chose endpoints more often than respondents from countries where a translated version was implemented.

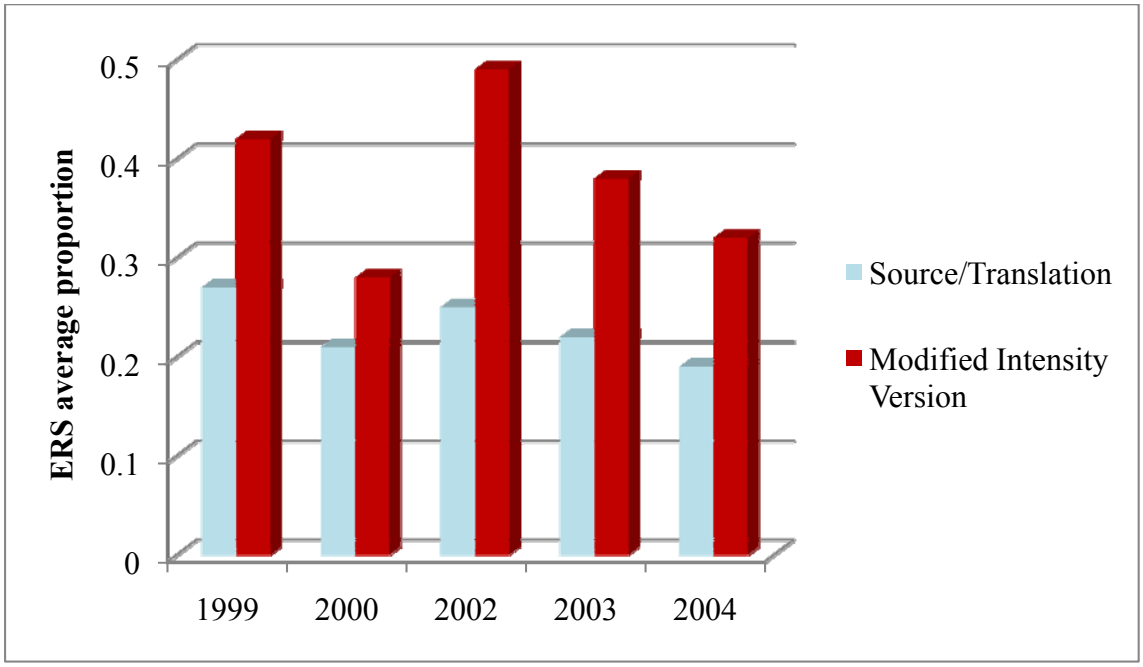

Figure 4.1. Average extreme response style proportion by answer scale version group

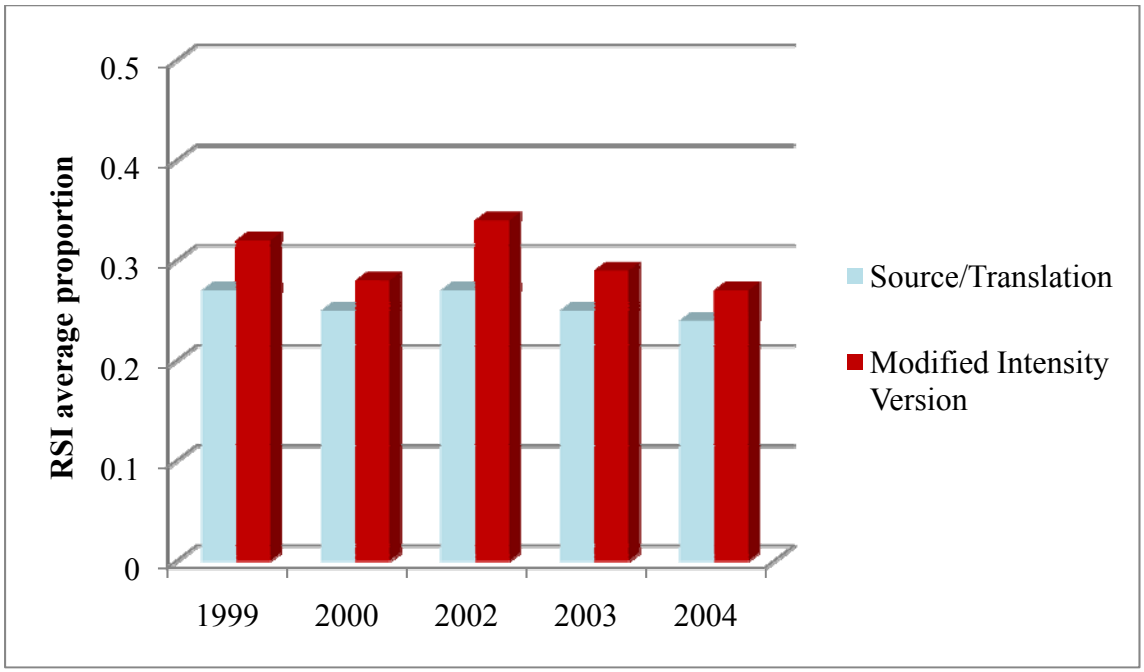

Figure 4.2. Average response style indicator proportion by answer scale version group

As can be seen on tables 4.5 through 4.9, answer scale label was the only predictor that showed a significant, consistent effect across all years and both indicators (ERS and RSI). Mode of data collection showed the expected impact on both indicators of extreme response style only in 1999. For that year, when the survey was administered by an interviewer, respondents used the endpoints more often (Wald(1) = 0.21, p = .02) than respondents that answered self-administered questionnaires. For all other years, the effect of more on extreme response style was nonsignificant both for ERS and RSI.

Demographic variables. Education has a significant effect for all years on ERS. However, the direction of the effect varies depending on the year. For 2000 (the environment module) and 2002 (the module on family and changing gender roles), more educated respondents tended to use extremes mode than low educated respondents. For the remaining years, the relationship goes in the expected direction: high educated respondents had lower extreme response style scores. However, two of these expected relationships were not significant when the outcome was RSI. The impact of education on extreme response style did not vary significantly across countries for any year (results not presented). This does not mean that the effect is necessarily constant for all countries, but that when taking all countries as a whole, the overall differences do not reach statistical significance.

Similarly, gender and age showed significant relationships both with RSI and ERS, but the sign of the effect differed across years.

Cultural factors. Few of the relationships between the cultural dimensions and extreme response style indicators were significant. Uncertainty avoidance is positively

related to extreme response style in four out of ten “full” models (in 1999 for ERS, in 2000 for RSI, and in 2004 for both indicators).

Table 4.5.
Multilevel models for 1999, Extreme Response Style as outcome: estimates, standard errors and fit statistics

1999 ERS Models	**Model 1 (Baseline) ICC=0.24**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0154***	0.0041	0.0104***	0.0029	0.0076***	0.0008	0.0019**	0.0007
Residual variance	0.0495***	0.0004	0.0501***	0.0004	0.0507***	0.0003	0.0468***	0.0005
Pseudo-R^2 – *Country variance*			32.45%		50.58%		87.92%	
Pseudo-R^2 – *Individual variance*							5.46%	
Number Countries	29		27		24		15	
Number Total Cases	33116		30217		26560		16427	
Predictors								
Intercept	0.3054***	0.0232	0.2664***	0.0230	0.1934***	0.0308	0.0615	0.1417
Methodological factors – Country level								
Answer scale version			0.1574***	0.0450	0.1458***	0.0441	0.2043***	0.0571
Mode of data collection					0.1243***	0.0371	0.2148*	0.0909
Demographic variables – Individual level								
Age							0.0000	0.0001
Education (years)							-0.0026***	0.0005
Gender							-0.0048	0.0034
Cultural factors – Country level								
Power distance							0.0002	0.0006
Individualism/collectivism							0.0025	0.0019
Uncertainty avoidance							-0.0005	0.0012
Masculinity/femininity							-0.0014	0.0008
Fit statistics								
R^2							0.34	
ML deviance	-5399		-4590		-3688		-3627	
AIC	-5393		-4582		-3678		-3603	
BIC	-5389		-4576		-3673		-3594	

*p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.6.
Multilevel models for 2000, Extreme Response Style as outcome: estimates, standard errors and fit statistics

2000 ERS Models	**Model 1 (Baseline) ICC=0.16**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0063***	0.0017	0.0046***	0.0013	0.0026***	0.0008	0.0016**	0.0006
Residual variance	0.0330***	0.0003	0.0339***	0.0003	0.0335***	0.0003	0.0314***	0.0003
Pseudo-R^2 – *Country variance*			27.69%		59.24%		74.35%	
Pseudo-R^2 – *Individual variance*							4.90%	
Number Countries	29		26		24		19	
Number Total Cases	29455		25531		23568		17769	
Predictors								
Intercept	0.2209***	0.0148	0.2096***	0.0152	0.1769***	0.0162	-0.0053	0.1463
Methodological factors – Country level								
Answer scale version			0.0668*	0.0316	0.0668*	0.0259	0.0868*	0.0350
Mode of data collection					0.0424*	0.0211	0.0247	0.0455
Demographic variables – Individual level								
Age							0.0000	0.0001
Education (years)							0.0026***	0.0004
Gender							-0.0094***	0.0027
Cultural factors – Country level								
Power distance							0.0012	0.0008
Individualism/collectivism							0.0015	0.0015
Uncertainty avoidance							0.0011	0.0008
Masculinity/femininity							-0.0005	0.0007
Fit statistics								
R^2							0.15	
ML deviance	-16705		-13841		-13092		-10987.5	
AIC	-16699		-13833		-13082		-10963.5	
BIC	-16695		-13828		-13076		-10952.1	

*p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.7.
Multilevel models for 2002, Extreme Response Style as outcome: estimates, standard errors and fit statistics

2002 ERS Models	**Model 1 (Baseline) ICC=0.29**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0167***	0.0038	0.0082***	0.0020	0.0082***	0.0021	0.0073***	0.0023
Residual variance	0.0403***	0.0003	0.0402***	0.0003	0.03996***	0.0003	0.0399***	0.0003
Pseudo-R^2 – *Country variance*			50.61%		50.79%		56.38%	
Pseudo-R^2 – *Individual variance*							0.92%	
Number Countries	38		33		31		21	
Number Total Cases	45917		42102		39668		27981	
Predictors								
Intercept	0.3036***	0.0210	0.2511***	0.01752	0.2416***	0.02756	0.3923	0.1881
Methodological factors – Country level								
Answer scale version			0.2374	0.0411	0.2249***	0.0444	0.2874***	0.0587
Mode of data collection					0.0079	0.0288	-0.0441	0.0696
Demographic variables – Individual level								
Age							-0.0008***	0.0001
Education (years)							0.0025***	0.0003
Gender							0.0385***	0.0024
Cultural factors – Country level								
Power distance							-0.0009	0.0012
Individualism/collectivism							-0.0025	0.0019
Uncertainty avoidance							0.0005	0.0012
Masculinity/femininity							0.0002	0.0011
Fit statistics								
R^2							0.50	
ML deviance	-16913.5		-15696.1		-14987.3		-10594.2	
AIC	-16907.5		-15688.1		-14977.3		-10570.2	
BIC	-16902.6		-15682.1		-14970.1		-10557.7	

*p < .05; ** p < 0.01; ***p < .001.

Table 4.8.
Multilevel models for 2003, Extreme Response Style as outcome: estimates, standard errors and fit statistics

2003 ERS Models	**Model 1 (Baseline) ICC=0.14**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0062***	0.0015	0.0021***	0.0005	0.0021***	0.0005	0.0016***	0.0005
Residual variance	0.0384***	0.0003	0.0363***	0.0003	0.0362***	0.0003	0.0346***	0.0003
Pseudo-R^2 – *Country variance*			66.58%		66.88%		73.85%	
Pseudo-R^2 – *Individual variance*							9.93%	
Number Countries	35		31		30		22	
Number Total Cases	44604		39918		38619		28528	
Predictors								
Intercept	0.2585***	0.0133	0.2221***	0.0092	0.2218***	0.0131	0.2584*	0.0890
Methodological factors – Country level								
Answer scale version			0.1532***	0.0208	0.1436***	0.0224	0.1120***	0.0334
Mode of data collection					0.0004	0.0168	-0.0416	0.0333
Demographic variables – Individual level								
Age							0.0002*	0.0001
Education (years)							-0.0012***	0.0003
Gender							-0.0210***	0.0022
Cultural factors – Country level								
Power distance							-0.0004	0.0006
Individualism/collectivism							-0.0008	0.0008
Uncertainty avoidance							0.0007	0.0006
Masculinity/femininity							0.0004	0.0005
Fit statistics								
R^2							0.21	
ML deviance	-18654.1		-18957.8		-18400.9		-14938.5	
AIC	-18648.1		-18949.8		-18390.9		-14914.5	
BIC	-18643.4		-18944		-18383.9		-14901.4	

*p < .05; ** p < 0.01; ***p < .001.

Table 4.9.
Multilevel models for 2004, Extreme Response Style as outcome: estimates, standard errors and fit statistics

2004 ERS Models	**Model 1 (Baseline) ICC=0.12**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0067***	0.0015	0.0042***	0.0010	0.0043***	0.0010	0.0028***	0.0008
Residual variance	0.0482***	0.0003	0.0480***	0.0003	0.0477***	0.0003	0.0460***	0.0004
Pseudo-R^2 – *Country variance*			36.61%		35.80%		58.21%	
Pseudo-R^2 – *Individual variance*							4.61%	
Number Countries	42							
Number Total Cases	48622							
Predictors								
Intercept	0.2148***	0.0127	0.1904***	0.0122	0.1705***	0.0209	0.0506***	
Methodological factors – Country level								
Answer scale version			0.1063***	0.0261	0.1024***	0.0281	0.1321***	0.0313
Mode of data collection					0.0282	0.0249	-0.0305	0.0333
Demographic variables – Individual level								
Age							0.0006***	0.0001
Education (years)							-0.0017***	0.0003
Gender							-0.0173***	0.0024
Cultural factors – Country level								
Power distance							0.0009	0.0007
Individualism/collectivism							0.0000	0.0009
Uncertainty avoidance							0.0016*	0.0007
Masculinity/femininity							-0.0003	0.0006
Fit statistics								
R^2							0.32	
ML deviance	-9258.2		-8281.3		-8094.9		-7411.5	
AIC	-9252.2		-8273.3		-8084.9		-7387.5	
BIC	-9246.9		-8266.9		-8077.1		-7372	

*p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.10.
Multilevel models for 1999, Response Style Indicator as outcome: estimates, standard errors and fit statistics

1999 RSI Models	**Model 1 (Baseline) ICC=0.21**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0016***	0.0004	0.0012***	0.0003	0.0008***	0.000	0.0002**	0.0001
Residual variance	0.0063***	0.0000	0.0063***	0.0001	0.0063***	0.000	0.0060***	0.0001
Pseudo-R^2 – *Country variance*			28.08%		53.26%		90.64%	
Pseudo-R^2 – *Individual variance*							4.66%	
Number Countries	29		27		24		15	
Number Total Cases	33116		30217		26560		16427	
Predictors								
Intercept	0.2834***	0.0075	0.2723***	0.0077	0.2451***	0.0097	0.2242***	0.0408
Methodological factors – Country level					0.0460**	0.0139	0.0448*	0.0165
Answer scale version			0.0487**	0.0151	0.0461***	0.0117	0.0377	0.0262
Mode of data collection								
Demographic variables – Individual level								
Age							0.0001**	0.0000
Education (years)							-0.0007**	0.0002
Gender							-0.0018	0.0012
Cultural factors – Country level								
Power distance							-0.0001	0.0002
Individualism/collectivism							0.0001	0.0005
Uncertainty avoidance							0.0004	0.0003
Masculinity/femininity							-0.0002	0.0002
Fit statistics								
R^2							0.28	
ML deviance	-73633.2		-67361.5		-59087.2		-37340	
AIC	-73627.2		-67353.5		-59077.2		-37316	
BIC	-73623.1		-67348.3		-59071.3		-37307.5	

***p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.11.
Multilevel models for 2000, Response Style Indicator as outcome: estimates, standard errors and fit statistics

2000 RSI Models	**Model 1 (Baseline) ICC=0.11**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0006***	0.0002	0.0005***	0.000	0.0003***	0.0001	0.0002**	0.0001
Residual variance	0.0047***	0.0000	0.0047***	0.000	0.0047***	0.0000	0.0045***	0.0000
Pseudo-R^2 – *Country variance*			17.21%		47.70%		70.66%	
Pseudo-R^2 – *Individual variance*							6.08%	
Number Countries	29		26		24		19	
Number Total Cases	29455		25531		23568		17769	
Predictors								
Intercept	0.2553***	0.0046	0.2522***	0.0051	0.2394***	0.0057	0.1702**	0.0489
Methodological factors – Country level								
Answer scale version			0.02257*	0.0106	0.0225*	0.0092	0.0253*	0.0118
Mode of data collection					0.0193**	0.0074	-0.0001	0.0153
Demographic variables – Individual level								
Age							0.0002***	0.0000
Education (years)							0.0013***	0.0001
Gender							-0.0050***	0.0010
Cultural factors – Country level								
Power distance							0.0004	0.0003
Individualism/collectivism							0.0004	0.0005
Uncertainty avoidance							0.0007**	0.0003
Masculinity/femininity							0.0000	0.0002
Fit statistics								
R^2							0.21	
ML deviance	-73915.1		-64017.5		-59394.4		-45719.4	
AIC	-73909.1		-64009.5		-59384.4		-45695.4	
BIC	-73905		-64004.5		-59378.5		-45684.1	

***p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.12.
Multilevel models for 2002, Response Style Indicator as outcome: estimates, standard errors and fit statistics

2002 RSI Models	**Model 1 (Baseline) ICC=0.21**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0014***	0.0003	0.0008***	0.0002	0.0008***	0.0002	0.0006***	0.0002
Residual variance	0.0045***	0.0000	0.0044***	0.0000	0.0044***	0.0000	0.0043***	0.0000
Pseudo-R^2 – *Country variance*			41.64%		45.11%		59.02%	
Pseudo-R^2 – *Individual variance*							5.03%	
Number Countries	38		33		31		28	
Number Total Cases	46053		42235		39800		28083	
Predictors								
Intercept	0.2851***	0.0060	0.2730***	0.0055	0.2654***	0.0084	0.306***	0.0526
Methodological factors – Country level								
Answer scale version			0.0642***	0.01286	0.0599***	0.0135	0.0794***	0.0164
Mode of data collection					0.0084	0.0088	-0.0037	0.0195
Demographic variables – Individual level								
Age							-0.0002***	0.0000
Education (years)							0.0008***	0.0001
Gender							0.0122***	0.0009
Cultural factors – Country level								
Power distance							-0.0002	0.0003
Individualism/collectivism							-0.0008	0.0005
Uncertainty avoidance							0.0002	0.0003
Masculinity/femininity							0.0001	0.0003
Fit statistics								
R^2							0.48	
ML deviance	-117602		-108992		-102654		-73191	
AIC	-117596		-108984		-102644		-73167	
BIC	-117591		-108978		-102637		-73155	

***p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.13.
Multilevel models for 2003, Response Style Indicator as outcome: estimates, standard errors and fit statistics

2003 RSI Models	**Model 1 (Baseline) ICC=0.12**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0007***	0.0002	0.0004***	0.0001	0.0003***	0.0001	0.0002***	0.0001
Residual variance	0.0052***	0.0000	0.0050***	0.0000	0.0050***	0.0000	0.0047***	0.0000
Pseudo-R^2 – *Country variance*			46.30%		49.26%		70.41%	
Pseudo-R^2 – *Individual variance*							8.95%	
Number Countries	35		31		30		22	
Number Total Cases	44604		39918		38619		28528	0.0536
Predictors								
Intercept	0.2637	0.0044	0.2537	0.0038	0.2491***	0.0054	0.2703***	0.0313
Methodological factors – Country level								
Answer scale version			0.0438	0.0087	0.0395***	0.0091	0.0226	0.0117
Mode of data collection					0.0083	0.0069	-0.0129	0.0117
Demographic variables – Individual level								
Age							0.0002***	0.0000
Education (years)							0.0000	0.0001
Gender							-0.0088***	0.0008
Cultural factors – Country level								
Power distance							-0.0002	0.0002
Individualism/collectivism							-0.0006	0.0003
Uncertainty avoidance							0.0003	0.0002
Masculinity/femininity							-0.0004*	0.0002
Fit statistics								
R^2							0.49	
ML deviance	-108051		-98428.1		-95255.5		-71806.6	
AIC	-108045		-98420.1		-95245.5		-71782.6	
BIC	-108040		-98414.3		-95238.5		-71769.5	

***p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.14.
Multilevel models for 2004, Response Style Indicator as outcome: estimates, standard errors and fit statistics

2004 RSI Models	**Model 1 (Baseline) ICC=0.10**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0009***	0.0002	0.0007***	0.0002	0.0007***	0.0002	0.0005***	0.0001
Residual variance	0.0074***	0.0000	0.0073***	0.0001	0.0073***	0.0001	0.0070***	0.0001
Pseudo-R^2 – *Country variance*			21.11%		24.13%		46.87%	
Pseudo-R^2 – *Individual variance*							5.93%	
Number Countries	42		37		35		27	
Number Total Cases	48622		42690		40352		31096	
Predictors								
Intercept	0.2461	0.0046	0.2394***	0.0049	0.2279***	0.008196	0.1864***	0.0419
Methodological factors – Country level								
Answer scale version			0.0325**	0.0105	0.02865**	0.01103	0.0354**	0.0126
Mode of data collection					0.0176	0.009756	-0.0106	0.0134
Demographic variables – Individual level								
Age							0.0004**	0.0000
Education (years)							0.0003	0.0001
Gender							-0.0092**	0.0010
Cultural factors – Country level								
Power distance							0.0003	0.0003
Individualism/collectivism							-0.0002	0.0004
Uncertainty avoidance							0.0006*	0.0003
Masculinity/femininity							0.0005	0.0002
Fit statistics								
R^2							0.32	
ML deviance	-100162		-88540.7		-84089.8		-65973.8	
AIC	-100156		-88532.7		-84079.8		-65949.8	
BIC	-100151		-88526.3		-84072		-65934.3	

***p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Middle response style

Methodological factors. Contrary to my expectations, answer scale version did not have a consistent significant impact on middle response style. Only for 1999 was this relationship significant, and only when the variable was considered alone or together with mode. The effect disappeared after adding other predictors to the model. Figure 4.3 shows that respondents answering modified intensity answer scales are only slightly less likely to select the middle response.

Demographic variables. Age, gender and education had a significant impact on the middle response style indicator. The effect was negative for education and age. As expected, more educated respondents tend to choose the middle response option less often than respondents with less years of formal schooling. In addition, older respondents also answer with the middle option less often than younger respondents. The effect of education did not significantly vary across countries.

The direction varied for gender across years. In general, females chose the middle point more often than males (years 2000, 2003, and 2004). For the Family and Changing Gender Roles, as expected, the relationship changed: females were more likely than men to give answers other than *neither agree nor disagree.* Given that more items present statements about women than about men, the change of direction is not surprising. This finding suggests that the mechanism related to the changes in direction observed in other variables could be driven by the content of the questionnaire.

Cultural factors. Findings regarding cultural dimensions are also inconsistent across years for middle response style. The most consistent is masculinity vs. femininity,

that shows a negative relationship in three years. Uncertainty avoidance is also negatively related to middle response style for 1999 and 2000.

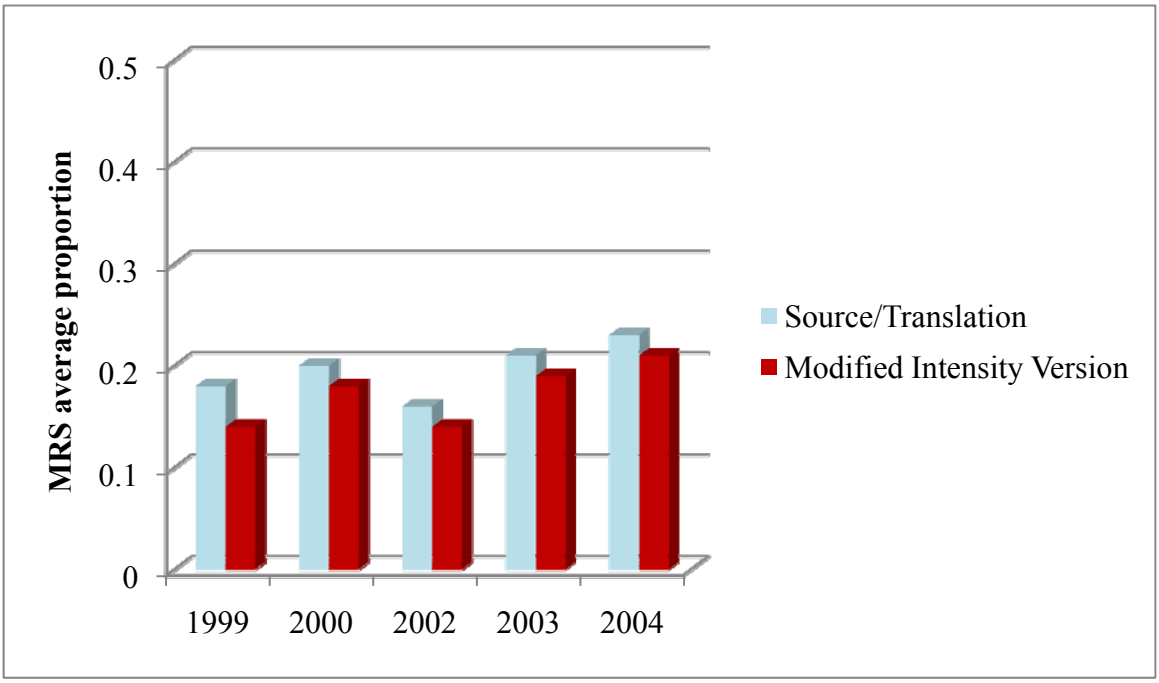

Figure 4.3. Average middle response style proportion by answer scale version group

Table 4.15.
Multilevel models for 1999, Middle Response Style as outcome: estimates, standard errors and fit statistics

1999 MRS Models	**Model 1 (Baseline) ICC=0.09**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0023***	0.0006	0.0019***	0.0005	0.0012***	0.0004	0.0002**	0.0001
Residual variance	0.0242***	0.0002	0.0236***	0.0002	0.0237***	0.0002	0.0250***	0.0003
Pseudo-R^2 – *Country variance*			18.57%		48.14%		92.33%	
Pseudo-R^2 – *Individual variance*							-3.56%	
Number Countries	29		27		24		15	
Number Total Cases	33116		30217		26560		16427	
Predictors								
Intercept	0.1723***	0.0090	0.1783***	0.0098	0.2133***	0.0122	0.1664***	0.0455
Methodological factors – Country level								
Answer scale version			-0.0384*	0.0191	-0.0386*	0.0175	0.0268	0.0184
Mode of data collection					-0.0592***	0.0148	0.0636**	0.0293
Demographic variables – Individual level								
Age							-0.0005***	0.0001
Education (years)							0.0001	0.0004
Gender							0.0024	0.0025
Cultural factors – Country level								
Power distance							0.0004	0.0002
Individualism/collectivism							0.0020***	0.0006
Uncertainty avoidance							-0.0021***	0.0004
Masculinity/femininity							-0.0007***	0.0003
Fit statistics								
R^2							0.29	
ML deviance	-29155.7		-27310.1		-23916.5		-13918.7	
AIC	-29149.7		-27302.1		-23906.5		-13894.7	
BIC	-29145.6		-27296.9		-23900.6		-13886.2	

***p < .05; ** p < 0.01; ***p < .001.

Table 4.16.
Multilevel models for 2000, Middle Response Style as outcome: estimates, standard errors and fit statistics

2000 MRS Models	**Model 1 (Baseline) ICC=0.08**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0019***	0.0005	0.0015***	0.0004	0.0013***	0.0004	0.0008**	0.0003
Residual variance	0.0223***	0.0002	0.0218***	0.0002	0.0221***	0.0002	0.0221***	0.0002
Pseudo-R^2 – *Country variance*			17.96%		28.33%		55.43%	
Pseudo-R^2 – *Individual variance*							1.16%	
Number Countries	29		26		24		19	
Number Total Cases	29455		25531		23568		17769	
Predictors								
Intercept	0.1996***	0.0081	0.2008***	0.0088	0.2195***	0.0117	0.3145*	0.1055
Methodological factors – Country level								
Answer scale version			-0.0234	0.0184	-0.0230	0.0187	-0.0143	0.0254
Mode of data collection					-0.0348*	0.0152	0.0249	0.0330
Demographic variables – Individual level								
Age							-0.0008***	0.0001
Education (years)							-0.0028***	0.0003
Gender							0.0108***	0.0022
Cultural factors – Country level								
Power distance							-0.0003	0.0006
Individualism/collectivism							0.0001	0.0011
Uncertainty avoidance							-0.0017***	0.0006
Masculinity/femininity							-0.0004	0.0005
Fit statistics								
R^2							0.23	
ML deviance	-28282		-25074.5		-22882.4		-17283.2	
AIC	-28276		-25066.5		-22872.4		-17259.2	
BIC	-28271.9		-25061.4		-22866.6		-17247.9	

***p < .05; ** p < 0.01; ***p < .001.

Table 4.17.
Multilevel models for 2002, Middle Response Style as outcome: estimates, standard errors and fit statistics

2002 MRS Models	**Model 1 (Baseline) ICC=0.**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0019***	0.0005	0.0016***	0.0004	0.0014***	0.0003	0.0005***	0.0002
Residual variance	0.0154***	0.0001	0.0149***	0.0001	0.0152***	0.0001	0.0147***	0.0001
Pseudo-R^2 – *Country variance*			17.34%		30.35%		73.61%	
Pseudo-R^2 – *Individual variance*							4.86%	
Number Countries	38		33		31		21	
Number Total Cases	45917		42102		39668		27981	
Predictors								
Intercept	0.1629***	0.0072	0.1588***	0.0078	0.1798***	0.0112	0.1682***	0.0508
Methodological factors – Country level								
Answer scale version			-0.0191	0.0182	-0.0143	0.0181	-0.0301	0.0159
Mode of data collection					-0.0255*	0.0118	-0.0305	0.0189
Demographic variables – Individual level								
Age							0.0000	0.0000
Education (years)							-0.0005*	0.0002
Gender							-0.0101***	0.0015
Cultural factors – Country level								
Power distance							-0.0002	0.0003
Individualism/collectivism							0.0009	0.0005
Uncertainty avoidance							-0.0003	0.0003
Masculinity/femininity							-0.0003	0.0003
Fit statistics								
R^2							0.32	
ML deviance	-61041.7		-57381.2		-53266.4		-38627.7	
AIC	-61035.7		-57373.2		-53256.4		-38603.7	
BIC	-61030.8		-57367.2		-53249.2		-38591.1	

***p < .05; ** p < 0.01; ***p < .001. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.18.
Multilevel models for 2003, Middle Response Style as outcome: estimates, standard errors and fit statistics

2003 MRS Models	**Model 1 (Baseline) ICC=0.12**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0027***	0.0007	0.0026***	0.0007	0.0024***	0.0006	0.0013***	0.0004
Residual variance	0.0205***	0.0001	0.0200***	0.0001	0.0201***	0.0001	0.0201***	0.0002
Pseudo-R^2 – *Country variance*			3.48%		11.76%		51.76%	
Pseudo-R^2 – *Individual variance*							2.05%	
Number Countries	35		31		30		22	
Number Total Cases	44604		39918		38619		28528	
Predictors								
Intercept	0.2038***	0.0089	0.2073***	0.0103	0.2256***	0.0142	0.1772*	0.0800
Methodological factors – Country level								
Answer scale version			-0.0220	0.0234	-0.0143	0.0241	0.0215	0.0300
Mode of data collection					-0.0327	0.0182	0.0100	0.0300
Demographic variables – Individual level								
Age							-0.0007***	0.0001
Education (years)							-0.0011***	0.0002
Gender							0.0143***	0.0017
Cultural factors – Country level								
Power distance							0.0003	0.0005
Individualism/collectivism							0.0014	0.0008
Uncertainty avoidance							-0.0004	0.0005
Masculinity/femininity							-0.0014***	0.0004
Fit statistics								
R^2							0.27	
ML deviance	-46686.6		-42683.6		-41053.2		-30465	
AIC	-46680.6		-42675.6		-41043.2		-30441	
BIC	-46675.9		-42669.9		-41036.2		-30427.9	

***p < .05; ** p < 0.01; ***p < .001.

Table 4.19.
Multilevel models for 2004, Middle Response Style as outcome: estimates, standard errors and fit statistics

2004 MRS Models	**Model 1 (Baseline) ICC=0.08**		**Model 2 (Answer Scale Version predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0035***	0.0008	0.0036***	0.0008	0.0033***	0.0008	0.0023***	0.0006
Residual variance	0.0382***	0.0002	0.0374***	0.0003	0.0373***	0.0003	0.0372***	0.0003
Pseudo-R^2 – *Country variance*			-2.31%		5.70%		34.84%	
Pseudo-R^2 – *Individual variance*			2.17%				2.59%	
Number Countries	42		37		35		27	
Number Total Cases	48622		42690		40352		31096	
Predictors								
Intercept	0.2306***	0.0092	0.2329***	0.0112***	0.2590		0.3053***	0.0938
Methodological factors – Country level								
Answer scale verion			-0.0235	0.0241	-0.0122		-0.0094	0.0281
Mode of data collection					-0.0422		0.0116	0.0299
Demographic variables – Individual level								
Age							-0.0011***	0.0001
Education (years)							-0.0027***	0.0003
Gender							0.0197***	0.0022
Cultural factors – Country level								
Power distance							-0.0003	0.0006
Individualism/collectivism							0.0008	0.0008
Uncertainty avoidance							-0.0007	0.0007
Masculinity/femininity							-0.0017***	0.0006
Fit statistics								
R^2							0.23	
ML deviance	-20539.5		-18961.5		-18031.7		-13961.5	
AIC	-20533.5		-18953.5		-18021.7		-13937.5	
BIC	-20528.3		-18947.1		-18013.641		-13921.9	

***p < .05; ** p < 0.01; ***p < .001.

Acquiescence

Methodological factors. Mode of data collection was hypothesized to affect acquiescence measures. The results suggest that when considered alone, data collection method has an impact on acquiescence. Respondents in interviewer-administered settings are more likely to agree vs. disagree than respondents in self-administered settings (Figure 4.4). This finding is consistent with the deference hypothesis. However, when including the cultural dimensions in the model, this effect is attenuated and becomes only marginally significant. At same time, cultural variables become less significant after including the demographic variables in the model.

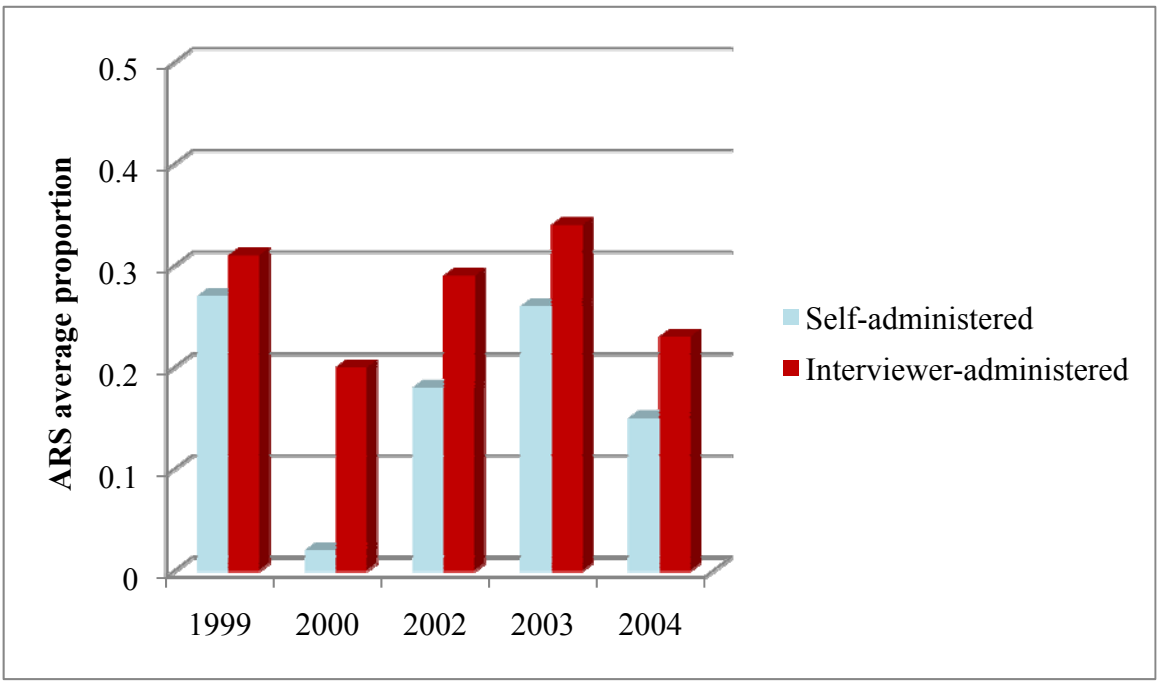

Figure 4.4. Average acquiescent response style proportion by mode of data collection

Demographic variables. The effect of age is consistent across all five years, and supports the cognitive ability hypothesis (Knaüper, 1999). Older respondents were more likely to acquiesce than younger respondents. The effect of education is also consistent with previous research, respondents with higher education tend to acquiesce less than respondents with lower education. The effect of education did not significantly vary across countries. Gender was significant for three out of five years; contrary to what a (small) majority of studies find, females were less likely to agree than males.

Cultural factors. An important finding concerned the relationship between individualism/ collectivism and acquiescence. Consistent with previous research and theoretical predictions, individualism/collectivism was significantly related to acquiescence for all the years when considered alone. Respondents in individualistic countries were less likely to acquiesce than respondents in collectivistic countries. In most years, the country variance accounted for by this predictor was larger than 50%. However, when introducing the demographic variables, this effect was considerable reduced, become nonsignificant for all years but 2002.

Table 4.20.
Multilevel models for 1999, Acquiescent Response Style as outcome: estimates, standard errors and fit statistics

1999 ARS Models	**Model 1 (Baseline) ICC=0.11**		**Model 2 (Data Collection Mode predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0115***	0.0033	0.0121***	0.0037	0.0132***	0.0042	0.0026**	0.0010
Residual variance	0.0898***	0.0007	0.0889***	0.007	0.0891***	0.0008	0.0887***	0.0010
Pseudo-R^2 – *Country variance*			-5.22%		-9.09%		80.30%	
Pseudo-R^2 – *Individual variance*							0.45%	
Number Countries	29		26		24		15	
Number Total Cases	33116		29459		26560		16427	
Predictors								
Intercept	0.2894***	0.0200	0.2702***	0.0333	0.2640***	0.0405	0.4000*	0.1673
Methodological factors – Country level								
Answer scale version					0.0201	0.0580	0.0858	0.0675
Mode of data collection			0.0427	0.0441	0.0452	0.0488	0.0579	0.1074
Demographic variables – Individual level								
Age							0.0009***	0.0002
Education (years)							-0.0125***	0.0007
Gender							0.0051	0.0047
Cultural factors – Country level								
Power distance							0.0003	0.0008
Individualism/collectivism							-0.0024	0.0022
Uncertainty avoidance							-0.0008	0.0014
Fit statistics								
R^2							0.29	
ML deviance	14318.6		12423.7		11254.2		6868.8	
AIC	14324.6		12431.7		11264.2		6892.8	
BIC	14328.7		12436.7		11270.0		6901.3	

***$p < .05$; ** $p < 0.01$; ***$p < .001$.

Table 4.21.
Multilevel models for 2000, Acquiescent Response Style as outcome: estimates, standard errors and fit statistics

2000 ARS Models	**Model 1 (Baseline) ICC=0.20**		**Model 2 (Data Collection Mode predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0188***	0.0050	0.0107***	0.0030	0.0085***	0.0025	0.0039**	0.0013
Residual variance	0.0754***	0.0006	0.0758***	0.0006	0.0778***	0.0007	0.0718***	0.0008
Pseudo-R^2 – *Country variance*			43.09%		20.56%		64.71%	
Pseudo-R^2 – *Individual variance*							7.71%	
Number Countries	29		26		24		19	
Number Total Cases	29455		27492		23568		17769	
Predictors								
Intercept	0.1193***	0.0256	0.0210	0.0289	0.0315	0.0293	-0.2299	0.2268
Methodological factors – Country level								
Answer scale version					-0.0591	0.0469	0.0172	0.0542
Mode of data collection			0.1855***	0.0401	0.2934***	0.0381	0.1455*	0.0705
Demographic variables – Individual level								
Age							0.0026***	0.0001
Education (years)							-0.0167***	0.0006
Gender							-0.0226***	0.0041
Cultural factors – Country level								
Power distance							0.0038**	0.0012
Individualism/collectivism							0.0006	0.0024
Uncertainty avoidance							0.0021	0.0013
Fit statistics								
R^2							0.52	
ML deviance	7616.4		7235.9		6803.6		3695.6	
AIC	7622.4		7243.9		6813.6		3719.6	
BIC	7626.5		7249.0		6819.5		3731.0	

***$p < .05$; ** $p < 0.01$; ***$p < .001$.

Table 4.22.
Multilevel models for 2002, Acquiescent Response Style as outcome: estimates, standard errors and fit statistics

2002 ARS Models	**Model 1 (Baseline) ICC=0.16**		**Model 2 (Data Collection Mode predictor only)**		**Model 3 (Methodological predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0081***	0.0019	0.0044***	0.0011	0.0038***	0.0010	0.0019**	0.0006
Residual variance	0.0440***	0.0003	0.0441***	0.0003	0.0440***	0.0003	0.0412***	0.0003
Pseudo-R^2 – *Country variance*			45.68%		13.64%		50.00%	
Pseudo-R^2 – *Individual variance*							6.36%	
Number Countries	38		35		31		21	
Number Total Cases	45917		43407		39668		27981	
Predictors								
Intercept	0.2605***	0.0147	0.1799***	0.0181	0.1768***	0.0188	0.3466**	0.0975
Methodological factors – Country level								
Answer scale version					0.0488	0.0304	0.0722*	0.0306
Mode of data collection			0.1059***	0.0199	0.1090***	0.0197	0.0174	0.0363
Demographic variables – Individual level								
Age							0.0024***	0.0001
Education (years)							-0.0080***	0.0003
Gender							-0.0072**	0.0025
Cultural factors – Country level								
Power distance							-0.0002	0.0006
Individualism/collectivism							-0.0033**	0.0010
Uncertainty avoidance							0.0008	0.0006
Fit statistics								
R^2							0.45	
ML deviance	-12874.2		-12187.8		-11164.6		-9785.3	
AIC	-12868.2		-12179.8		-11154.6		-9761.3	
BIC	-12863.3		-12173.6		-11147.4		-9748.7	

***$p < .05$; ** $p < 0.01$; ***$p < .001$. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

Table 4.23.
Multilevel models for 2003, Acquiescent Response Style as outcome: estimates, standard errors and fit statistics

2003 ARS Models	**Model 1 (Baseline) ICC=0.14**		**Model 2 (Data Collection Mode predictor only)**		**Model 3 (Cultural predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0084***	0.0020	0.0071***	0.0018	0.0064***	0.0017	0.0023***	0.0007
Residual variance	0.0535***	0.0004	0.0534***	0.0003	0.0519***	0.0004	0.0467***	0.0004
Pseudo-R^2 – *Country variance*			15.48%		23.81%		72.62%	
Pseudo-R^2 – *Individual variance*							12.71%	
Number Countries	35		33		30		22	
Number Total Cases	44604		41999		38619		28528	
Predictors								
Intercept	0.3127***	0.0155	0.2627***	0.0234	0.2521***	0.0231	0.2154	0.1066
Methodological factors – Country level								
Answer scale version					0.0686	0.0395	-0.0682	0.0340
Mode of data collection			0.0841**	0.0301	0.0744*	0.0297	0.0223	0.0399
Demographic variables – Individual level								
Age							0.0021***	0.0001
Education (years)							-0.0128***	0.0026
Gender							-0.0122***	0.0026
Cultural factors – Country level								
Power distance							0.0003	0.0007
Individualism/collectivism							-0.0011	0.0010
Uncertainty avoidance							0.0015*	0.0007
Fit statistics								
R^2							0.40	
ML deviance	-3845.7		-3662.3		-4507.2		-6227.4	
AIC	-3839.7		-3654.3		-4497.2		-6203.4	
BIC	-3835.1		-3648.3		-4490.2		-6190.3	

***$p < .05$; ** $p < 0.01$; ***$p < .001$.

Table 4.24.
Multilevel models for 2004, Acquiescent Response Style as outcome: estimates, standard errors and fit statistics

2004 ARS Models	**Model 1 (Baseline) ICC=0.08**		**Model 2 (Data Collection Mode predictor only)**		**Model 3 (Cultural predictors only)**		**Model 4 (All hypothesized predictors**	
	Estimate	SE	Estimate	SE	Estimates	SE	Estimates	SE
Total country level variance	0.0079***	0.0017	0.0062***	0.0014	0.0062***	0.0015	0.0041***	0.0011
Residual variance	0.0953***	0.0006	0.0947***	0.0006	0.0955***	0.0007	0.0955***	0.0001
Pseudo-R^2 – *Country variance*			21.52%		21.52%		33.87%	
Pseudo-R^2 – *Individual variance*							0.00%	
Number Countries	42		38		35		27	
Number Total Cases	48622		45282		40352		31096	
Predictors								
Intercept	0.2127***	0.0139	0.1482***	0.0239	0.1568***	0.0252	0.2068	0.1256
Methodological factors – Country level								
Answer scale version					0.0105	0.0340	-0.0042	0.0377
Mode of data collection			0.0843**	0.0285	0.0678*	0.0301	0.0695	0.0401
Demographic variables – Individual level								
Age							0.0014***	0.0001
Education (years)							-0.0141***	0.0005
Gender							-0.0068	0.0035
Cultural factors – Country level								
Power distance							-0.0013	0.0008
Individualism/collectivism							0.0001	0.0011
Uncertainty avoidance							-0.0001	0.0009
Fit statistics								
R^2							0.22	
ML deviance	23892.9		21947.8		19865.2		15327.3	
AIC	23898.9		21955.8		19875.2		15351.3	
BIC	23904.2		21962.3		19883.0		15.366.8	

***$p < .05$; ** $p < 0.01$; ***$p < .001$.

Chapter 5. Effects of Answer Scale Modifications on Response Distributions

This chapter focuses on the impact of different translations on the probability distributions of response categories in agreement answer scales. The objective was to compare the psychometric properties of attitudinal scales across countries that used different answer scale versions (see table 5.1 for verbal labels of the five answer scale points of the countries examined in this chapter): Great Britain, as case study using the source text; Germany and Spain, as case studies of countries using a closely translated answer scale; Denmark, as case study where the answer scale introduced an intensity modifier, and Japan, as a case study where the answer scale suffered a different type of modification. Germany, Japan and Great Britain were selected because the results of this study can be compared against those findings regarding intensity scores obtained in the calibration study by Smith et al. (2009). Given that the English calibration of this study was conducted on a U.S. sample, it would have been ideal to test these models using the ISSP U.S. respondents. However, the year for which these analyses were possible was the year in which the agreement scale of the United States had four scale points instead of five. Therefore, Great Britain was selected instead. Denmark was selected among other countries that used the "modified intensity" answer scale version because it was the most similar in other characteristics to Great Britain and Germany.

Table 5.1.

Agreement answer scales for Great Britain, Germany, Japan, Denmark, and Spain in 2002

Great Britain	**Germany**	**Japan**	**Denmark**	**Spain**
Strongly Agree	Stimme voll und ganz zu	Sou omou	Helt enig	Muy de acuerdo
Agree	Stimme zu	Dochiraka to ieba sou omou	Delvis enig	De acuerdo
Neither Agree Nor Disagree	Weder noch	Dochira tomo ienai	Hverken enig eller uenig	Ni de acuerdo ni en desacuerdo
Disagree	Stimme nicht zu	Dochiraka to ieba sou omwanai	Delvis uenig	En desacuerdo
Strongly Disagree	Stimme überhaupt nicht zu	Sou omowanai	Helt uenig	Muy en desacuerdo
Can’t Choose	Kann ich nicht sagen	- - - -	Ved ikke	NS | NC

Motivation for the Analysis

Table 5.2 and 5.3 reproduce calibration data found by Smith et al. (2009). Table 5.2 shows the intensity means associated with each of the response categories that are commonly used in ISSP surveys for German (obtained from German participants), English (obtained from U.S. participants) and Japanese (obtained from Japanese participants). Table 5.3 shows the labels for each country that define similar intensity intervals between scale points and across countries. If using the calibration technique for multilingual answer scale design, verbal labels with these properties would probably be chosen.

Table 5.2.
Means of response categories commonly used in ISSP surveys

Item IDs D/US	German Expressions	American Expressions	Japanese Expressions	Mean Germany	Mean USA	Mean Japan
A20/v	Stimme voll und ganz zu	Strongly agree	Hijouni sou omou	19,87	18,80	17,70
A16/b	Stimme zu	Agree	Sou Omou	19,05	16,00	15,70
A4/p	Stimme weder zu noch lehne ab	Neither agree nor disagree	Dochira tomo ieanai	9,77	9,90	10,00
A3/j	Lehne ab	Disagree	Sou omowanai	2,41	3,50	3,70
A5/w	Lehne stark ab	Strongly disagree	Kesshite sou wa omowanai	1,21	1,50	2,70
A9/e	Kann ich nicht sagen	Can't choose	Wakaranai	9,42	9,80	8,30

Smith et al. 2009

Table 5.3.

Means of response categories commonly used in ISSP surveys

Item IDs D/US	German Expressions	American Expressions	Japanese Expressions	Mean Germany	Mean USA	Mean Japan
A17/h	Stimme bestimmt zu	Definitely agree	Zettai	19,22	19,00	18,80
A1/a	Stimme im Grunde zu	Basically Agree	Kihonteki ni wa	14.93	13.80	14.70
A4/p	Stimme weder zu noch lehne ab	Neither agree nor disagree	Dochira tomo ieanai	9.77	9.90	10,00
A21/o	Lehne maessig ab	Moderately Disagree	Amari	6.63	6.40	6.00
A2/i	Lehne stark ab	Definitely disagree	Zettai/Zenzen/ Mattaku	1.21	1.00	2.10

Smith et al. 2009

It is worth noting that the mean intensity rating of the Japanese label for the endpoints is lower than for the other two countries. An intensity reduction seems to have been what the Japanese researchers intended when they chose the answer scale verbal labels, in order to compensate for the Japanese avoidance of the endpoints. Smith et al. (2009) hypothesize that the use of these ratings "should facilitate creation of an optimal response scale", but caution against assuming that they will perform better and call for empirical evidence of how these verbal labels perform when used in surveys is necessary. Ratings were done one verbal label at a time, rather than in relation to other scale points, as it would be in any given survey. The goal here is to provide a test of how the wording of verbal labels in table 5.1 performed when used in the British, German, and Japanese ISSP 2002 module.

Method

Eight items from the "Family and Changing Gender Roles" module related to the role of women in the labor force were selected (see table 5.4). I focused on the samples from the United Kingdom (*n* = 1960), Spain (*n* = 2471), Japan (*n* = 1132), Germany (*n* = 1367) and Denmark (*n* = 1379).

Table 5.4.
Support of women in the labor force scale

Item 4: A working mother can establish just as warm and secure a relationship with her children as a mother who does not work
Item 5: A pre-school child is likely to suffer if his or her mother works
Item 6: All in all, family life suffers when the woman has a full-time job
Item 7: A job is all right, but what most women really want is a home and children
Item 8: Being a housewife is as fulfilling as working for pay
Item 9: Having a job is the best way for a woman to be an independent person.
Item 10: Both the man and the woman should contribute to the household income.
Item 11: A man's job is to earn money; a woman's job is to look after the home and family

Measurement Models

In multilingual studies and in translation and adaptation of existing instruments for use in different languages and cultures, a traditional approach to examining comparability of scales in each group has been to compare psychometric indices from classical test theory. If obtained indices such as Cronbach's alpha or factorial structures were similar for each language, this was taken as evidence that the scales worked comparably in all the groups with similar indices. Even though newer approaches to estimation of measurement error have been proposed and are widely used in some contexts, classical test theory is still presented as preliminary evidence of comparability. In order to evaluate whether this information can indeed be valuable, I first examined

psychometric indices from classical test theory. Then, I used a family of models that imposes fewer assumptions about the distribution of the data and the relationship between the items and the construct of interest, namely, Item Response Theory (IRT). These models allow assessing the dimensionality of the scale and examining which items are better suited to measure the latent construct, as well as for what kind of people (in terms of their value on the latent variable).

After examining these two approaches separately, Cronbach's alpha values were compared to reliability indices from IRT. In the IRT context, conjoint scaling leads to an interesting development of the concept of reliability. Under this framework, one does not only care about obtaining a high level of reliability for a given scale, but wants that reliability to cover the target population. Items are not (necessarily) equally informative across the full range of the latent trait, as was assumed in classical test theory. Instead, the items (and therefore perhaps the scale) are more reliable for a particular range of the population. An ideal scale will include items that maximally distinguish across individuals that are the target population of a study. In survey research, this is very often the general population, and therefore scales are intended to cover as much of the latent trait as possible.

Polytomous IRT models of the eight items were estimated for each country using Mplus 5.1. These models, as compared to models where items are assumed to be continuous, allow to more carefully assess the impact of different categories on measurement. An assumption is made that these categories are ordered with respect to each other, where agreeing with one item reflects less of the attitude than strongly agreeing with it. Given the graded response nature of the answer scale, Samejima's

(1969) model with a logistic response function was the preferred model. This model provides information about the probability distribution of each response category across the latent construct of interest. Because of the expectation that answer scale labels would affect how respondents use each of the response categories, this model is particularly relevant. Two-parameter and one-parameter IRT models were compared for each country.

Results

Item statistics and classical test theory measurement indicators

Tables 5.5 and 5.6 present item means and standard deviations for all eight items in each country, as well as information about the relationship of each item with the rest of the scale and Cronbach's alpha for each country. For most items, means were higher in the Denmark sample. The range of means for Great Britain and Spain suggests that items were neither too easy nor too difficult for the sample (from about 2.8 to about 3.9), except for item 10 in Spain, that seems somewhat easier than the rest. In Denmark, four items have means above 4, suggesting that the scale was "easier" for them than for the other countries. That is, the Danish sample shows an overall higher level of support for women in the labor force. The mean of item 8 was particularly low for Japan (2.03).

Table 5.5.

Item-total statistics – Great Britain, Spain and Denmark

	UK. Cronbach's α = 0.73					Spain. Cronbach's α = 0.74					Denmark. Cronbach's α = 0.73				
	M	SD	% Missing	Corrected Item-Total Correlation	α if Item Deleted	M	SD	% Missing	Corrected Item-Total Correlation	α if Item Deleted	M	SD	% Missing	Corrected Item-Total Correlation	α if Item Deleted
Item 4	3.57	1.17	3	0.49	0.69	3.45	1.20	1.9	0.42	0.71	4.09	1.16	4.1	0.43	0.70
Item 5	3.09	1.12	3.3	0.59	0.67	2.86	1.10	2.8	0.48	0.70	3.50	1.40	5.8	0.57	0.66
Item 6	3.10	1.15	3.3	0.58	0.67	2.76	1.10	2.5	0.49	0.70	3.67	1.46	5.9	0.55	0.67
Item 7	3.34	1.07	5.1	0.45	0.70	3.10	1.13	5.0	0.50	0.69	3.54	1.36	9.1	0.50	0.68
Item 8	2.81	1.07	4.6	0.30	0.73	3.09	1.18	4.6	0.31	0.73	3.07	1.44	11.6	0.34	0.72
Item 9	3.40	1.00	3.8	0.20	0.75	3.91	0.93	2.5	0.28	0.73	4.17	1.20	6.2	0.09	0.76
Item 10	3.64	0.98	3.2	0.28	0.73	4.09	0.74	1.6	0.39	0.72	4.16	1.08	2.5	0.29	0.72
Item 11	3.61	1.07	2.2	0.53	0.69	3.62	1.19	1.5	0.58	0.68	4.25	1.21	2.7	0.59	0.67

Table 5.6.

Item-total statistics – Germany and Japan

	Germany Cronbach's α = 0.79					Japan Cronbach's α = 0.62				
	M	SD	% Missing	Corrected Item-Total Correlation	α if Item Deleted	M	SD	% Missing	Corrected Item-Total Correlation	α if Item Deleted
Item 4	4.22	1.04	3%	.41	.78	4.12	1.26	3%	.26	.36
Item 5	2.89	1.27	6%	.59	.75	3.36	1.45	5%	.32	.32
Item 6	3.07	1.31	5%	.64	.74	3.31	1.44	3%	.42	.27
Item 7	3.65	1.22	8%	.58	.75	2.78	1.50	7%	.24	.36
Item 8	3.24	1.35	9%	.50	.76	2.03	1.18	8%	-.25	.54
Item 9	4.01	.98	4%	.24	.80	3.44	1.43	5%	.04	.45
Item 10	3.89	1.03	4%	.33	.79	3.37	1.43	2%	.09	.43
Item 11	3.66	1.20	4%	.61	.74	3.40	1.48	1%	.34	.31

In all countries except for Japan there were five items that showed item-remainder correlations above 0.4, and three that seemed to be more weakly related to the rest of the scale: items 9, 10 and 8. Item 9, particularly, was weakly correlated with the rest of the scale in Denmark (0.09) and Japan (0.04). This finding was not too surprising when the

content of the item is examined ("Having a job is the best way for a woman to be an independent person"). Agreeing with such an item implies that one believes women need to "gain" independency. This may strongly contrast with views that support the presence of women in the labor force, thus making people from both ends of this latent trait be likely to agree or disagree with the item. In addition, deletion of item 9 would increase the reliability of the scale in Denmark, Germany and Great Britain. Therefore, it was considered appropriate to eliminate the item from further analyses.

Even more problematic is item 8 in Japan. A negative correlation suggests that the item is interpreted in the opposite direction that it was intended, so that people that agree that being a housewife is as fulfilling as working for pay tend to be in favor of the integration of women in the labor force. This item certainly deserves closer attention, so as to determine what is driving the unexpected direction. However, this item was dropped in later analyses because of poor functioning across countries; therefore, this discussion is out of scope for this dissertation.

Tables 5.7 through 5.11 present the Pearson correlations among all items for Great Britain, Germany, Japan, Denmark, and Spain. Given their low correlations with the rest, items 8 and 10 may be measuring a different dimension than the other five. However, their deletion would not improve Cronbach's α for any of the countries, and this could indicate that the items are measuring different facets of the same dimension. The explanation may also be related to the one for item 9. A point of view that women should choose the path that they prefer may make people that support women in the labor force agree with the item "Being a housewife is as fulfilling as working for pay". These two items (8 and 9, bold font in tables 5.7 through 5.11) may be measuring a different

dimension such as "women's freedom of choice", rather than support for participation in the labor force, therefore neither of them will be used in further analyses. However, item 10 seems theoretically consistent with the support of women in labor force topic, and has higher correlations with items other than items 8 and 9. For that reason, item 10 will be examined using an IRT model, to see whether it performs adequately.

Table 5.7.

Inter-item correlation matrix – Great Britain

	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11
Item 4	1.00							
Item 5	.51	1.00						
Item 6	.47	.62	1.00					
Item 7	.24	.37	.36	1.00				
Item 8	**.07**	**.13**	**.17**	**.32**	**1.00**			
Item 9	**.16**	**.08**	**.09**	**.00**	**.21**	**1.00**		
Item 10	.23	.18	.18	.03	.12	.33	1.00	
Item 11	.32	.43	.45	.52	.27	.03	.14	1.00

Table 5.8.

Inter-item correlation matrix – Germany

	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11
Item 4	1.00							
Item 5	.37	1.00						
Item 6	.37	.64	1.00					
Item 7	.22	.44	.49	1.00				
Item 8	**.13**	**.34**	**.40**	**.52**	**1.00**			
Item 9	**.22**	**.12**	**.11**	**.13**	**.15**	**1.00**		
Item 10	.33	.22	.28	.10	.21	.28	1.00	
Item 11	.33	.42	.50	.58	.49	.14	.19	1.00

Table 5.9.

Inter-item correlation matrix – Japan

	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11
Item 4	1.00							
Item 5	.24	1.00						
Item 6	.33	.46	1.00					
Item 7	.09	.19	.27	1.00				
Item 8	**-.26**	**-.14**	**-.17**	**-.08**	**1.00**			
Item 9	**.09**	**-.04**	**-.02**	**-.08**	**-.09**	**1.00**		
Item 10	.10	-.01	.54	-.11	-.04	.38	1.00	
Item 11	.24	.26	.27	.40	-.10	-.55	-.04	1.00

Table 5.10.

Inter-item correlation matrix – Denmark

	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11
Item 4	1.00							
Item 5	.43	1.00						
Item 6	.44	.63	1.00					
Item 7	.22	.40	.39	1.00				
Item 8	**.06**	**.21**	**.17**	**.36**	**1.00**			
Item 9	**.06**	**-.03**	**-.03**	**.02**	**.13**	**1.00**		
Item 10	.26	.13	.15	.10	.18	.25	1.00	
Item 11	.31	.47	.46	.49	.34	.05	.20	1.00

Table 5.11

Inter-item correlation matrix – Spain

	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11
Item 4	1.00							
Item 5	.38	1.00						
Item 6	.39	.52	1.00					
Item 7	.20	.29	.36	1.00				
Item 8	**.06**	**.13**	**.12**	**.33**	**1.00**			
Item 9	**.16**	**.12**	**.10**	**.17**	**.19**	**1.00**		
Item 10	.31	.19	.18	.19	.17	.32	1.00	
Item 11	.31	.34	.34	.48	.35	.20	.33	1.00

Psychometric Analysis – Item Response Theory

The starting point was the estimation of a model that contained six items for each country. For each model, six *a* parameters or discriminations were estimated (one per item) and twenty-four *b* parameters or difficulties (four per item) were computed. In order to identify the model, the latent trait metric was obtained by fixing the mean to zero and the variance to one. These parameters are reported on tables 5.12 through 5.16.

Conceptually, the *b* parameter of a response graded model can be thought of as the point on the latent construct (e.g., the attitude supporting women's participation in the labor force) where the probability of choosing a given response category is equal to 0.50. This indicates what "range" of the attitude a given item is tapping into. Because in graded response models the *b* parameter is estimated for each response option of each item, it can be interpreted as the likely level of support for the participation of women in the labor force of a respondent who selected that response option in that item. For example, for Great Britain (see table 5.12), answering "strongly agree" to item 10 is most likely to happen among respondents with a high level of support for women in the labor force (b = 2.99, the highest value of all items and response options). Therefore, the *b* parameter gives us a sense of what range of the attitude a certain response category is measuring best.

The *a* parameter is the slope at the point of the latent trait scale where the response category best discriminates between those of low or high levels on the support attitude. A high *a* parameter value indicates that the response cateogory of the item can sharply differentiate between people with an attitude more positive than the latent trait score at that point and people with an attitude more negative than that point. Let us

assume that a very steep slope (thus a high *a* parameter value) is obtained for response option "agree" of item 5 exactly at the middle point of the latent trait. This would indicate that this "agree" response category of item 5 can differentiate very accurately between respondents with a slightly positive attitude and respondents with a slightly negative attitude. This renders the response category useful to measure the latent trait, in combination with the specific item. A low parameter *a* is less desirable; a response category with a flat slope would be as likely to be chosen by someone with a very high attitude level as by someone with a very low attitude level. Therefore, the response option would not allow the researcher to capture the variability across individuals regarding his or her variable of interest.

Constraints can be set regarding how different the values of the parameters are allowed to be across items. In one-parameter models, the slopes of response categories are assumed to be equally "informative" of the underlying attitude in all items. This means that all response categories are taken to discriminate equally well in all items. This is a testable hypothesis. Forcing the parameters to be equal across items made the models significantly for all the countries considered. Therefore, all countries needed models where items have both different thresholds and different discriminations (see table 5.17).

Table 5.12.

Model parameters – Great Britain

	Unstandardized estimates					
Item	Loading	SE Loading	Threshold	SE Threshold	pred a	pred b
item4$1	1.59	0.08	-3.93	0.14	0.94	-2.47
item4$2	1.59	0.08	-1.59	0.08	0.94	-1.00
item4$3	1.59	0.08	-0.87	0.07	0.94	-0.55
item4$4	1.59	0.08	1.84	0.08	0.94	1.16
item5$1	2.76	0.14	-4.97	0.21	1.62	-1.81
item5$2	2.76	0.14	-1.00	0.10	1.62	-0.36
item5$3	2.76	0.14	0.54	0.09	1.62	0.19
item5$4	2.76	0.14	4.44	0.20	1.62	1.61
item6$1	2.46	0.12	-4.33	0.17	1.45	-1.76
item6$2	2.46	0.12	-1.09	0.09	1.45	-0.44
item6$3	2.46	0.12	0.42	0.09	1.45	0.17
item6$4	2.46	0.12	3.84	0.16	1.45	1.56
item7$1	1.21	0.07	-3.83	0.14	0.71	-3.16
item7$2	1.21	0.07	-1.36	0.07	0.71	-1.12
item7$3	1.21	0.07	0.07	0.06	0.71	0.06
item7$4	1.21	0.07	2.29	0.09	0.71	1.89
item10$1	**0.47**	**0.05**	**-4.42**	**0.20**	**0.28**	**-9.32**
item10$2	0.47	0.05	-1.90	0.07	0.28	-4.02
item10$3	0.47	0.05	-0.46	0.05	0.28	-0.98
item10$4	0.47	0.05	1.42	0.06	0.28	2.99
item11$1	1.53	0.08	-3.98	0.14	0.90	-2.60
item11$2	1.53	0.08	-1.93	0.08	0.90	-1.26
item11$3	1.53	0.08	-0.68	0.07	0.90	-0.44
item11$4	1.53	0.08	2.04	0.09	0.90	1.33

Table 5.13

Model parameters – Germany

	Unstandardized estimates					
Item	Loading	SE Loading	Threshold	SE Threshold	pred a	pred b
item4$1	1.15	0.09	-4.27	0.20	0.67	-3.73
item4$2	1.15	0.09	-2.39	0.11	0.67	-2.08
item4$3	1.15	0.09	-2.03	0.10	0.67	-1.77
item4$4	1.15	0.09	-0.03	0.07	0.67	-0.02
item5$1	2.35	0.14	-3.31	0.16	1.38	-1.41
item5$2	2.35	0.14	-0.19	0.10	1.38	-0.08
item5$3	2.35	0.14	0.77	0.10	1.38	0.33
item5$4	2.35	0.14	3.29	0.17	1.38	1.40
item6$1	2.97	0.19	-3.99	0.22	1.75	-1.34
item6$2	2.97	0.19	-0.82	0.12	1.75	-0.28
item6$3	2.97	0.19	0.35	0.12	1.75	0.12
item6$4	2.97	0.19	3.37	0.20	1.75	1.14
item7$1	1.68	0.10	-3.65	0.16	0.99	-2.17
item7$2	1.68	0.10	-1.82	0.10	0.99	-1.08
item7$3	1.68	0.10	-0.98	0.09	0.99	-0.58
item7$4	1.68	0.10	1.37	0.09	0.99	0.81
item10$1	**0.70**	**0.07**	**-4.30**	**0.22**	**0.41**	**-6.15**
item10$2	0.70	0.07	-1.93	0.09	0.41	-2.76
item10$3	0.70	0.07	-1.14	0.07	0.41	-1.62
item10$4	0.70	0.07	0.91	0.07	0.41	1.30
item11$1	1.70	0.10	-3.69	0.16	1.00	-2.17
item11$2	1.70	0.10	-1.94	0.11	1.00	-1.14
item11$3	1.70	0.10	-0.92	0.09	1.00	-0.54
item11$4	1.70	0.10	1.42	0.09	1.00	0.83

Table 5.14

Model parameters – Japan

	Unstandardized estimates					
Item	Loading	SE Loading	Threshold	SE Threshold	pred a	pred b
item4$1	1.00	0.10	-2.86	0.14	0.59	-2.85
item4$2	1.00	0.10	-2.20	0.11	0.59	-2.20
item4$3	1.00	0.10	-1.27	0.09	0.59	-1.27
item4$4	1.00	0.10	-0.31	0.07	0.59	-0.31
item5$1	1.55	0.13	-2.53	0.14	0.91	-1.63
item5$2	1.55	0.13	-1.12	0.10	0.91	-0.72
item5$3	1.55	0.13	0.23	0.09	0.91	0.15
item5$4	1.55	0.13	0.88	0.09	0.91	0.57
item6$1	2.15	0.21	-3.09	0.21	1.26	-1.44
item6$2	2.15	0.21	-1.14	0.12	1.26	-0.53
item6$3	2.15	0.21	0.38	0.10	1.26	0.18
item6$4	2.15	0.21	1.16	0.12	1.26	0.54
item7$1	0.87	0.09	-1.02	0.08	0.51	-1.17
item7$2	0.87	0.09	-0.25	0.07	0.51	-0.29
item7$3	0.87	0.09	0.87	0.08	0.51	1.01
item7$4	0.87	0.09	1.46	0.09	0.51	1.69
item10$1	**0.04**	**0.07**	**-1.58**	**0.08**	**0.02**	**-39.45**
item10$2	**0.04**	**0.07**	**-1.08**	**0.07**	**0.02**	**-26.88**
item10$3	0.04	0.07	0.04	0.06	0.02	1.08
item10$4	**0.04**	**0.07**	**0.83**	**0.07**	**0.02**	**20.68**
item11$1	1.06	0.10	-2.09	0.11	0.62	-1.97
item11$2	1.06	0.10	-0.96	0.08	0.62	-0.91
item11$3	1.06	0.10	0.09	0.07	0.62	0.08
item11$4	1.06	0.10	0.66	0.08	0.62	0.63

Table 5.15

Model parameters – Denmark

	Unstandardized estimates					
Item	Loading	SE Loading	Threshold	SE Threshold	pred a	pred b
item4$1	1.50	0.07	-3.99	0.13	0.88	-2.66
item4$2	1.50	0.07	-0.99	0.06	0.88	-0.66
item4$3	1.50	0.07	-0.73	0.06	0.88	-0.48
item4$4	1.50	0.07	2.08	0.08	0.88	1.39
item5$1	1.99	0.10	-4.13	0.15	1.17	-2.07
item5$2	1.99	0.10	0.11	0.07	1.17	0.06
item5$3	1.99	0.10	0.83	0.07	1.17	0.42
item5$4	1.99	0.10	4.27	0.16	1.17	2.14
item6$1	2.04	0.10	-3.79	0.14	1.20	-1.86
item6$2	2.04	0.10	0.28	0.07	1.20	0.14
item6$3	2.04	0.10	1.06	0.07	1.20	0.52
item6$4	2.04	0.10	4.69	0.18	1.20	2.30
item7$1	1.40	0.07	-3.86	0.13	0.82	-2.76
item7$2	1.40	0.07	-0.46	0.06	0.82	-0.33
item7$3	1.40	0.07	0.30	0.06	0.82	0.21
item7$4	1.40	0.07	2.87	0.10	0.82	2.05
item10$1	**0.93**	**0.06**	**-5.80**	**0.31**	**0.55**	**-6.23**
item10$2	0.93	0.06	-3.25	0.10	0.55	-3.49
item10$3	0.93	0.06	-2.23	0.07	0.55	-2.40
item10$4	0.93	0.06	1.25	0.06	0.55	1.35
item11$1	1.55	0.08	-3.95	0.13	0.91	-2.55
item11$2	1.55	0.08	-1.52	0.07	0.91	-0.98
item11$3	1.55	0.08	-0.92	0.06	0.91	-0.60
item11$4	1.55	0.08	1.62	0.07	0.91	1.05

Table 5.16

Model parameters – Spain

	Unstandardized estimates					
Item	Loading	SE Loading	Threshold	SE Threshold	pred a	pred b
item4$1	1.37	0.09	-3.57	0.15	0.80	-2.62
item4$2	1.37	0.09	-2.20	0.10	0.80	-1.61
item4$3	1.37	0.09	-1.84	0.10	0.80	-1.35
item4$4	1.37	0.09	0.04	0.07	0.80	0.03
item5$1	2.76	0.18	-4.36	0.23	1.62	-1.58
item5$2	2.76	0.18	-1.54	0.13	1.62	-0.56
item5$3	2.76	0.18	-0.56	0.12	1.62	-0.20
item5$4	2.76	0.18	1.10	0.12	1.62	0.40
item6$1	2.68	0.18	-3.95	0.21	1.57	-1.48
item6$2	2.68	0.18	-1.80	0.14	1.57	-0.67
item6$3	2.68	0.18	-1.05	0.12	1.57	-0.39
item6$4	2.68	0.18	0.38	0.11	1.57	0.14
item7$1	1.24	0.08	-2.77	0.12	0.73	-2.23
item7$2	1.24	0.08	-1.27	0.08	0.73	-1.02
item7$3	1.24	0.08	-0.26	0.07	0.73	-0.21
item7$4	1.24	0.08	0.78	0.08	0.73	0.63
item10$1	**0.45**	**0.06**	**-3.48**	**0.16**	**0.26**	**-7.78**
item10$2	0.45	0.06	-2.29	0.09	0.26	-5.11
item10$3	0.45	0.06	-1.28	0.07	0.26	-2.85
item10$4	0.45	0.06	-0.10	0.06	0.26	-0.23
item11$1	1.92	0.13	-4.43	0.21	1.13	-2.31
item11$2	1.92	0.13	-2.82	0.15	1.13	-1.47
item11$3	1.92	0.13	-1.87	0.12	1.13	-0.97
item11$4	1.92	0.13	-1.00	0.10	1.13	-0.52

Table 5.17
Model comparison: two parameter vs. one parameter models

	Great Britain	Germany	Japan	Denmark	Spain
Chi-Square test for difference testing	989.67	363.01	317.12	762.55	517.02
Degrees of freedom	4	4	4	4	4
p-value	0.00	0.00	0.00	0.00	0.00

Global model fit as indicated by chi-square was not interpretable for any of the countries. This is often the case because the value is based on the comparison of the original response pattern and how the model reproduces it; this leads to a large number of cells, where many have low frequencies. This makes the approximation not valid and other forms of model fit assessment become preferable. One way to solve this problem is to look at the bivariate standardized residuals, so as to identify the source of the lack of adequate fit in the measurement model. The bivariate standardized residuals were significant in most cases, even when the difference between the frequency estimated by the model is considerably close to the observed in the data. No particular items were identified as causing the misfit. This suggests lack of unidimensionality of the construct as measured by this scale, that is, that the set of items is measuring more than just support of women in the labor force (assuming that the scale indeed measures that construct). In general one can say that model fit is less than optimal, and that, based on model fit statistics, there seem to be problems with the way the items perform together.

Another way of studying model fit is to inspect visual information of the scale as a whole and of the individual items. Figures 5.1 through 5.5 show the histograms that represent both people and item categories on the latent trait for each country. The vertical

axis in the figures represent the attitudinal continuum, that ranges between two arbitrary values. The bars in the histograms represent frequencies: the bars on the left side represent how many item difficulties (or *b* parameters) belong in that range of the attitudinal continuum, and the bars on the right side represent how many respondents in the sample are placed in each given interval. An instrument adequately covers the attitudinal range of the sampled population if the bars in both sides "overlap", that is, if they show similar frequencies at the same intervals of the attitudinal continuum.

Imagine that most participants showed very low support for women in the labor force, with tall bars were observed in the upper part of the figure, but that most item difficulties were very high, with taller bars in the lower part of the histogram. Most respondents would consistently disagree with items that were strongly pro-women in labor force, and survey time would be wasted on items that cannot differentiate among the respondents in the sample. The instrument thus would not be well designed to measure that particular population.

As can be seen from figures 5.1. through 5.5., the instrument does not perfectly cover the range of ability that the population exhibits in any of the countries. Individuals tend to have a higher level of attitude than what the items are meant to measure.

Items have generally low difficulties (as can be seen either in tables 5.12 to 5.16 or in figures 5.1. to 5.5). Item 10, however, seems to be a really "easy" item (that even people that support women having jobs to a very little extent would endorse), because for all five countries the item difficulty lies in a very low range, clearly differentiated from the other items. However, when the sample (the "participants" histogram) does not cover

that particular range of the attitude, the finding must be interpreted with care, because the standard errors for the value will be very large.

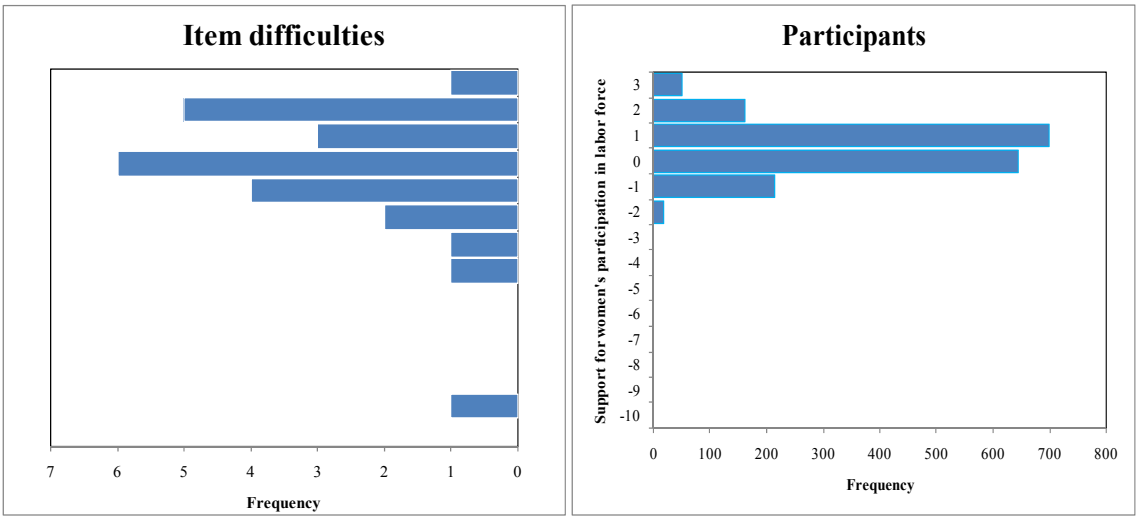

Figure 5.1. Histogram of people and item difficulty distribution – Great Britain

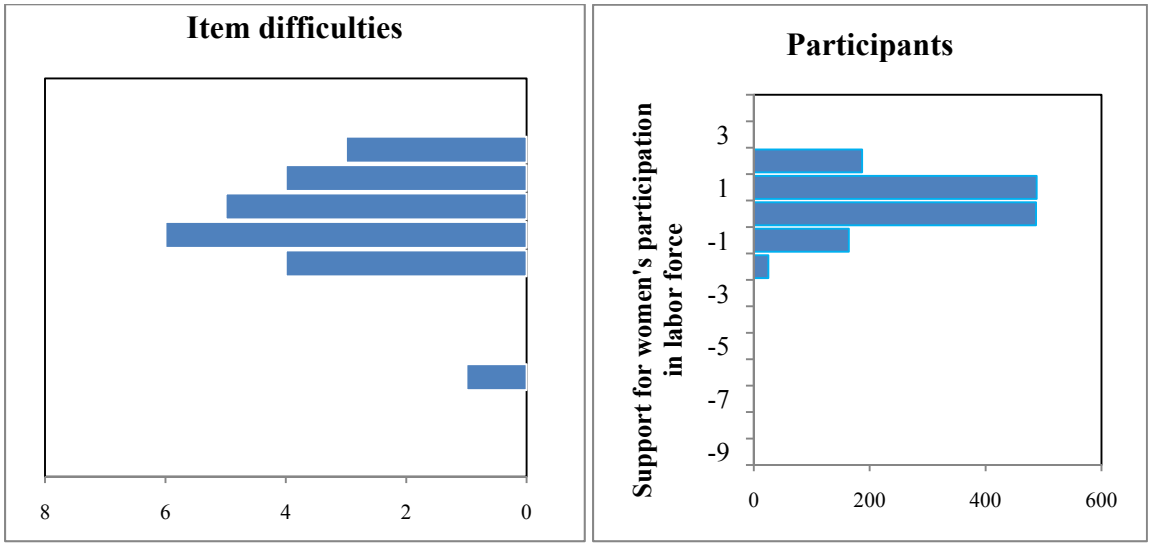

Figure 5.2. Histogram of people and item difficulty distribution – Germany

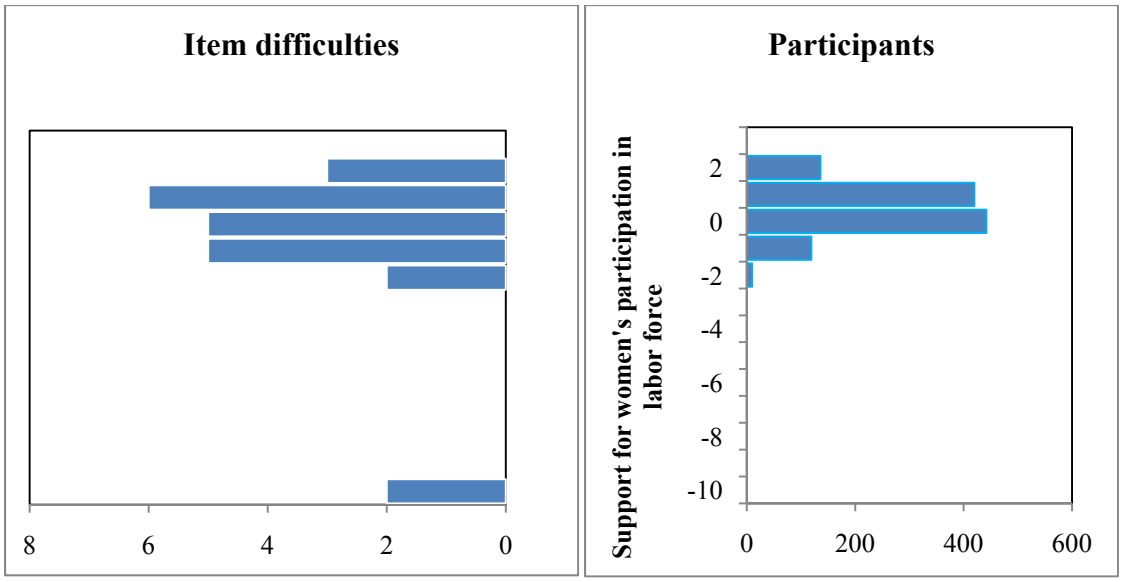

Figure 5.3. Histogram of people and item difficulty distribution – Japan

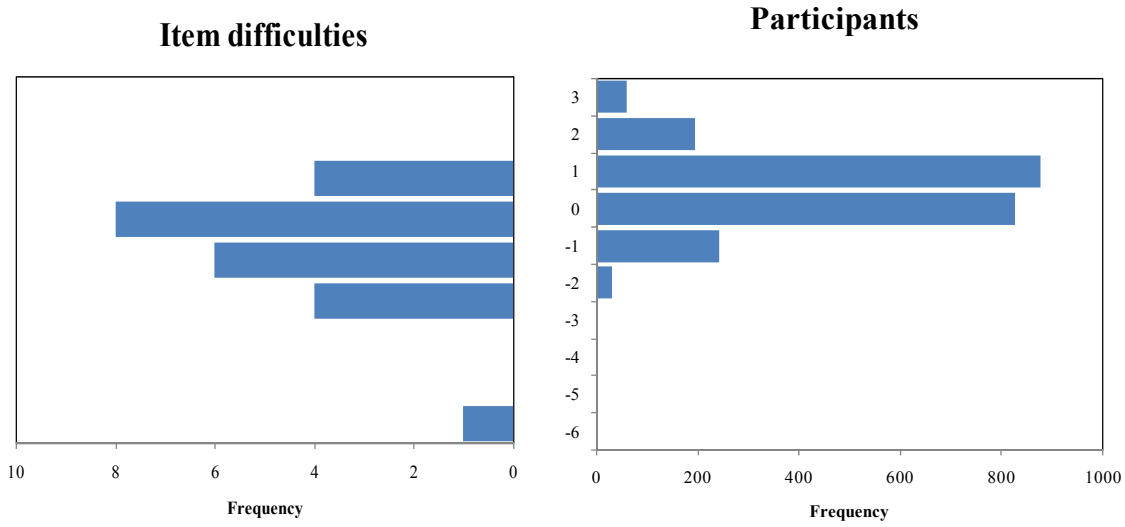

Figure 5.4. Histogram of people and item difficulty distribution – Denmark

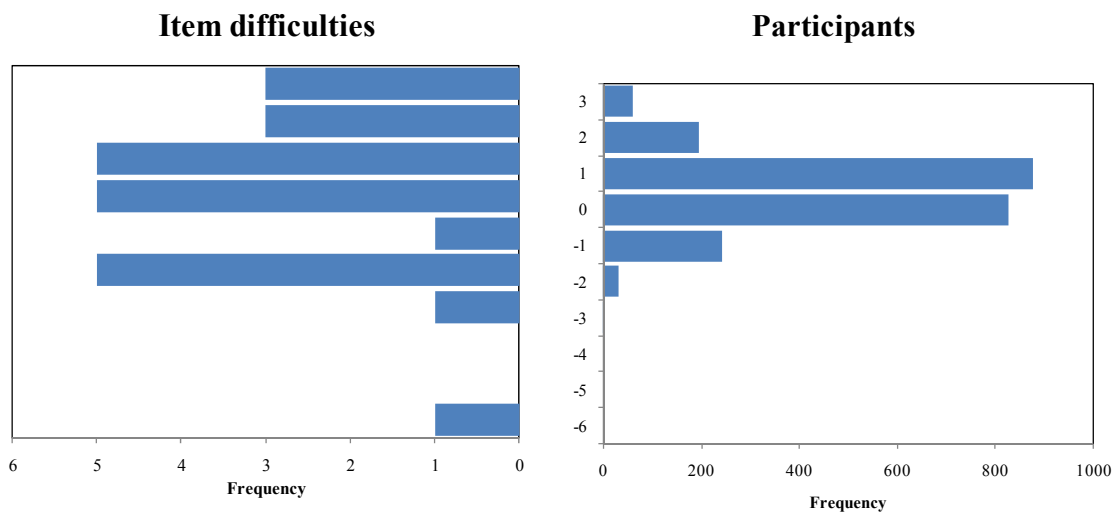

Figure 5.5. Histogram of people and item difficulty distribution – Spain

Total information curves

Figures 5.6 to 5.10 present the Total Information Curve for each country. These graphs represent reliability of the full scale across the attitudinal spectrum. The higher the curve, the better the scale covers that particular area of the attitude, and the wider the curve, the larger the attitudinal spectrum that is measured by the scale.

The scale in Denmark seemed to be reliable for a smaller range of people than in Spain or Great Britain. The lack of items that measure a level of attitude around -1 observed in the histograms for Spain (Figure 5.5) is reflected now on a bump in the total information curve. That portion of the latent attitude is less reliably measured than other parts. Similarly, Great Britain curve showed bumps on both sides, reflecting the lack of agreement between people and item difficulty distribution at the +2 and -2 standard deviations from the mean level of support observed in figure 5.1. The items work best as a set in Germany and Denmark, where they reliably measure a larger part of the latent attitude than the same items in the other countries.

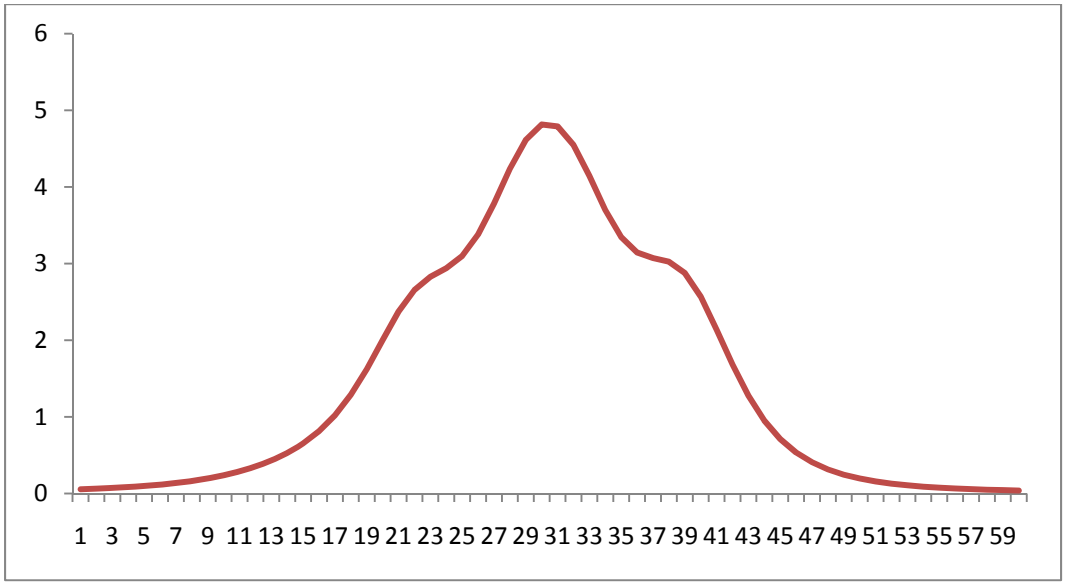

Figure 5. 6.Total information curve – Great Britain

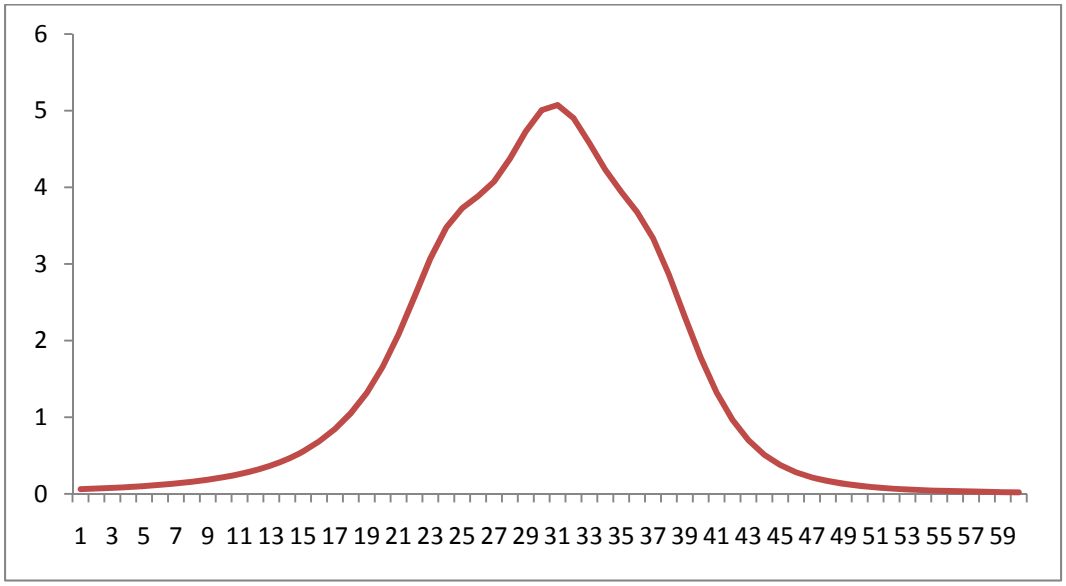

Figure 5. 7.Total information curve – Germany

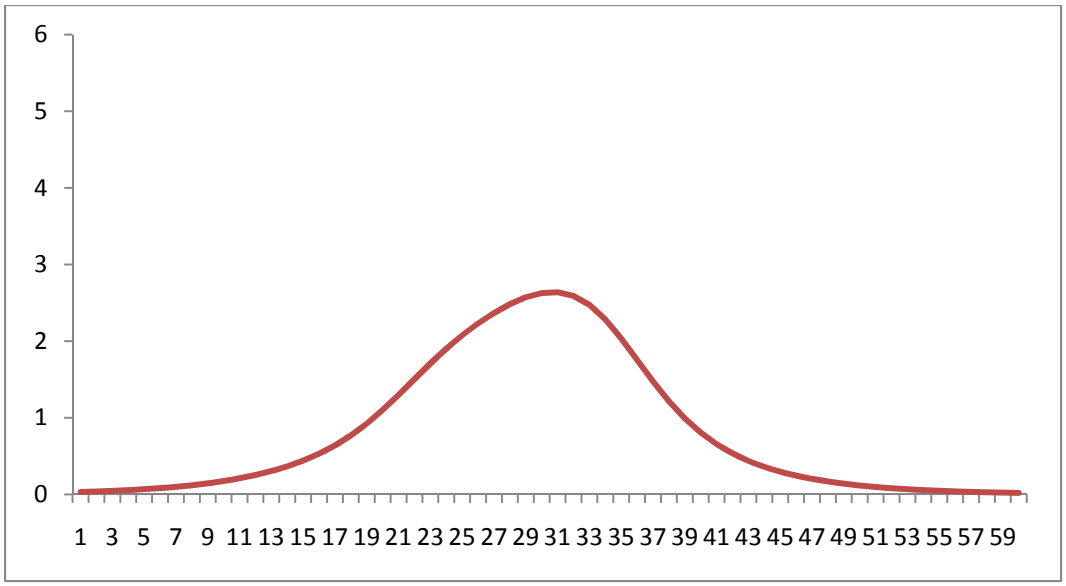

Figure 5.8. Total information curve – Japan

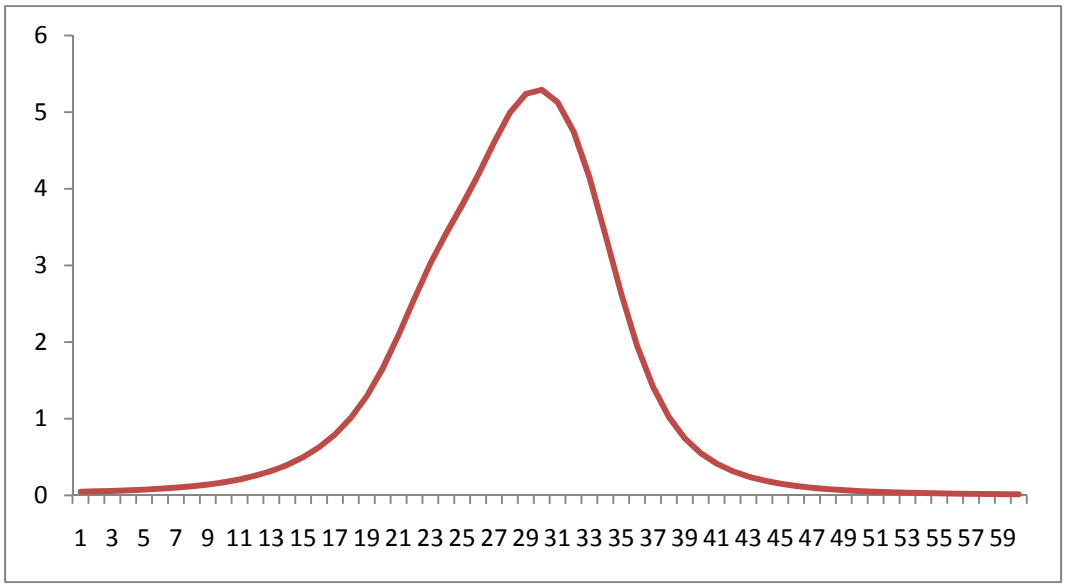

Figure 5.9. Total information curve – Denmark

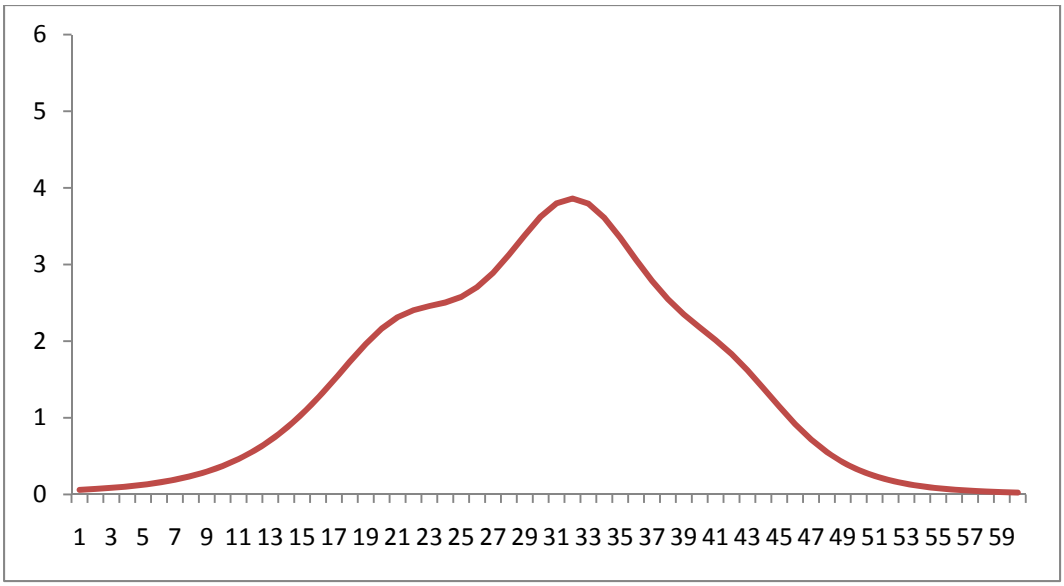

Figure 5.10. Total information curve – Spain

Response Characteristic Curves

Response characteristic curves represent graphically the *a* and *b* parameters estimated in the model. Figure 5.11 can serve as an example to illustrate the meaning of all the figures of response characteristic curves (figures 5.11 to 5.31). In figure 5.11 (and following), each line represents one response category. The distribution reflects what level of the attitudinal latent variables is best measured by the response category. Response categories with steeper, taller curves are preferable, because they better discriminate respondents of a particular level from others. Figure 5.11 depicts the curves for item 5 in Great Britain. The graph was chosen because the curves of most response categories are close to ideal. The only exception is the middle category. This category is not most likely to be chosen at any level of the attitudinal continuum, which makes it less useful for the purpose of the item—to measure the attitude and provide a score with which to differentiate individuals.

Figures 5.12 through 5.26 present the response characteristic curves for three items in each of the countries. Such response characteristic curves represent graphically the *a* and *b* parameters presented in tables 5.12 through 5.16).

The graphs show that response category 3 (*neither agree nor disagree*) was virtually never the most likely to be selected for any range of the underlying attitude for all countries, which questions the usefulness of using such category in this particular context, at least in the way it was worded.

There were marked differences between the response characteristic curves for categories two and four in Denmark as opposed to the countries that use a closely translated scale (Germany and Spain) or a questionnaire in the source language (Great Britain). In Spain, Germany, and in Great Britain all four response categories seemed to be most used by a considerable part of the sample, and for several items the division between *agree* and *disagree* is quite even (see, for example, item 5 in Great Britain). However, in Denmark those two response categories (*somewhat agree* and *somewhat disagree*) were less discriminant and have overlapping difficulties. Overlapping difficulties mean that a person with a certain level of the latent construct is as likely to select either one of the two labels. Having only one of two labels with a lot of overlapping would be a design consequence of such findings.

In general, these response characteristic curves of Great Britain, Germany and Spain discriminate reasonably well cross the spectrum of attitudes towards women being in the labor force, whereas the two countries where a modification of the scale took place. Denmark and Japan, the curves overlap considerably.

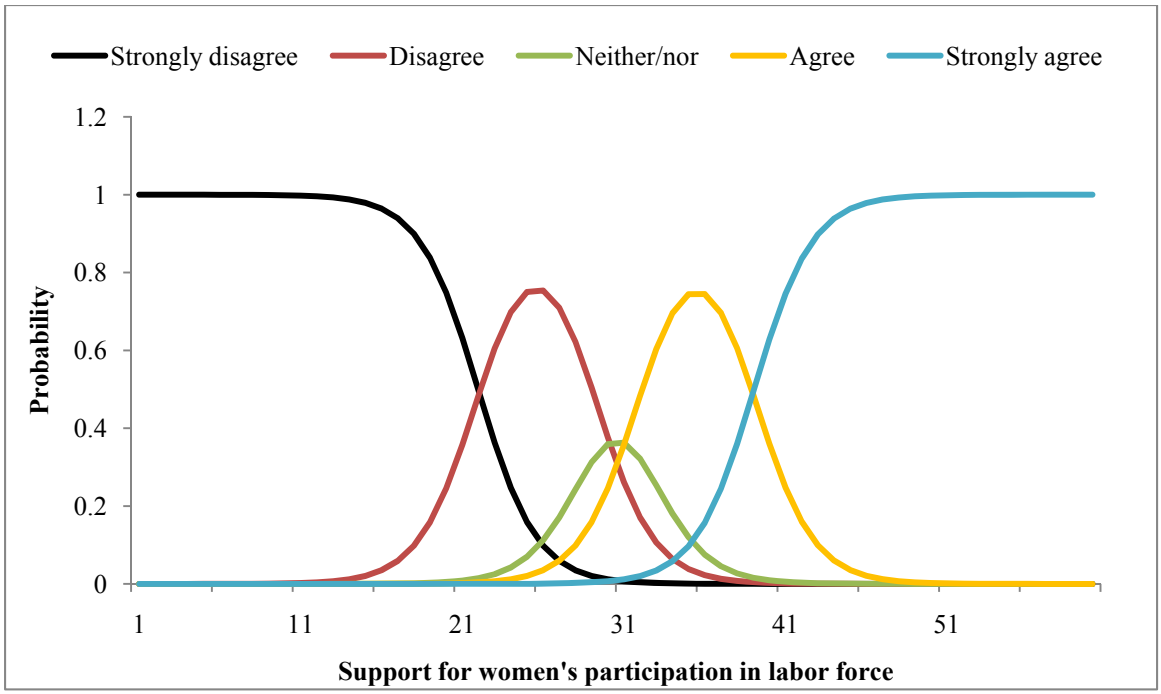

Figure 5.11. Response characteristic curves, item 5 – Great Britain

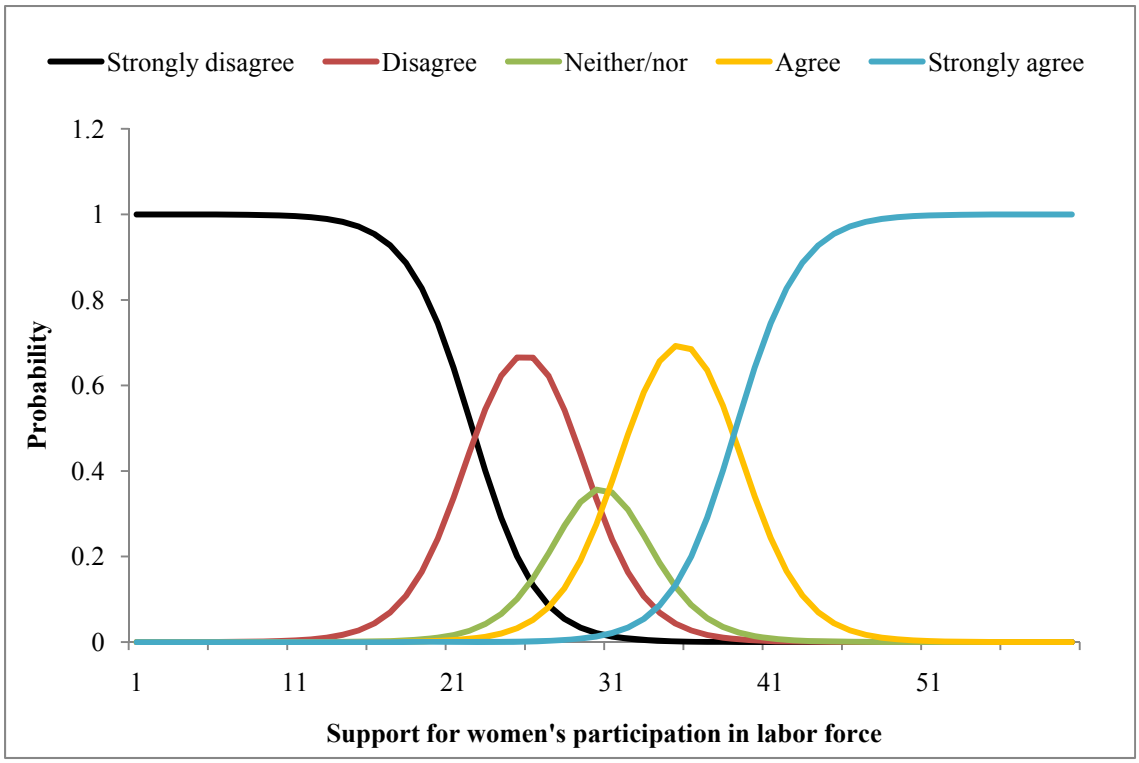

Figure 5.12. Response characteristic curves, item 6 – Great Britain

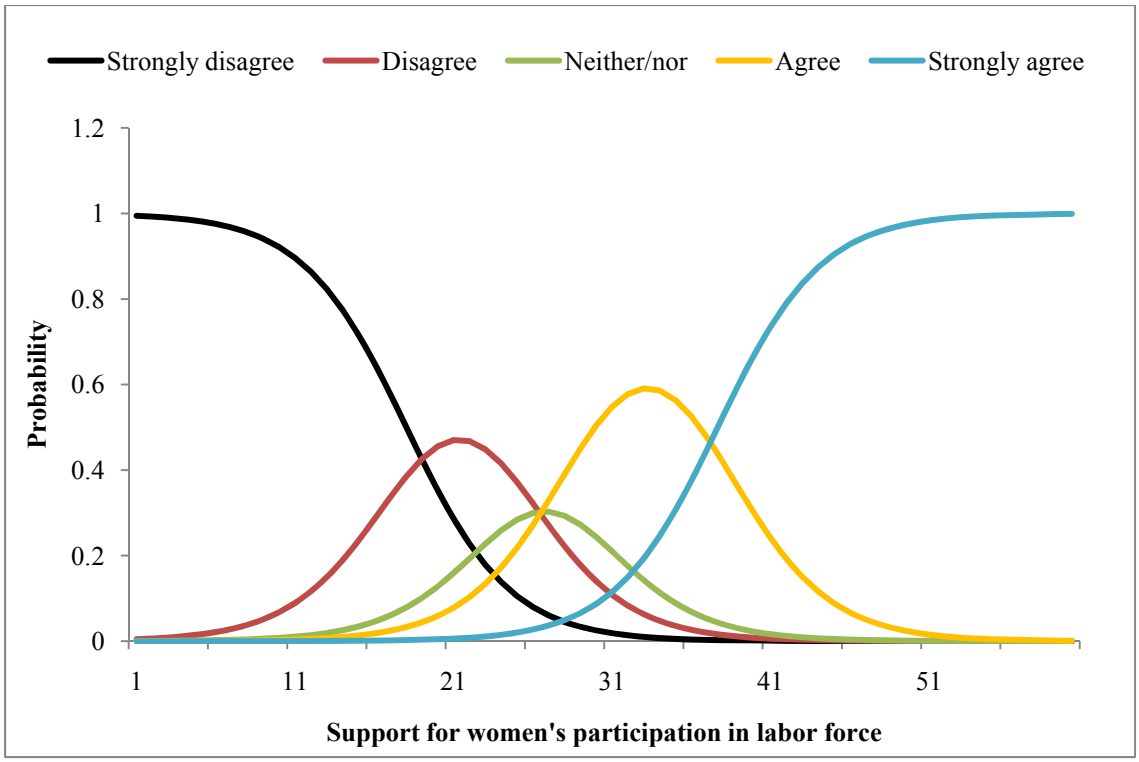

Figure 5.13. Response characteristic curves, item 11 – Great Britain

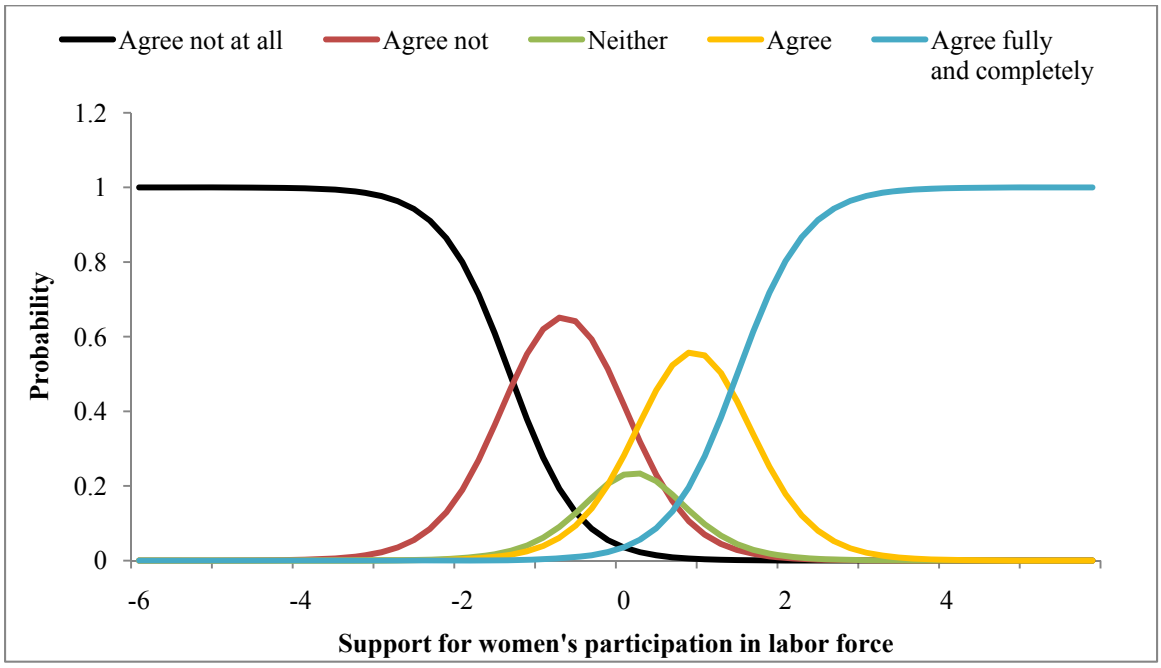

Figure 5.14. Response characteristic curves, item 5 – Germany

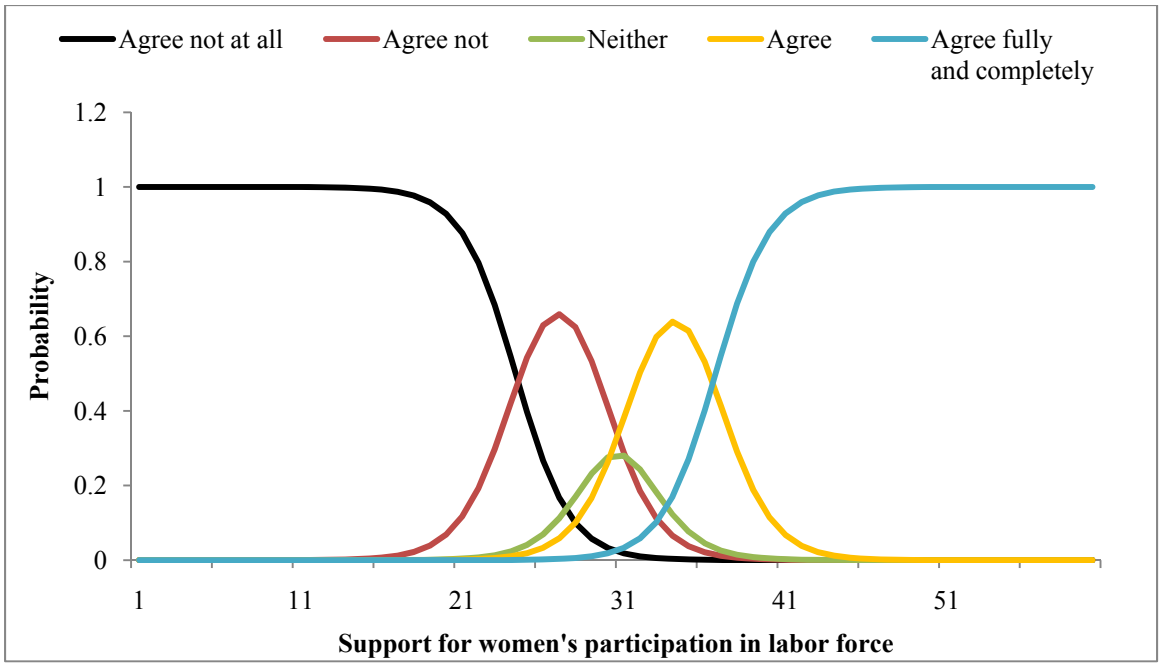

Figure 5.15. Response characteristic curves, item 6 – Germany

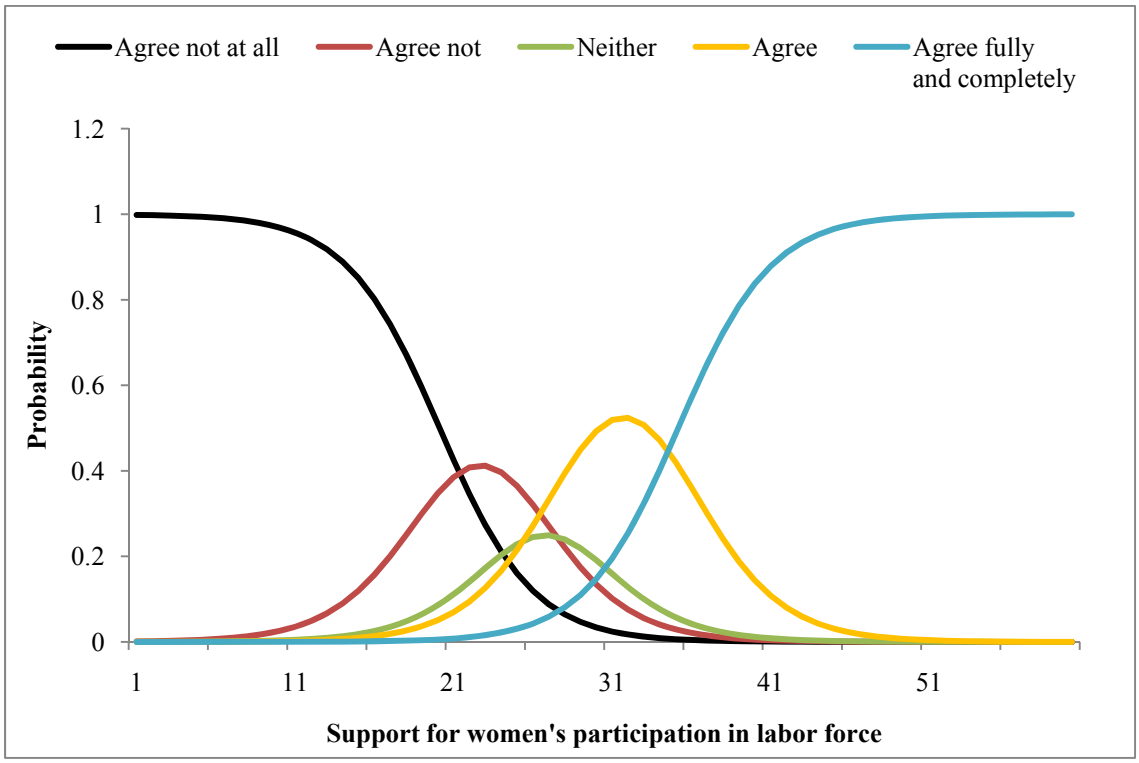

Figure 5.16. Response characteristic curves, item 11 – Germany

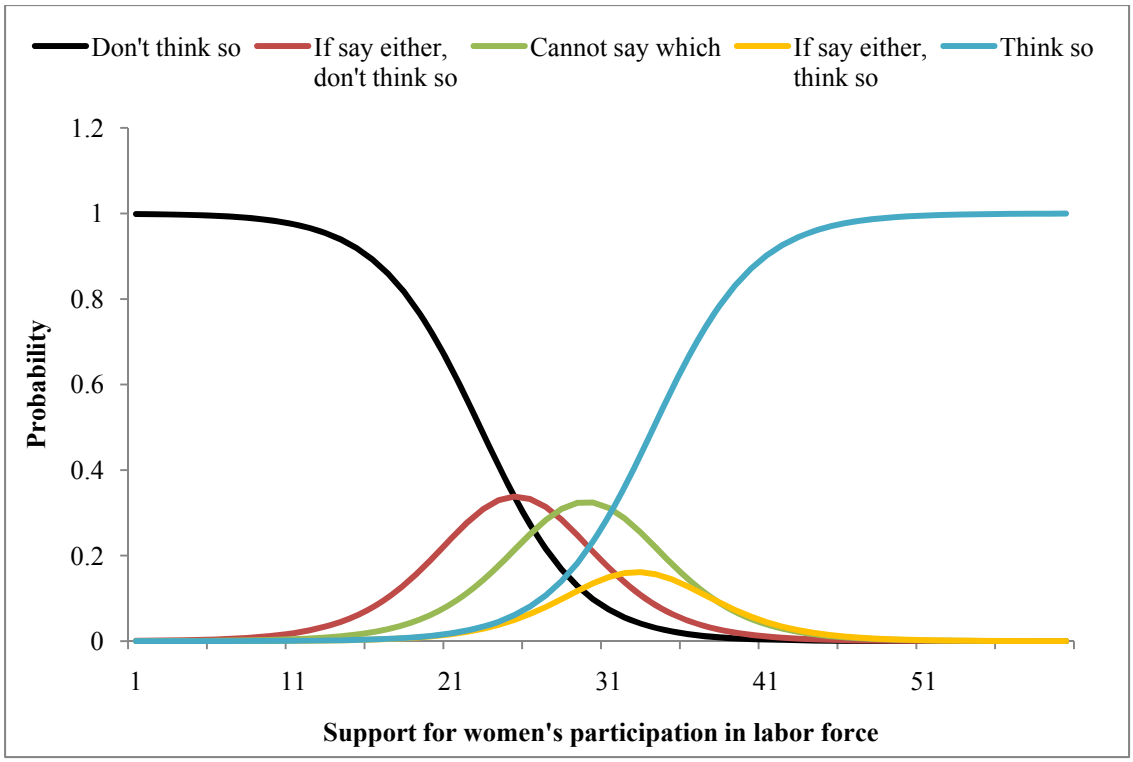

Figure 5.17. Response characteristic curves, item 5 – Japan

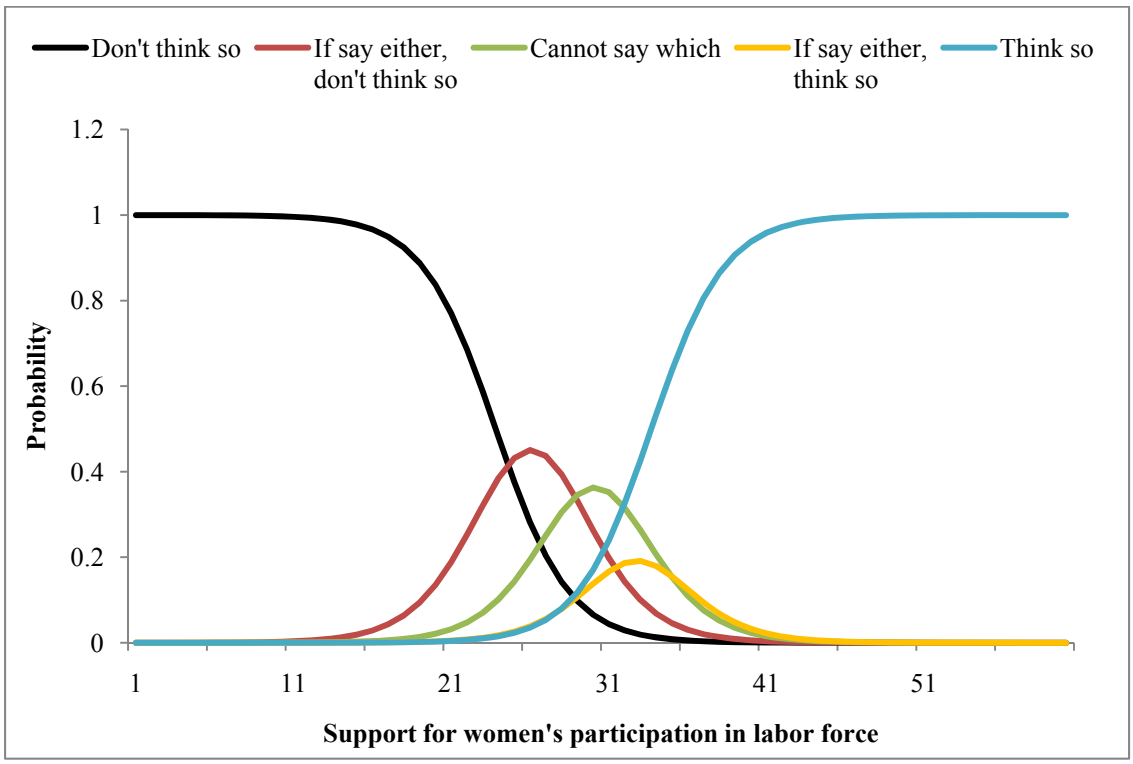

Figure 5.18. Response characteristic curves, item 6 – Japan

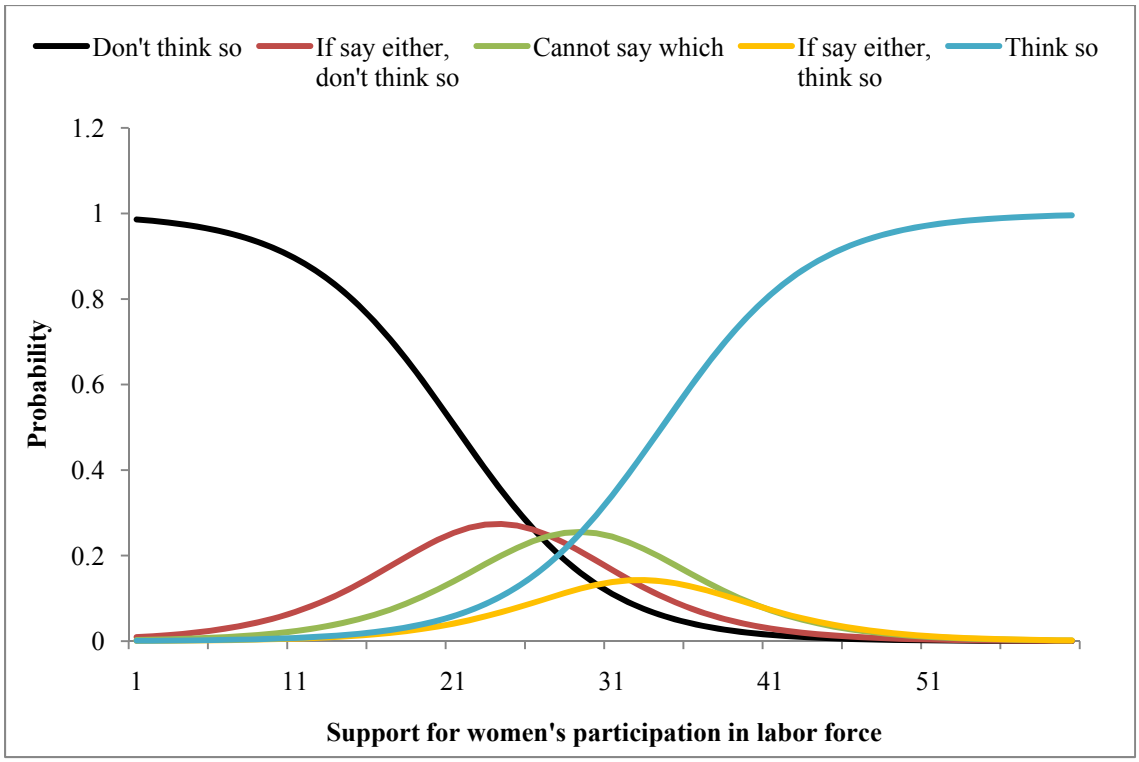

Figure 5.19. Response characteristic curves, item 11 – Japan

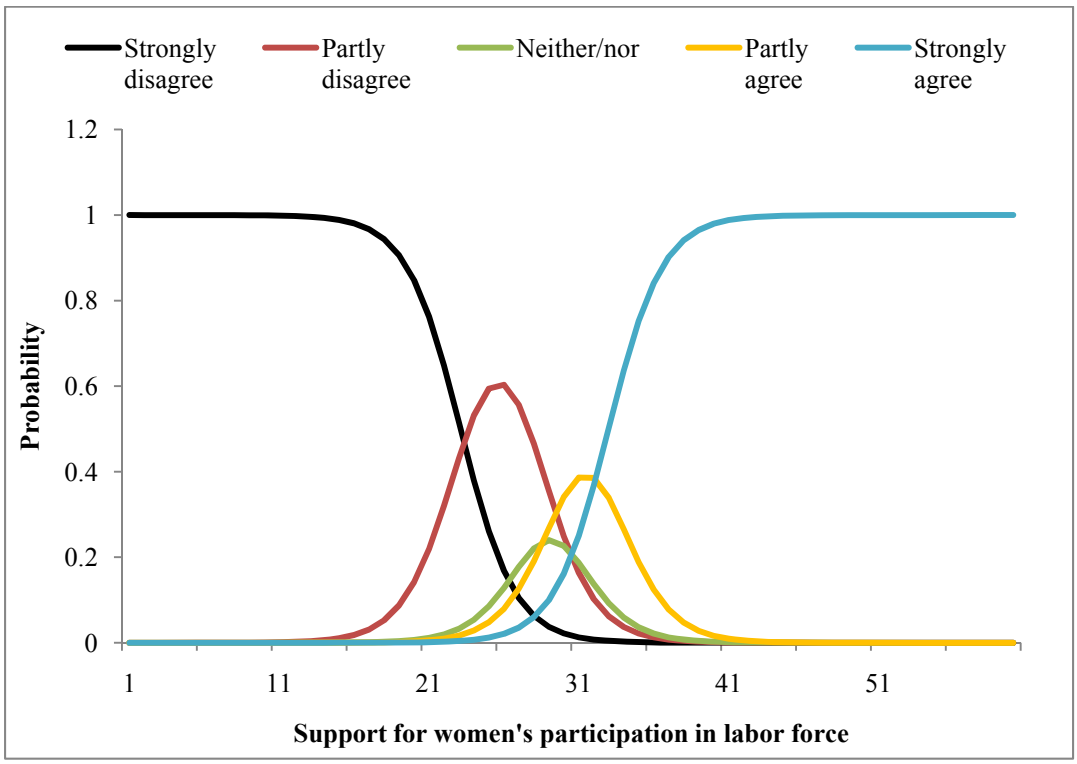

Figure 5.20. Response characteristic curves, item 5 – Denmark

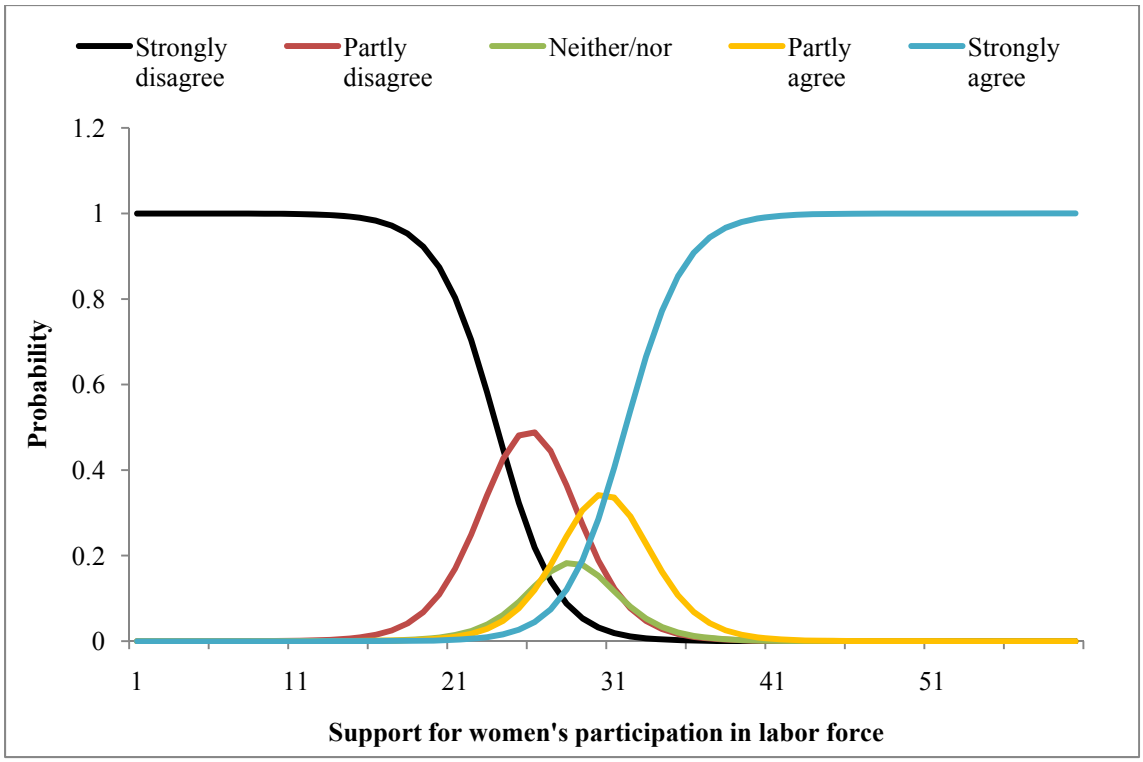

Figure 5.21. Response characteristic curves, item 6 – Denmark

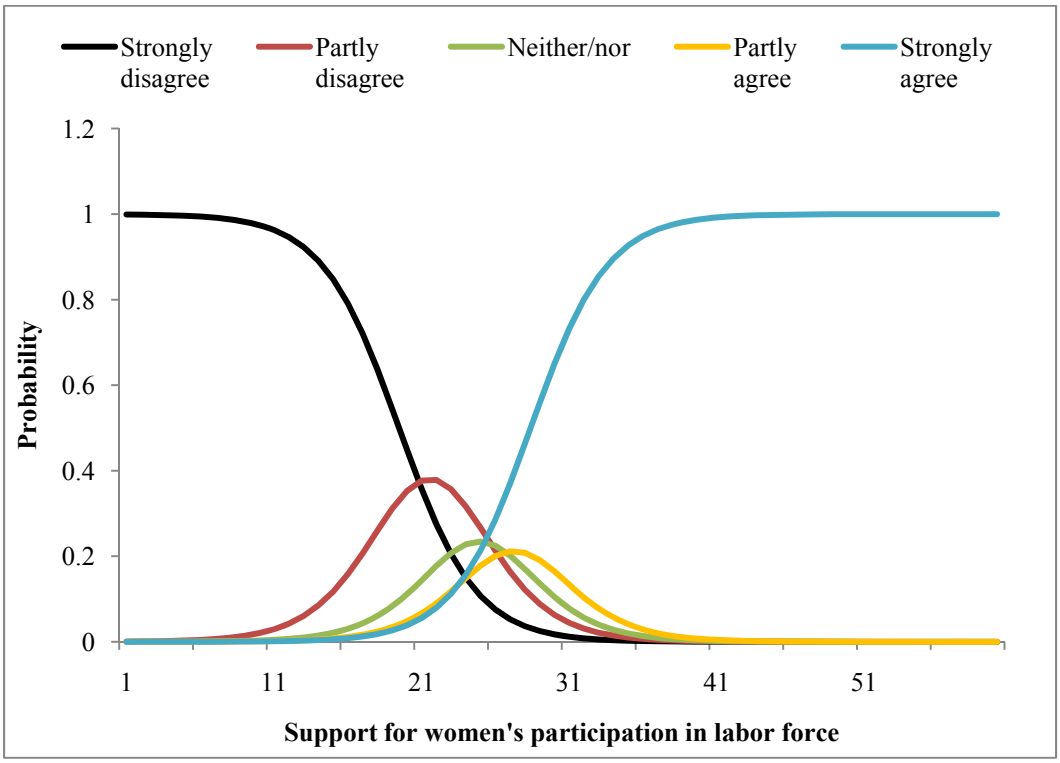

Figure 5.22. Response characteristic curves, item 11 – Denmark

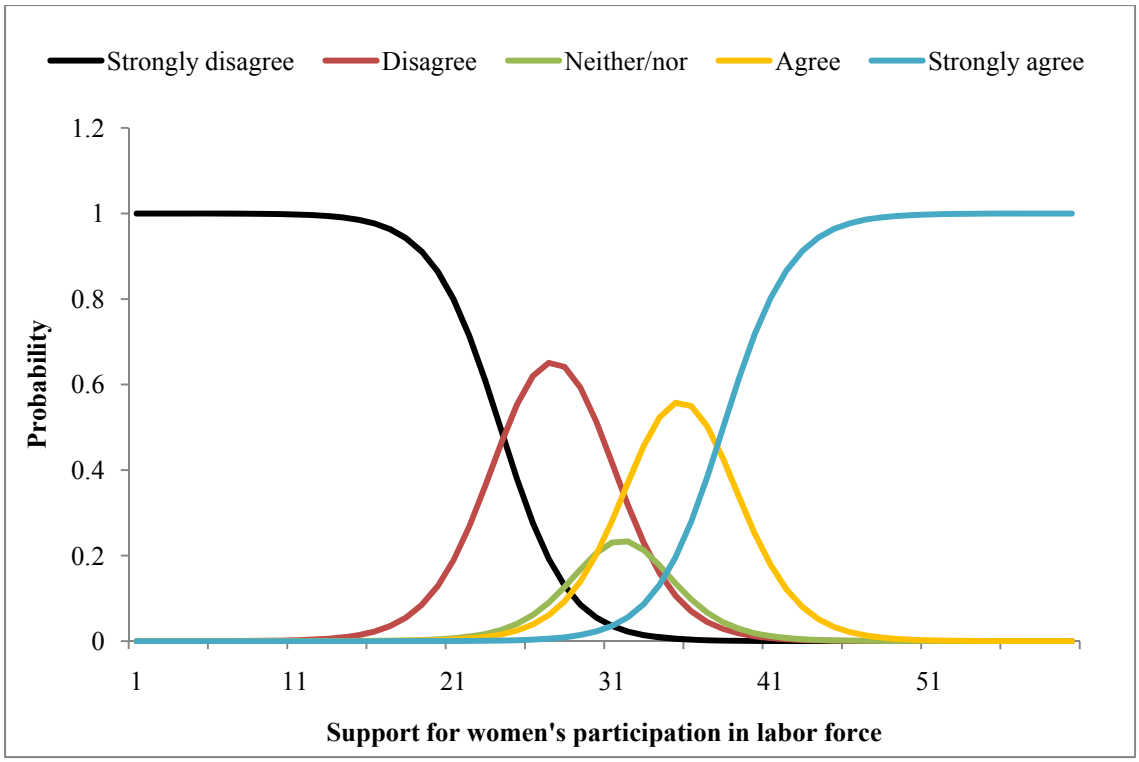

Figure 5.23. Response characteristic curves, item 5 - Spain

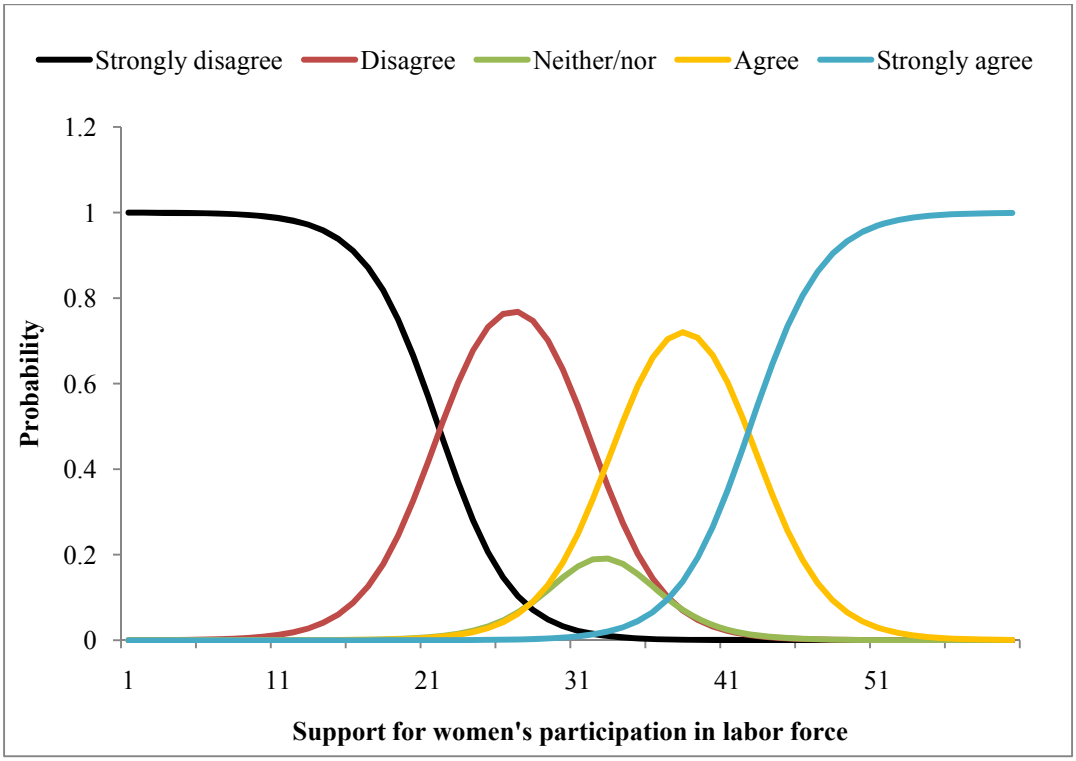

Figure 5.24. Response characteristic curves, item 6 – Spain

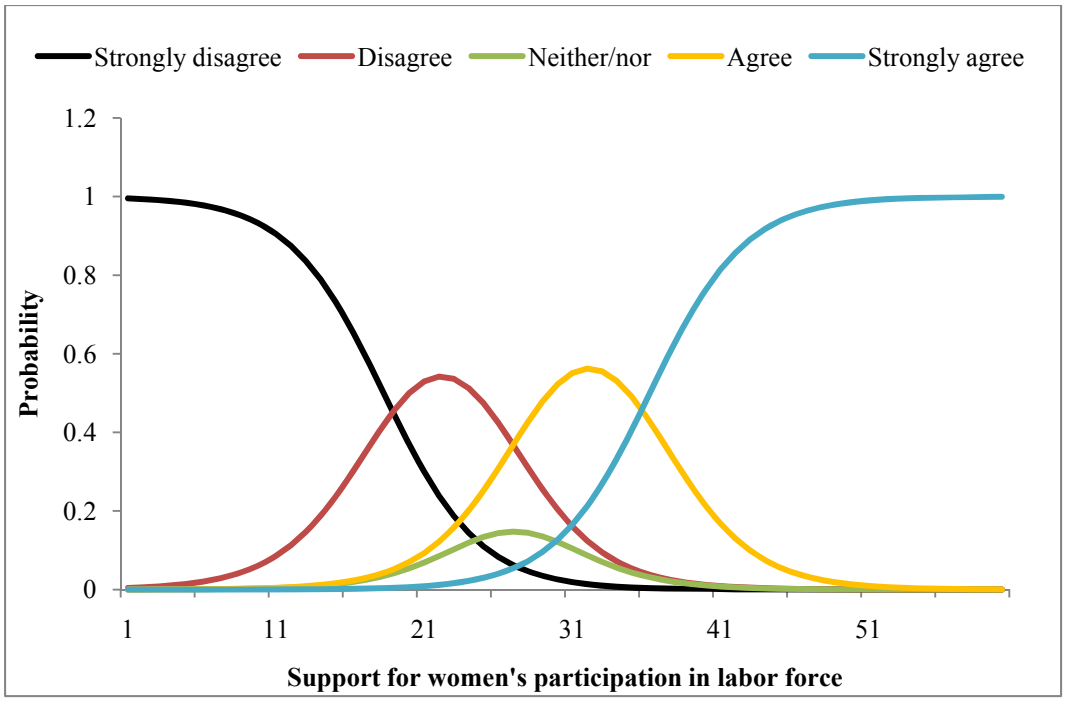

Figure 5.25. Response characteristic curves, item 11 – Spain

Discussion

The IRT models tested here suggest that the scale does not have internal consistency, even though Cronbach's alpha was acceptable and very similar for all countries except Japan. Similar value of Cronbach's alpha across language versions of an instrument are sometimes taken as evidence for equivalence of translations, but the IRT analyses showed that such conclusions may be misleading. Figures 5.6 through 5.10 showed that reliability differed considerably across countries. Furthermore, analyses of invariance all revealed that the scales were not invariant across countries, and therefore no evidence was found that supported the notion of comparable scales.

On the other hand, the scale items did not seem to work well together. There might be wording problems with some of the items, as illustrated with items 8 and 9. Women roles have changed so rapidly in the past decades that the questions, written in

the nineties, cannot reflect today's discourse on the topic. The idea of outdated items is reinforced by the fact that the scale seemed to be more cohesive and items tended to work better in Spain, a country were gender roles are somewhat more traditional than in other European countries.

Braun's work on these ISSP items (Braun & Scott, 1998; Braun & Harkness, 2003) suggest that the items are understood to mean different things in different countries. In societies with high female labor-force participation, questions may be given an economic interpretation, rather than an ideological one. In societies where paid work is less common for women, a gender-ideology interpretation is more likely.

Overall, the findings seem to support the idea that response distributions are affected by the way an answer scales are designed. The findings on the use of the second and fourth categories of the answer scale (*agree/somewhat agree, disagree/somewhat disagree*) point to the need of more careful consideration of translation and adaptation effects when conducting multilingual surveys. In Denmark, the two most useful categories to distinguish between people supportive of women working from those who have a more traditional view of women's roles were *strongly agree* and *strongly disagree*, whereas the other two categories were much less helpful in differentiating people on the latent trait.

The results for Japan are particularly interesting. All seems to indicate that the set of items does not work well in Japan. Part of it could be the modified answer scale. However, the shape of the curves is similar to that in Denmark, and yet the group of final items does reach high reliability in Denmark.

One of the limitations of the previous results is that the true values of the individuals in each nation may be driving the observed differences. Countries differing in how they see women and their roles in society is, inevitably, an alternative explanation to these results. In order to gain some perspective on the role of verbal labels in these differences, an additional set of IRT models was estimated that combined countries that used translation of the answer scale and compared them to a combination of countries that had added an intensity modifier. Care was put into combining countries that were geographically and culturally distant in each category (modified intensity vs. translation), and having countries from the same regional and cultural groups represented in both categories. The selected countries were Brazil, Denmark and the Czech Republic as representative of the modified intensity answer scale, and Mexico, Germany, and Poland for the translated scales.

The results found before with respect to information and difficulties were generally the same as those country by country. The most important piece of information comes from the response characteristic curves, depicted in figures 5.26 to 5.31. The same differences in response category curves seen between Great Britain and Denmark were found for these two groups. As can be seen in figures 5.29 to 5.31, the second and fourth answer scale points were less "useful" to measure the attitude of interest for the group of countries that used a modified answer scale (figures 5.26 to 5.28). Differences are clear, and they replicate the results from the selected individual countries. Having used countries with such different backgrounds in gender roles and family structures, it seems unlikely that substantive differences can account for the different response characteristic

curves observed for the three countries were translation was applied and those for the three countries were an intensity descriptor was added.

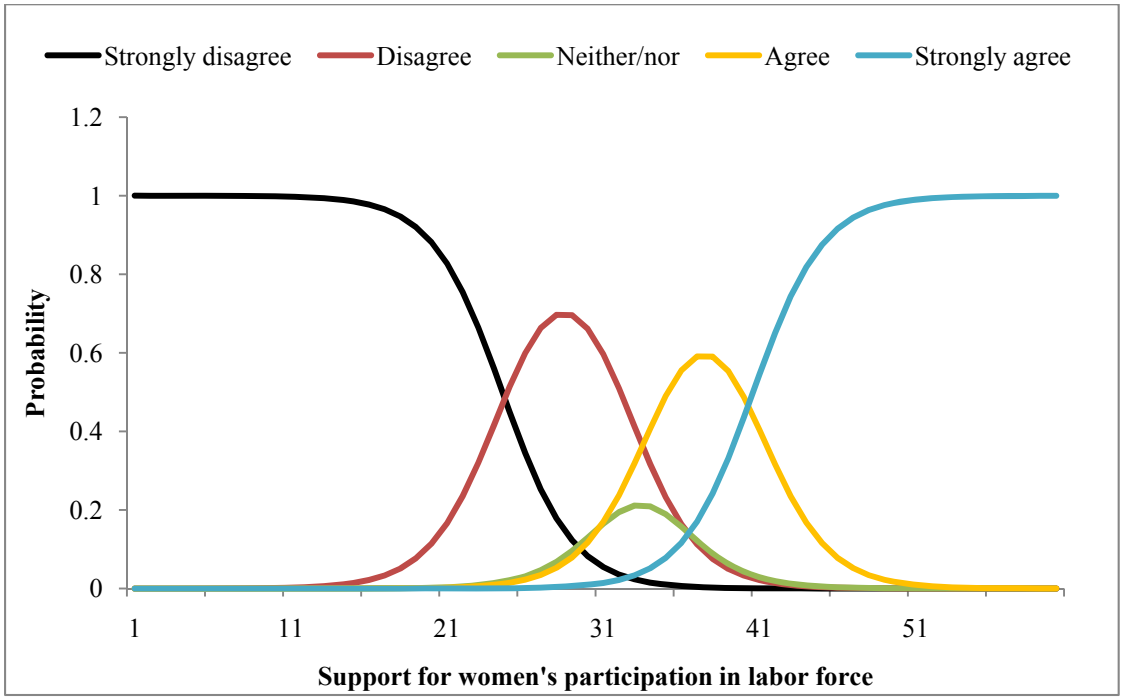

Figure 5.26. Response characteristic curves, item 5 – Closely translated

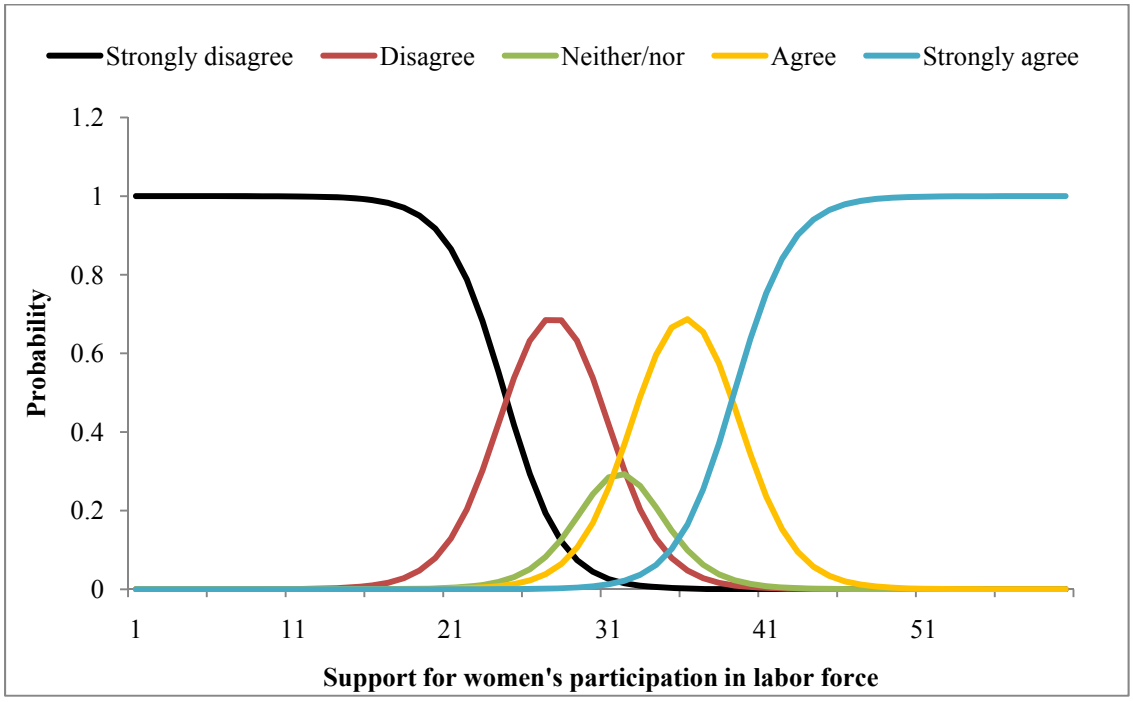

Figure 5.27. Response characteristic curves, item 6 – Closely translated

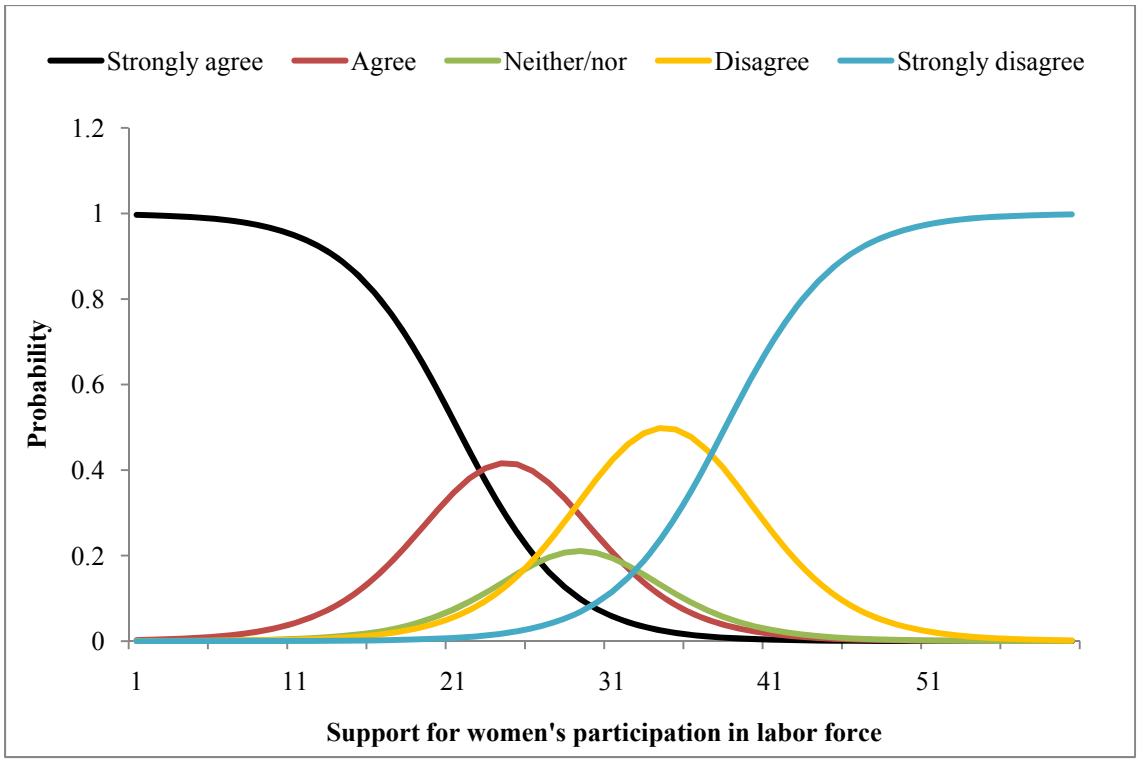

Figure 5.28. Response characteristic curves, item 11 – Closely translated

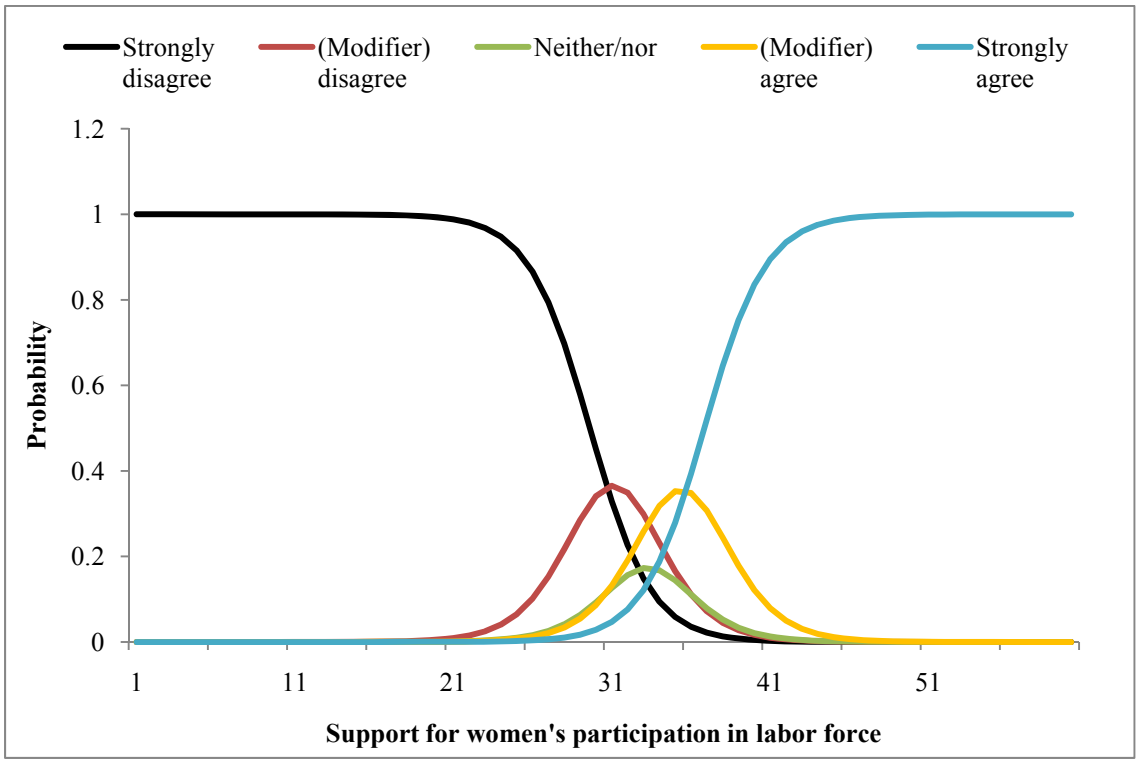

Figure 5.29. Response characteristic curves, item 5 – Modified Intensity

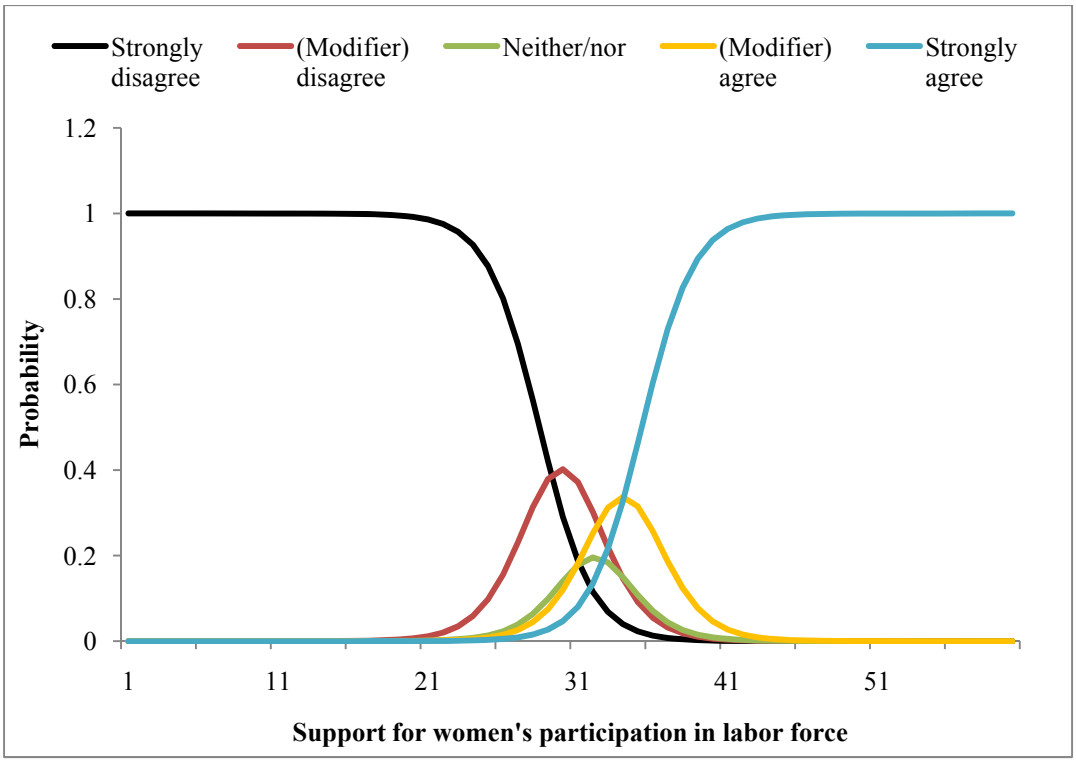

Figure 5.30. Response characteristic curves, item 6 – Modified Intensity

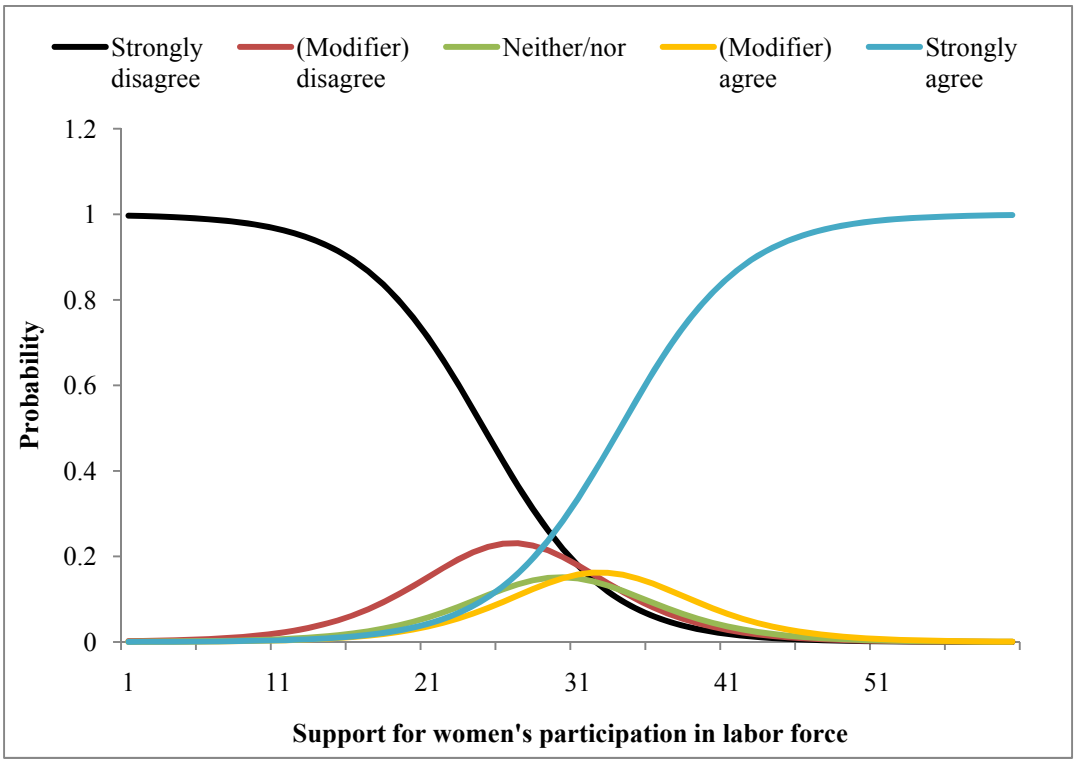

Figure 5.31. Response characteristic curves, item 11 – Modified Intensity

Chapter 6. Discussion and Conclusions

This research provides evidence for the presence of methodological, cultural and individual predictors of response styles. The effect of the answer scale version is consistent throughout the five ISSP modules analyzed. Furthermore, it was the only predictor that consistently had a large effect on variance reduction, and in the same direction (more extreme response when a modifier is added to *agree*). All other variables showed either lack of significant effect, or a variable effect.

The inconsistent findings, in line with the existing empirical research, where predictors are sometimes found to be significant and sometimes not, can be interpreted as a sign that response styles indicators differ across topic of research. This, in turn, can mean two things. Either the traditional (and not so traditional) response style indicators are not indications of response styles as defined in the literature (stable, pervasive, personal tendencies), or response styles as such do not exist. The findings from this dissertation question traditional views of response styles and response style research. With regard to cultural differences in response styles, I have shown that a methodological variable was the strongest of the included predictors of extreme response style differences. Confirming expectations, a substantial amount of the variance across countries in extreme response style, previously attributed to personal factors and cultural dimensions, was due to methodological aspects of the survey.

Countries' decision to modify the scale points rather than translate answer scales has consequences in data quality. Future research needs to investigate how to design adequate answer scale points, both for monolingual and multilingual research. Therefore, such decisions should not be based on experts' intuition but on empirical evidence for comparability of verbal labels across languages.

These findings also suggest that respondents "pay attention" to the verbal labels offered in answer scales. The likelihood that a respondent selects the scale point *strongly agree* is affected by how the adjacent scale point is labeled. Therefore, verbal labels affect the matching process between the respondents' subjective response categories offered by the researcher. Furthermore, this effect seems to happen in several countries.

The results presented here are specially relevant for cross-cultural research that involves translation, whether such research has a a substantive nature (e.g., comparing environmental attitudes across different populations) or a methodological one, such as the reviewed literature on cultural differences in response styles. The findings are also relevant for multilingual research within countries, for monolingual research across countries, and even for monolingual research within countries. Even when the U.S. General Social Survey was still a monolingual survey, changes in verbal labels of the agreement answer scale were found throughout the questionnaire: answer scales had 2, 4, or 5 scale points; the endpoints were sometimes labeled as *agree* and sometimes *agree strongly*. The second and fourth points of a 5-point answer scale were sometimes labeled *agree,* sometimes *slightly agree*, and yet sometimes *somewhat agree*.

The effect of mode on acquiescence was less clear. Mode helped explain country variance in response styles, but the effect was reduced after controlling for cultural

variables. At the same time, individualism/collectivism was significant when the effect was estimated alone or with other cultural dimensions. This could be suggesting an interaction effect between mode and individualism/collectivism. Because the collectivistic cultural syndrome focuses on maintaining harmony with the group, the presence of the interviewer could have a stronger effect than for individualistic nations. However, the small sample size at country level when including the cultural dimensions made testing an interaction effect of mode and individualism/collectivism unadvisable.

Researchers studying cultural differences in response styles for multilingual contexts should carefully examine how answers scales were translated, as well as other methodological features that may vary across the countries or cultures. This affects both primary and secondary research, but may be more dangerous in secondary research for two reasons: a) if documentation is scarce, the researcher will not be able to appraise the quality of the methodology applied—in which case it may be advisable not to conduct analysis involving these questions; b) when data are publicly available, the potential for data exploitation and therefore for wrong inference is larger than when data are kept private. Future research could benefit from incorporating methodological manipulations in modeling the causes of response styles.

The evidence presented in this dissertation encourages a number of considerations relevant both for substantive research and for methodological research:

a) Researchers in large cross-national programs need to carefully document key methodological aspects of their studies (Harkness, 2008; Jowell, 1998; Mohler, et al., in press). Availability of instruments and related materials in all languages is increasing but it should become the norm.

b) When it comes to answer scales that are repeatedly used throughout a questionnaire and often times across studies, such as the agreement scale discussed in this study, better developmental procedures are needed, both in monolingual and multilingual research.

c) Researchers approaching secondary data should be aware of the potential implications of translation and adaptation. Rather than assuming comparability in the translations of instruments, empirical evidence should be sought. The IRT model presented in chapter 5 is an example of a statistical technique that can help discern whether questions and answer scales are comparable across countries. One recommendation found in the literature is to look at the translation when something in the data "doesn't make sense", but this dissertation has shown how differences that seemed to match researchers' conceptions of the cultures involved can be to some extent due to a methodological artifact. At the same time, having a complete assessment of all the instruments involved in the study is rather cumbersome and costly. There is little collective knowledge regarding what survey elements may be more susceptible to mistakes in survey translation. This dissertation, in line with previous research (Braun, 2003; Harkness, 2003; Harkness, et al., 2004; Harkness, et al., 2008; Harkness, et al., 2009) indicates that answer scales are key elements to investigate, because: a) answer scales tend to affect several questions, therefore the impact of a translation mistake will carry over, rather than distorting only one question; b) modifications to answer scales exist in numerous important cross-national studies where available documentation

exists, but that no documented rationale is provided for those changes and most likely there is no empirically-driven rationale for such decision, decision that is made at country level. One exception is the work by the QoL group, where the methodology applied to translate answer scales is decided and guided by the steering group.

Findings from this dissertation question the validity of the current definition and operationalization of response styles. Except answer scale version, none of the predictors was consistently related to extreme response style. For acquiescence, it was only age and education that had a systematic effect on acquiescence. The lack of consistency of significance of predictors across years suggests that the indicators are not driven by the same variables each year. This questions the notion that traditionally indicators are actually measuring stable response styles. Unless one argues that all the indicators used in this dissertation—including traditional and newer measures—are unrelated to the hypothesized response style construct, this can be taken as additional evidence that response styles are artifactual constructs, rather than inherent, stable response tendencies.

Limitations of this dissertation

In order to clearly disentangle the effect that different wording of labels of answer scales has on how respondents answer surveys more research is needed where the same or equivalent populations are exposed to both forms of the answer scale for the same items. That way, cultural differences, substantive differences and any other potential confounding factors would be controlled for, and better insight could be gained.

In addition, a criticism applicable to this and other studies of response styles lies on the way that they are measured. This is related to the lack of external reference for response styles. Empirical evidence reflecting the effects of the hypothesized response styles in other situations and contexts is needed. Such evidence could serve as external validation of the findings.

Numerous approaches could be taken for this. For example, discourse analyses could be used to evaluate how labels commonly used in survey research perform in more ordinary usages of language. In connection to this, corpora and frequency dictionaries can also be consulted so that richer information about the usage of certain terms in each language can be compared. Corpora analyses could reveal patterns that no other methodological approach could cover.

Different pretesting techniques could also be used to investigate answer scale design and verbal labels design. Behavioral coding of survey interviews could be used to examine latency of response, intensity in the tone of voice when providing the answer, or response clarifications that respondents spontaneously provide. Cognitive interviews are also great resources from which information can be gathered about how respondents interpret verbal labels that are used in surveys.

More sophisticated statistical techniques that the ones used in this dissertation have been proposer to obtain response style indicators. Structural equation modeling (Cheung & Rensvold, 2000; Holbrook et al, 2006; Welkenhuyzen-Gybels, Billiet, & Cambre, 2003), latent class analysis (Moors, 2003; 2004; Morren et al., 2008), or IRT (Tanzer, 2005) have been used in attempts to “extract” response style scores that are independent from the content of the items.

In sum, there is no way of guaranteeing that the more or less frequent choice of endpoints is due to bias rather than substantive differences, just as it is not possible, presently, to guarantee that substantive interpretations are free from bias. The combination of sophisticated statistical techniques with the design recommendation of including low-correlated, content-diverse items (Greenleaf, 1992) could provide better indicators of what is intended to measure (a stylistic tendency that manifests across topics), and thus help advance response styles research.

Finally, this dissertation represents only a step towards better understanding how to design answer scales, both in monolingual and multilingual contexts. The findings of this dissertation are limited to 5-point bipolar agreement answer scales. No conclusions can be made regarding that the effect of the addition of downtoning intensity modifiers would have the same magnitude when used on answer scales that measure different dimensions (e.g., satisfaction, certainty, importance, and so on).

If differences are found, as they are found in this study from year to year, explanations of why the magnitude of the effect varies could be also investigated. Future research expanding findings to more answer scale dimensions can provide insights on why variation occurs, and contribute towards a theory of how respondents use and interpret language in the survey context.

Appendix

This appendix contains detailed tables for the above analyses. It also contains a more detailed account of design features of the ISSP. The appendix tables are ordered by chapter.

Table A.1.
Question wording 1999. Inequality module

□	In [Country] people get rewarded for their effort
□	In [Country] people get rewarded for their intelligence and skills
□	To get all the way to the top in [Country] today, you have to be corrupt
□	Inequality continues to exist because it benefits the rich and powerful
□	No one would study for years to become a lawyer or doctor unless they expected to earn a lot more than ordinary workers
□	Large differences in income are necessary for [Country's] prosperity
□	Inequality continues to exist because people don't join together to get rid of it
□	Differences in income in [Country] are too large
□	It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes
□	Present economic differences between rich and poor countries are too large
□	People in wealthy countries should make an additional tax contribution to help people in poor countries

Table A.2.
Question wording 2000. Environment module

□	Private enterprise is the best way to solve [Country's] economic problems
□	It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes
□	We believe too often in science, and not enough in feelings and faith
□	Overall, modern science does more harm than good
□	Modern science will solve our environmental problems with little change to our way of life
□	We worry too much about the future of the environment and not enough about prices and jobs today
□	Almost everything we do in modern life harms the environment
□	People worry too much about human progress harming the environment
□	In order to protect the environment [Country] needs economic growth
□	It is right to use animals for medical testing if it might save human lives
□	Economic growth always harms the environment
□	The earth simply cannot continue to support population growth at its present rate
□	It is just too difficult for someone like me to do much about the environment
□	I do what is right for the environment, even when it costs more money or takes more time
□	There are more important things to do in life than protect the environment
□	There is no point in doing what I can for the environment unless others do the same
□	Many of the claims about environmental threats are exaggerated
□	For environmental problems, there should be international agreements that [Country] and other countries should be made to follow
□	Poorer countries should be expected to make less effort than richer countries to protect the environment
□	Economic progress in [Country] will slow down unless we look after the environment better
□	Government should redistribute income from the better-off to those who are less well-off
□	There is little that people can do to change the course of their lives
□	One of the problems with people today is that they challenge authority too often
□	People with money should be left to enjoy it
□	There are times when people should follow their consciences even if it means breaking the law
□	Private enterprise needs to be controlled to protect everyone's needs
□	All societies have inequalities which it is better not to interfere with
□	Taking everything into account, the world is getting better

Table A.3.
Question wording 2002. Family and Gender Roles module

- □ A working mother can establish just as warm and secure a relationship with her children as a mother who does not work
- □ A pre-school child is likely to suffer if his or her mother works
- □ All in all, family life suffers when the woman has a full-time job
- □ A job is all right, but what most women really want is a home and children
- □ Being a housewife is just as fulfilling as working for pay
- □ Having a job is the best way for a woman to be an independent person
- □ Both the man and the woman should contribute to the household income
- □ A man's job is to earn money; a woman's job is to look after the home and family
- □ Men ought to do a larger share of household work than they do now
- □ Men ought to do a larger share of childcare than they do now
- □ Married people are generally happier than unmarried people
- □ It is better to have a bad marriage than no marriage at all
- □ People who want children ought to get married
- □ One parent can bring up a child as well as two parents together
- □ It is all right for a couple to live together without intending to get married
- □ It is a good idea for a couple who intend to get married to live together first
- □ Divorce is usually the best solution when a couple can't seem to work out their marriage problems
- □ Watching children grow up is life's Greatest joy
- □ People who have never had children lead empty lives
- □ Working women should receive paid maternity leave when they have a baby
- □ Families should receive financial benefits for child-care when both parents work
- □ There are so many things to do at home, I often run out of time before I get them all done
- □ My life at home is rarely stressful
- □ There are so many things to do at work, I often run out of time before I get them all done
- □ My job is rarely stressful

Table A.4.
Question wording 2003. National Identity module

□	I would rather be a citizen of [Country] than of any other country in the world
□	There are some things about [Country] today that make me feel ashamed of [Country]
□	The world would be a better place if people from other countries were more like the [Country nationality]
□	General speaking, [Country] is a better country than most other countries
□	People should support their country even if the country is in the wrong
□	When my country does well in international sports, it makes me proud to be [Country nationality]
□	I am often less proud of [Country] than I would like to be
□	[Country] should limit the import of foreign products in order to protect its national economy
□	For certain problems, like environmental pollution, international bodies should have the right to enforce solutions
□	[Country] should follow its own interests, even if this leds to conflicts with other nations
□	Foreigners should not be allowed to buy land in [Country]
□	[Country] television should give preference to [Country] films and programmes
□	Large international companies are doing more and more damage to local businesses in [Country]
□	Free trade leads to better products becoming available in [Country]
□	In general, [Country] should follow the decisions of international organizations to which it belongs, even if the government does not agree with them
□	International organizations are taking away too much power from the [Country nationality] government
□	Increased exposure to foreign films, music, and books is damaging our national and local cultures
□	A benefit of the Internet is that it makes information available to more and more people worldwide
□	It is impossible for people who do not share [Country] customs and traditions to become fully [Country's nationality]
□	Ethnic minorities should be given government assistance to preserve their customs and traditions
□	Immigrants increase crime rates
□	Immigrants are generally good for [Country] economy
□	Immigrants take jobs away from people who were born in [Country]
□	Immigrants improve [Country nationality] society by bringing in New ideas and cultures
□	Government spends too much money assisting immigrants
□	Children born in [Country] of parents who are not citizens should have the right to become [Country nationality] citizens
□	Children born abroad should have the right to become [Country nationality] citizens if at least one of their parents is a [Country nationality] citizen.
□	Legal immigrants to [Country] who are not citizens should have the same rights as [Country nationality] citizens
□	[Country] should take stronger measures to exclude illegal immigrants

Table A.5.
Question wording 2004. Citizenship module

□	People like me don't have any say about what the government does
□	I don't think the government cares much what people like me think
□	I feel I have a pretty good understanding of the important political issues facing [Country]
□	I think most people in [Country] are better informed about politics and government than I am
□	Most of the time we can trust people in government to do what is right
□	Most politicians are in politics only for what they can get out of it personally
□	Political parties encourage people to become active in politics
□	Political parties do not give voters real policy choices
□	Referendums are a good way to decide important political questions

Table A.6.
Reliability and Item Total Statistics for 2000. All items

Cronbach's Alpha = 0.774

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v4 Solve economic problems: Priv enterprise	77.99	136.649	.172	.127	.774
v5 Responsib gov: reduce income difference	78.11	133.994	.258	.249	.770
v46 International agreements should be made	78.95	142.558	-.037	.143	.780
v47 Poorer countries to make less effort	77.68	131.195	.337	.170	.765
v48 Economic progress will slow down unless	77.94	134.422	.273	.156	.769
v8 Science: believe too often in	77.89	135.227	.238	.145	.771
v9 Science: more harm than good	77.18	130.407	.418	.312	.761
v10 Science: solve environmental problems	77.38	130.839	.393	.216	.763
v11 Worry: about future environment	77.45	128.729	.449	.328	.759
v12 Environment: modern life harms the	77.84	134.502	.255	.252	.770
v13 Worry: progress harming environment	77.47	129.704	.438	.311	.760
v14 Environment: protect by economic growth	78.05	131.672	.353	.206	.765
v15 Animals: medical testing if save lives	78.20	136.010	.189	.120	.773
v16 Economic growth: harms the environment	77.50	131.840	.362	.290	.764
v17 Earth cannot continue support pres. rate	78.15	139.265	.073	.088	.779
v22 To do about environment: too difficult	77.35	127.723	.487	.330	.757
v23 Do what is right costs money takes time	78.04	139.918	.071	.111	.778
v4 Solve economic problems: Priv enterprise	77.99	136.649	.172	.127	.774
v24 More important than protect environment	77.50	132.783	.326	.254	.766
v25 No point unless others do the same	77.44	129.567	.395	.291	.762
v26 Many about environment exaggerated	77.41	131.045	.371	.350	.764
v65 Government should redistribute income	77.92	133.176	.270	.307	.769

Table A.6 (continued)
Reliability and Item Total Statistics for 2000. All items

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v66 People can do little to change lives	77.12	127.706	.487	.306	.757
v67 People challenge authority too often	77.53	132.391	.322	.163	.766
v68 People with money should be left enjoy	78.07	139.105	.088	.142	.778
v69 People follow conscience even break law	77.80	135.627	.200	.082	.773
v70 Private enterprise needs control	78.34	136.006	.226	.195	.771
v71 All societies have inequalities	77.45	132.107	.352	.228	.765
v72 Taking everything into account, better	77.54	136.039	.207	.141	.772

Table A.7.
Reliability and Item Total Statistics for 2000. Low-Correlated Items

Cronbach's Alpha = 0.297

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v4 Solve economic problems: Priv enterprise	22.99	13.478	.138	.066	.259
v5 Responsib gov: reduce income difference	23.16	13.956	.065	.070	.298
v8 Science: believe too often in	22.93	13.602	.140	.050	.259
v12 Environment: modern life harms the	22.89	13.543	.133	.081	.262
v15 Animals: medical testing if save lives	23.21	13.895	.075	.043	.292
v17 Earth cannot continue support pres. rate	23.18	13.783	.102	.044	.278
v23 Do what is right costs money takes time	23.06	13.911	.146	.044	.259
v46 International agreements should be made	24.00	14.694	.115	.040	.275
v68 People with money should be left enjoy	23.10	14.014	.093	.080	.282
v72 Taking everything into account, better	22.52	14.173	.070	.062	.293

Table A.8.
Reliability and Item Total Statistics for 2002. All items

Cronbach's Alpha = 0.553

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v4 Workg mom: warm relation child ok	59.91	65.031	.039	.129	.559
v5 Workg mom: pre school child suffers	59.46	60.377	.279	.277	.525
v6 Workg woman: family life suffers	59.23	60.408	.268	.271	.526
v7 What women really want is home & kids	59.60	60.130	.303	.326	.521
v8 Household satisfies as much as paid job	59.64	62.383	.187	.205	.539
v9 Work is best for womens independence	59.82	62.079	.227	.144	.534
v10 Both should contribute to hh income	60.34	63.392	.197	.173	.539
v11 Mens job is work, womens job household	59.46	60.301	.233	.413	.531
v12 Men should do larger share of hh work	59.45	62.998	.146	.388	.544
v13 Men should do larger share of childcare	59.39	63.010	.146	.432	.544
v18 Marriage: married people gen. happier	59.52	61.814	.200	.257	.536
v19 Bad marriage better than no marriage	58.47	61.806	.199	.214	.537
v20 Marriage better, if people want kids	59.73	62.538	.158	.284	.543
v21 Single parent can raise child as well	59.57	63.339	.118	.101	.549
v22 Couple livg together without marriage	59.10	65.819	-.030	.511	.573
v23 Couple live together bef. get married	59.35	64.427	.040	.425	.561
v24 Divorce best solution w marr. problems	59.33	62.054	.155	.193	.543
v25 Children: watching up is greatest joy	60.60	64.008	.213	.161	.539
v26 People without kids: lead empty lives	59.36	58.994	.316	.182	.517
v27 Workg women shld: paid maternity leave	60.62	63.816	.214	.293	.539
v28 Workg parents shld: financial benefits	60.25	62.026	.244	.283	.532
v44 So many things to do at home	59.26	62.683	.139	.116	.546

Table A.8 (continued)
Reliability and Item Total Statistics for 2002. All items

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v45 My life at home is rarely stressful	59.32	65.294	.020	.039	.562
v46 So many things to do at work	59.30	65.136	.026	.016	.561

Table A.9.
Reliability and Item Total Statistics for 2002. Low-Correlated Items

Cronbach's Alpha = 0.218

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v4 Workg mom: warm relation child ok	23.44	13.929	.090	.102	.192
v6 Workg woman: family life suffers	22.77	14.501	-.008	.070	.248
v9 Work is best for women independence	23.38	13.273	.198	.065	.136
v12 Men should do larger share of hh work	23.01	13.718	.107	.084	.183
v18 Marriage: married people gen. happier	23.03	14.698	-.022	.059	.255
v21 Single parent can raise child as well	23.09	14.000	.057	.054	.210
v24 Divorce best solution w marr. problems	22.91	12.879	.145	.089	.155
v27 Workg women shld: paid maternity leave	24.16	14.463	.151	.039	.175
v45 My life at home is rarely stressful	22.88	14.220	.047	.008	.215
v46 So many things to do at work	22.85	14.344	.028	.004	.226

Table A.10.
Reliability and Item Total Statistics for 2003. All items

Cronbach's Alpha = 0.69

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v19 I would rather be a citizen of[Country] than of any other country in the world	72.89	96.715	.328	.283	.675
v20 There are some things about[Country] today that make me feel ashamed of[Country]	72.32	102.034	.049	.181	.697
v21 The world would be a better place if people from other countries were more like the[Country Nationality]	71.87	95.210	.376	.354	.671
v22 Generally speaking,[Country] is a better country than most other countries	72.32	97.264	.281	.353	.679
v23 People should support their country even if the country is in the wrong.	71.76	94.467	.365	.220	.671
v24 When my country does well in international sports, it makes me proud to be[Country Nationality]	73.02	97.745	.331	.195	.677
v25 I am often less proud of[Country] than I would like to be	72.15	99.880	.152	.211	.689
v36 [Country] should limit the import of foreign products in order to protect its national economy.	72.28	94.022	.400	.302	.668
v37 For certain problems, like environment pollution, international bodies should have the right to enforce solutions	72.70	100.288	.165	.099	.687

Table A.10. (continued)
Reliability and Item Total Statistics for 2003. All items

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v38 [Country] should follow its own interests, even if this leads to conflicts with other nations	72.20	95.623	.347	.186	.673
v39 Foreigners should not be allowed to buy land in[Country]	71.96	93.858	.375	.286	.670
v40 [Country?s] television should give preference to[Country] films and programmes	72.10	92.689	.449	.343	.664
v41 Large international companies are doing more and more damage to local businesses in[Country].	72.42	96.144	.344	.244	.674
v42 Free trade leads to better products becoming available in[Country].	72.51	101.096	.136	.135	.689
v43 In general,[Country] should follow the decisions of international organizations to which it belongs, even if the govern	72.01	101.214	.111	.130	.691
v44 International organizations are taking away too much power from the[Country Nationality] government.	72.16	96.457	.349	.236	.674
v45 Increased exposure to foreign films, music, and books is damaging our national and local cultures.	71.80	92.921	.437	.321	.665
v46 A benefit of the Internet is that it makes information available to more and more people worldwide.	73.09	102.938	.054	.093	.693

Table A.10. (continued)
Reliability and Item Total Statistics for 2003. All items

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v47 It is impossible for people who do not share [Country?s] customs and traditions to become fully [Country?s nationality]	72.20	96.376	.284	.204	.678
v48 Ethnic minorities should be given government assistance to preserve their customs and traditions	72.03	100.244	.115	.238	.692
v50 Immigrants increase crime rates	72.19	96.982	.267	.363	.680
v51 Immigrants are generally good for [Country?s] economy	71.96	105.393	-.092	.331	.705
v52 Immigrants take jobs away from people who were born in[Country]	71.89	94.682	.367	.384	.671
v53 Immigrants improve[Country Nationality] society by bringing in new ideas and cultures	72.14	106.193	-.129	.333	.708
v54 Government spends too much money assisting immigrants	72.14	98.020	.225	.406	.683
v59 Children born in[Country] of parents who are not citizens should have the right to become[Country Nationality] citizens	72.68	102.601	.035	.239	.697
v60 Children born abroad should have the right to become[Country Nationality] citizens if at least one of their parents is	72.81	102.177	.084	.203	.692
v61 Legal immigrants to[Country] who are not citizens should have the same rights as[Country Nationality] citizens.	72.08	104.077	-.041	.222	.705
v62 [Country] should take stronger measures to exclude illegal immigrants?	72.89	98.960	.206	.200	.684

Table A.11.
Reliability and Item Total Statistics for 2003. Low-Correlated Items

Cronbach's Alpha = 0.008

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
v20 There are some things about[Country] today that make me feel ashamed of[Country]	23.95	11.999	-.018	.022	.022
v22 Generally speaking,[Country] is a better country than most other countries	23.94	11.442	.074	.066	-.045[a]
v47 It is impossible for people who do not share [Country?s] customs and traditions to become fully [Country?s nationality]	23.85	11.028	.086	.151	-.061[a]
v48 Ethnic minorities should be given government assistance to preserve their customs and traditions	23.61	12.170	-.056	.106	.053
v50 Immigrants increase crime rates	23.84	11.332	.060	.319	-.038[a]
v51 Immigrants are generally good for [Country?s] economy	23.55	12.649	-.074	.327	.057
v53 Immigrants improve[Country Nationality] society by bringing in new ideas and cultures	23.74	12.643	-.077	.319	.061
v54 Government spends too much money assisting immigrants	23.80	11.761	.009	.349	.002
v61 Legal immigrants to[Country] who are not citizens should have the same rights as[Country Nationality] citizens.	23.67	12.330	-.071	.162	.064
v62 [Country] should take stronger measures to exclude illegal immigrants?	24.54	11.286	.106	.178	-.067[a]
v50 Immigrants increase crime rates	84.93	14255.190	.429	.355	.697
v54 Government spends too much money assisting immigrants	78.21	13425.497	.403	.354	.709

a. The value is negative due to a negative average covariance among items.

Table A.12.
Mean proportion of extreme response style per country and year, ordered from highest to lowest

1999		2000		2002		2003		2004	
ERS	Country	ERS	Country	ERS	Country	ERS	Country	ERS	Country
.6159	Bulgaria	.4012	Bulgaria	.7823	Brazil	.4359	Japan	.4835	Brazil
.5575	Russia	.3592	Russia	.5668	Denmark	.4338	Bulgaria	.3550	France
.5454	Portugal	.3533	Denmark	.5547	Japan	.4227	Denmark	.3491	Bulgaria
.4688	Slovakia	.3464	Japan	.4310	France	.4200	Russia	.3470	Japan
.4287	Denmark	.3395	Israel - Arabic	.4285	Slovakia	.3552	France	.3135	Israel-other
.4048	Israel-other	.2984	Czech republic	.3987	United States	.3428	Israel	.3016	Mexico
.4041	Japan	.2696	Switz - German	.3686	Hungary	.3387	Austria	.2946	Denmark
.3830	Czech republic	.2480	Portugal	.3584	Austria	.3112	South Africa	.2818	Slovenia
.3606	France	.2480	Switz - French	.3582	Israel-other	.2936	Hungary	.2648	Austria
.3478	Hungary	.2454	Germany	.3514	Czech republic	.2895	Israel Arabic	.2578	Czech republic
.3215	Latvia	.2427	Austria	.3430	Portugal	.2809	Czech Republic	.2563	Uruguay
.3052	Israel-Hebrew	.2339	Mexico	.3418	Israel-Hebrew	.2610	Philippines	.2530	Israel-Jewish
.2853	Israel-Arabic	.2228	Canada	.3404	Switz - French	.2564	Uruguay	.2492	Russia
.2837	Austria	.2186	Spain	.3369	Mexico	.2481	Venezuela	.2449	Germany
.2753	Philippines	.2080	Finland	.3209	Bulgaria	.2479	Portugal	.2427	Poland
.2705	Slovenia	.2033	Slovenia	.3125	Germany	.2430	United States	.2289	Slovakia
.2570	Poland	.2012	Sweden	.2999	Sweden	.2362	R Chile	.2283	Portugal
.2360	Chile	.1880	Israel - Jewish	.2958	Israel-Arabic	.2356	Australia	.2254	Venezuela
.2308	Spain	.1741	New Zealand	.2845	Finland	.2342	Canada	.2159	South Korea
.2231	Germany	.1715	Latvia	.2609	Russia	.2274	Slovakia	.2145	Flanders
.2147	Canada	.1654	Switz - Italian	.2599	Philippines	.2244	New Zealand	.2145	Latvia
.2129	United States	.1549	United States	.2587	Slovenia	.2238	Slovenia	.2137	Israel-Arabic
.2125	Sweden	.1546	Norway	.2558	Northern Ireland	.2218	Sweden	.2135	South Africa
.2073	New Zealand	.1509	Philippines	.2459	Chile	.2218	Finland	.2110	Hungary
.1906	Norway	.1419	North Ireland	.2432	Spain	.2202	Germany	.1919	Ireland
.1676	North Ireland	.1367	Great Britain	.2311	Norway	.2072	South Korea	.1822	Switz - French
.1625	Cyprus	.1206	Netherlands	.2224	Latvia	.2056	Latvia	.1774	Spain

Table A.12. (continued)
Mean proportion of extreme response style per country and year, ordered from highest to lowest

1999		2000		2002		2003		2004	
ERS	Country	ERS	Country	ERS	Country	ERS	Country	ERS	Country
.1541	Great Britain	.1191	Chile	.2218	Australia	.2037	Norway	.1735	Finland
.1440	Australia	.0893	Ireland	.2179	Flanders	.2015	Poland	.1641	Australia
				.2089	Great Britain	.1992	Great Britain	.1618	Canada
				.2025	Switz - Italian	.1903	Netherlands	.1594	United States
				.1995	New Zealand	.1696	Ireland	.1576	Philippines
				.1949	Netherlands	.1633	Taiwan	.1570	Norway
				.1947	Poland	.1490	Switzerland	.1443	Sweden
				.1776	Switz - German	.1317	Spain	.1426	Great Britain
				.1760	Ireland			.1343	New Zealand
				.1267	Taiwan			.1323	Chile
				.0629	Cyprus			.1295	Netherlands
								.1142	Taiwan
								.0925	Switz - German
								.0814	Switz - Italian
								.0540	Cyprus

Table A.13.
Mean proportion of response style indicator per country and year, ordered from highest to lowest

1999		2000		2002		2003		2004	
RSI	Country	RSI	Country	RSI	Country	RSI	Country	RSI	Country
.3790	Bulgaria	.3015	Bulgaria	.4343	Brazil	.3197	Bulgaria	.3415	Brazil
.3650	Portugal	.2999	Denmark	.3635	Denmark	.3174	Denmark	.3032	Venezuela
.3638	Russia	.2998	Russia	.3423	Japan	.3157	Russia	.2883	France
.3365	Israel-other	.2888	Mexico	.3169	France	.3096	Venezuela	.2832	Bulgaria
.3315	Slovakia	.2819	Israel - Arabic	.3115	Portugal	.3009	Japan	.2767	Mexico
.3220	Denmark	.2761	Switz - German	.3104	Mexico	.2981	South Africa	.2736	Denmark
.3041	Czech republic	.2733	Czech republic	.3090	Slovakia	.2905	Israel	.2732	Israel-other
.2982	France	.2692	Germany	.3013	Switz - French	.2884	Austria	.2731	Uruguay
.2975	Latvia	.2690	Portugal	.3004	United States	.2857	France	.2703	Austria
.2932	Israel-Hebrew	.2670	Austria	.2997	Austria	.2721	Uruguay	.2675	Poland
.2908	Hungary	.2670	Spain	.2963	Germany	.2698	Portugal	.2668	Japan
.2906	Japan	.2649	Japan	.2956	Israel-Hebrew	.2646	Chile	.2645	Germany
.2821	Chile	.2624	Switz- French	.2930	Hungary	.2639	Hungary	.2644	Ireland
.2792	Austria	.2619	Canada	.2910	Israel-other	.2581	Czech republic	.2599	Slovenia
.2791	Slovenia	.2592	Slovenia	.2895	Czech republic	.2577	Canada	.2597	Russia
.2721	Philippines	.2507	Finland	.2888	Bulgaria	.2573	Philippines	.2577	Portugal
.2713	Israel-Arabic	.2470	New Zealand	.2877	Spain	.2570	Ireland	.2555	South Africa
.2707	Poland	.2440	Israel - Jewish	.2859	Chile	.2569	Australia	.2536	Israel-Hebrew
.2694	Spain	.2394	Latvia	.2805	Philippines	.2556	Slovenia	.2512	Czech republic
.2623	Germany	.2393	Sweden	.2785	Finland	.2554	United States	.2422	Philippines
.2555	Canada	.2353	Chile	.2763	Slovenia	.2521	Germany	.2393	Slovakia
.2553	New Zealand	.2322	Norway	.2738	Russia	.2516	New Zealand	.2383	South Korea
.2453	Norway	.2316	Switz - Italian	.2736	Northern Ireland	.2475	Switzerland	.2366	Latvia
.2432	United States	.2306	Ireland	.2700	Sweden	.2469	Taiwan	.2356	Spain
.2369	Northern Ireland	.2257	Great Britain	.2622	Switz - German	.2458	Poland	.2323	United States
.2347	Cyprus	.2254	Philippines	.2615	Ireland	.2416	Norway	.2301	Canada
.2340	Sweden	.2254	United States	.2599	Norway	.2413	Finland	.2291	Flanders

Table A.13 (continued)
Mean proportion of response style indicator per country and year, ordered from highest to lowest

1999		2000		2002		2003		2004	
RSI	Country	RSI	Country	RSI	Country	RSI	Country	RSI	Country
.2331	Great Britain	.2177	Netherlands	.2596	Poland	.2406	South Korea	.2291	Australia
.2316	Australia	.2171	North Ireland	.2591	Switz - Italian	.2394	Netherlands	.2285	Switz - French
				.2576	Flanders	.2392	Latvia	.2280	Israel-Arabic
				.2558	Latvia	.2376	Great Britain	.2276	Finland
				.2546	Israel-Arabic	.2357	Sweden	.2261	Switz - Italian
				.2534	Taiwan	.2354	Slovakia	.2248	Hungary
				.2533	Great Britain	.2181	Spain	.2224	Chile
				.2509	Australia			.2189	New Zealand
				.2472	New Zealand			.2180	Norway
				.2429	Netherlands			.2178	Netherlands
				.2187	Cyprus			.2166	Great Britain
								.2141	Switz - German
								.2093	Taiwan
								.1965	Sweden
								.1879	Cyprus

Table A.14.
Mean proportion of middle response style per country and year, from highest to lowest

1999		2000		2002		2003		2004	
MRS	Country	MRS	Country	MRS	Country	MRS	Country	MRS	Country
.2767	Sweden	.2866	Japan	.2776	Israel-Arabic	.2859	Slovakia	.3582	Sweden
.2416	Japan	.2737	North Ireland	.2227	Netherlands	.2789	Sweden	.3117	Hungary
.2400	United States	.2533	United States	.2191	Sweden	.2593	Spain	.3025	Cyprus
.2236	Cyprus	.2499	Netherlands	.2182	Australia	.2565	Finland	.3015	Israel-Arabic
.2217	Great Britain	.2492	Philippines	.2096	New Zealand	.2490	Latvia	.2981	Flanders
.2202	Northern Ireland	.2442	Sweden	.1984	Latvia	.2487	Great Britain	.2851	Norway
.2175	Australia	.2388	Switz - Italian	.1972	United States	.2485	Czech republic	.2798	Japan
.2092	Norway	.2340	Great Britain	.1970	Great Britain	.2474	Israel -Hebrew	.2769	Taiwan
.2000	Israel-Arabic	.2258	Norway	.1966	Hungary	.2449	South Korea	.2764	Great Britain
.1927	Canada	.2140	Latvia	.1941	Israel-other	.2379	Hungary	.2718	Slovakia
.1868	Philippines	.2118	Israel - arab	.1932	Czech republic	.2371	Norway	.2682	Switz - French
.1862	New Zealand	.2118	Israel - Jewish	.1924	Slovakia	.2326	Netherlands	.2680	Latvia
.1846	Hungary	.2053	Czech republic	.1913	Norway	.2322	Japan	.2633	Finland
.1744	Poland	.2051	Finland	.1880	Cyprus	.2318	Philippines	.2625	South Korea
.1737	Germany	.1983	Switz - French	.1879	Flanders	.2212	United States	.2587	New Zealand
.1677	France	.1952	Bulgaria	.1849	Japan	.2182	Poland	.2583	Netherlands
.1668	Austria	.1860	New Zealand	.1697	Finland	.2181	New Zealand	.2532	Czech republic
.1665	Czech republic	.1779	Chile	.1659	Switz - Italian	.2124	France	.2478	Australia
.1541	Slovenia	.1750	Canada	.1650	Russia	.2117	Germany	.2426	Chile
.1532	Spain	.1748	Austria	.1649	Bulgaria	.2080	Australia	.2421	Slovenia
.1429	Slovakia	.1720	Portugal	.1629	France	.2036	Canada	.2416	Canada
.1405	Denmark	.1685	Germany	.1601	Northern Ireland	.2014	Slovenia	.2384	Israel-Jewish
.1323	Israel-Hebrew	.1670	Ireland	.1596	Israel-Hebrew	.1852	Austria	.2363	Switz - German
.1316	Latvia	.1665	Slovenia	.1583	Austria	.1809	Israel	.2350	Spain
.1075	Chile	.1650	Switz - German	.1562	Poland	.1779	R Chile	.2303	United States
.1024	Russia	.1599	Russia	.1537	Slovenia	.1757	Taiwan	.2207	Israel-other
.0999	Bulgaria	.1537	Denmark	.1377	Philippines	.1688	Portugal	.2163	Bulgaria

Table A.14. (continued)
Mean proportion of middle response style per country and year, from highest to lowest

1999		2000		2002		2003		2004	
MRS	Country	MRS	Country	MRS	Country	MRS	Country	MRS	Country
.0853	Portugal	.1507	Spain	.1352	Switz - French	.1678	Uruguay	.2104	Russia
.0589	Israel-other	.0789	Mexico	.1300	Ireland	.1592	Switzerland	.2018	France
				.1290	Switz - German	.1573	Russia	.2000	Denmark
				.1285	Germany	.1552	Bulgaria	.1977	Portugal
				.1130	Taiwan	.1531	Denmark	.1946	Mexico
				.1129	Denmark	.1418	Ireland	.1913	South Africa
				.1019	Chile	.1189	South Africa	.1889	Philippines
				.0971	Portugal	.0095	Venezuela	.1868	Germany
				.0955	Mexico			.1834	Austria
				.0924	Spain			.1772	Switz - Italian
				.0453	Brazil			.1729	Poland
								.1639	Uruguay
								.1345	Ireland
								.1176	Brazil
								.0127	Venezuela

Table A.15.
Mean proportion of acquiescence per country and year, from highest to lowest

1999		2000		2002		2003		2004	
ARS	Country	ARS	Country	ARS	Country	ARS	Country	ARS	Country
.7201	Israel-other	.4838	Israel - arab	.4485	Brazil	.5058	Russia	.4032	Poland
.4360	Poland	.4677	Portugal	.4259	Chile	.4831	Bulgaria	.3554	Bulgaria
.4316	Chile	.2951	Chile	.3943	Bulgaria	.4641	R Chile	.3527	Russia
.4186	Germany	.2820	Philippines	.3878	Portugal	.4531	Portugal	.3262	Israel-other
.4131	Israel-Arabic	.2673	Mexico	.3781	Israel-other	.4420	South Africa	.3212	Portugal
.4056	Portugal	.2470	Bulgaria	.3688	Russia	.4363	Israel	.3186	Cyprus
.3979	Spain	.2263	Slovenia	.3458	Hungary	.4163	Poland	.3079	South Africa
.3653	Austria	.1939	Spain	.3425	Mexico	.4026	Hungary	.3002	Latvia
.3636	Philippines	.1619	Russia	.3343	Poland	.3714	Uruguay	.2943	Flanders
.3269	Slovenia	.1423	Israel - Jewish	.3268	Czech republic	.3628	Taiwan	.2903	South Korea
.3164	Northern Ireland	.1224	Latvia	.3244	Taiwan	.3555	Czech republic	.2869	Ireland
.2846	Japan	.1157	Czech republic	.3128	Slovenia	.3455	South Korea	.2665	Brazil
.2714	Australia	.1097	North Ireland	.3073	Austria	.3290	Venezuela	.2543	Slovenia
.2703	France	.0747	Austria	.3065	Switz - Italian	.3199	Israel	.2361	Israel-Jewish
.2691	Sweden	.0690	Netherlands	.2909	Slovakia	.3185	Spain	.2356	Spain
.2537	Great Britain	.0682	Switz - Italian	.2908	Latvia	.3166	Philippines	.2333	Canada
.2341	Russia	.0600	Germany	.2841	Philippines	.3053	Slovakia	.2313	Switz - Italian
.2323	Israel-Hebrew	.0546	Great Britain	.2661	Spain	.3048	Japan	.2304	Austria
.2312	United States	.0489	Ireland	.2620	France	.2980	Slovenia	.2217	Israel-Arabic
.2279	Czech republic	.0371	Denmark	.2558	Germany	.2845	Australia	.2192	Australia
.2258	Bulgaria	.0314	Switz - German	.2546	Switz - French	.2711	New Zealand	.2173	Switz - French
.2177	Denmark	.0311	Sweden	.2377	Israel-Hebrew	.2660	Ireland	.2157	Hungary
.2070	Slovakia	.0194	Japan	.2361	Switz - German	.2659	Latvia	.2132	Czech republic
.1920	Norway	.0185	United States	.2150	Denmark	.2656	Great Britain	.2122	Finland
.1819	Cyprus	.0093	Switz - French	.2101	Finland	.2642	Austria	.2120	Germany
.1778	New Zealand	-.0203	Norway	.2045	United States	.2548	Canada	.2070	Slovakia
.1687	Latvia	-.0271	Canada	.1964	Cyprus	.2484	Denmark	.1952	Sweden
.1671	Canada	-.0447	Finland	.1921	Northern Ireland	.2444	United States	.1947	Switz - German
.1143	Hungary	-.0833	New Zealand	.1904	Flanders	.2187	Germany	.1914	Philippines

Table A.15 (continued)
Mean proportion of acquiescence per country and year, from highest to lowest

1999		2000		2002		2003		2004	
ARS	Country	ARS	Country	ARS	Country	ARS	Country	ARS	Country
				.1852	Sweden	.2174	France	.1776	United States
				.1804	Japan	.1990	Sweden	.1774	New Zealand
				.1764	Israel-Arabic	.1960	Finland	.1719	Great Britain
				.1716	Great Britain	.1822	Norway	.1573	Taiwan
				.1543	Ireland	.1794	Switzerland	.1496	Uruguay
				.1433	Australia	.1585	Netherlands	.1475	Chile
				.1340	Norway			.1128	Denmark
				.0834	Netherlands			.0769	Norway
				.0781	New Zealand			.0601	Venezuela
								.0521	Japan
								.0506	Mexico
								.0418	France
								.0389	Netherlands

Table A.16.
Hofstede's cultural dimensions' scores.

	Power Distance	Individualism /Collectivism	Uncertainty Avoidance
Australia	36	90	51
Germany	35	67	65
Great Britain	35	89	35
Northern Ireland	NA	NA	NA
United States	40	91	46
Austria	11	55	70
Hungary	NA	NA	NA
Italy	50	76	75
Ireland	28	70	35
Netherlands	38	80	53
Norway	31	69	50
Sweden	31	71	29
Czech Republic	NA	NA	NA
Slovenia	NA	NA	NA
Poland	NA	NA	NA
Bulgaria	NA	NA	NA
Russia	NA	NA	NA
New Zealand	22	79	49
Canada	39	80	48
R Philippines	94	32	44
Israel-Hebrew	13	54	81
Israel-Arabic	80	38	68
Japan	54	46	92
Spain	57	51	86
Latvia	NA	NA	NA
Slovakia	NA	NA	NA
France	68	71	86
Cyprus	NA	NA	NA
Portugal	63	27	104
Denmark	18	74	23
R Chile	63	23	86
Switzerland	34	68	58
Flanders	65	75	94
Brazil	69	38	76
South Africa	49	65	49
Finland	33	63	59
Mexico	81	30	82
Taiwan	58	17	69
Venezuela	81	12	76
South Korea	60	18	85
Uruguay	61	36	100
Dominican Republic	NA	NA	NA
Croatia	76	27	88

NA = Data not available

References

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research, 20*, 139-181.

American Association for Public Opinion Research. (2008). Standard definitions: Final dispositions of case codes and outcome rates for surveys. 5th edition. Lenexa, Kansas: AAPOR.

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme-response style. Improving meaning of translated and culturally adapted rating scales. *Educational and Psychological Measurement, 66*, 374-392.

Bachman, J. G., & O'Malley, P. M. (1984a). Black-White differences in self-esteem: Are they affected by response styles? *The American Journal of Sociology, 90*, 624-639.

Bachman, J. G., & O'Malley, P. M. (1984b). Yea-saying, nay-saying, and going to extremes: Black-White differences in response styles. *Public Opinion Quarterly, 48*, 491-509.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research, 38*, 143-156.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). Response biases in marketing research. In R. Grover & M. Vriens (Eds.), *The handbook of marketing research: Uses, misuses, and future advances* (Vol. 95-109). Thousand Oaks, CA: SAGE.

Berg, I. A., & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement, 13*, 164-169.

Block, J. (1965). *The challenge of response sets*. New York: Appleton-Century-Crofts.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions. The definitive guide to questionnaire design - For market research, political polls, and social and health questionnaires*. San Francisco, CA: Jossey-Bass.

Braun, M. (2003). Errors in comparative survey research: An overview. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-Cultural Survey Methods*. New York: Wiley.

Braun, M., & Uher, R. (2003). The ISSP and its approach to background variables. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in cross-national comparison. A European working book for demographic and socio-economic variables*. New York, NY: Kluwer Academic/Plenum Publishers.

Brengelmann, J. C. (1959). Differences in questionnaire responses between English and German nationals. *Acta Psychologica, 16*, 339-355.

Brengelmann, J. C. (1960). A note on questionnaire rigidity and extreme response set. *Journal of Mental Science, 106*, 187-192.

Braun, M. and Scott, J. (1998), Multidimensional scaling and equivalence: Is *having a job* the same as *working*? in J. A. Harkness (Ed.), *ZUMA-Nachrichten Spezial No. 3. Cross-Cultural Survey Equivalence,* Mannnheim: ZUMA.

Braun, M. and Harkness, J. A. (2005), Text and context: Challenges to comparability in survey questions, in J. H. P. Hoffmeyer-Zlotnik and J. A. Harkness (Eds.), *ZUMA-Nachrichten Spezial Band 11, Methodological Aspects in Cross-National Research*, pp. 95-108, Mannheim, Germany: ZUMA. Retrieved from http://www.gesis.org/fileadmin/upload /forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial11.pdf

Broen, W. E. J., & Wirt, R. D. (1958). Varieties of response sets. *Journal of Consulting Psychology, 22*, 237-240.

Campbell, A., Converse, P. E., Miller, W. E., & Strokes, D. E. (1960). *The American voter*. New York: Wiley.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological Methodology, 12*, 389-437.

Carr, L. G. (1971). The Srole items and acquiescence. *American Sociological Review, 36*, 287-293.

Chen, C., Lee, S.-y., & Stevenson, H. W. (1995). Response style and cross-cultural comparison of rating scales among East Asian and North American students. *Psychological Science, 6*, 170-175.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*.

Chun, K.-T., Campbell, J. B., & Yoo, J. H. (1974). Extreme response style in cross-cultural research. *Journal of Cross-Cultural Psychology, 5*, 465-480.

Clarke, I. (2000a). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality, 15*, 137-152.

Clarke, I. (2000b). Global marketing research: Is extreme response style influencing your results? *Journal of International Consumer Marketing, 12*, 91-111.

Clarke, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review, 18*.

Clarke, V. A., Ruffin, C. L., Hill, D. J., & Beamen, A. L. (1992). Ratings of orally presented verbal expressions of probability by a heterogeneous sample. *Journal of Applied Social Psychology, 22*, 638-656.

Cloud, J., & Vaughan, G. M. (1970). Using balanced scales to control acquiescence. *Sociometry, 33*, 193-202.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal & Social Psychology, 60*, 151-174.

Couper, M. P., & De Leeuw, E. (2003). Nonresponse in cross-cultural and cross-national surveys. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural survey methods*. New York: Wiley.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology, 33*.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475-494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3-31.

CSDI. (2009, March 02, 2009). Cross-cultural survey guidelines: Data collection Retrieved March 10, 2009, 2009, from http://ccsg.isr.umich.edu/datacoll.cfm

Culpepper, R. A., Zhao, L., & Lowery, C. (2002). Survey response bias among Chinese managers *Academy of Management Proceedings* (pp. J1-J6).

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research, 39*, 1-18.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys*. Hoboken, NJ: Wiley.

Federico, C. M. (2004). Predicting attitude extremity: The interactive effects of schema development and the need to evaluate and their mediation of evaluative integration. *Personality and Social Psychology Bulletin, 30*, 1281-1294.

Fisher, E. M. (1974). *Change in anomie in Detroit from the 1950s to 1971. Ph.D. dissertation* Ann Arbor, MI: University of Michigan.

Fowler, F. J. (1995). *Improving survey questions. Design and evaluation*. Thousand Oaks, CA: Sage.

Fowler, F. J., & Cosenza, C. (2008). Writing effective questions. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology*. New York: Erlbaum.

Friedman, H. H., & Amoo, T. (1999). Rating the rating scales. *The Journal of Marketing Management, 9*, 114-123.

Gibbons, J. L., Zellner, J. A., & Rudek, D. J. (1999). Effects of language and meaningfulness on the use of extreme response style by Spanish-English bilinguals. *Cross-Cultural Research, 33*, 369-381.

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research, 29*, 176-188.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley.

Haberstroh, S., Oyserman, D., Schwarz, N., Kuhnen, U., & Ji, L.-J. (2002). Is the interdependent self more sensitive to question context than the independent self? Self-construal and the observation of conversational norms. . *Journal of Experimental Social Psychology, 38*, 323-329.

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences, 44*, 932-942.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin, 69*, 192-203.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural survey methods*. New York: Wiley.

Harkness, J. A. (2005). Report to the ISSP General Assembly on behalf of the translation group. Mannheim: ZUMA.

Harkness, J. A. (2008). Comparative survey research: goal and challenges. In E. De Leeuw, J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 56-77). Hyattsville, VA: Lawrence Erlbaum.

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

Harkness, J. A., & McKinney, T. (2009). *Communication and comparative survey research*. Paper presented at the International Workshop in Comparative Survey Design and Implementation, Ann Arbor, MI.

Harkness, J. A., Mohler, P. P., Smith, T. W., & Davis, J. A. (1997). Final Report on the Project 'Research into the Methodology of Intercultural Surveys' (MINTS) for the German-American Transcoop Programme. Mannheim, Chicago: ZUMA, NORC.

Harkness, J. A., Mohler, P. P., & van de Vijver, F. J. R. (2003). Comparative research. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural survey methods*. New York: Wiley.

Harkness, J. A., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaire* (pp. 453-472). New York: Wiley.

Harkness, J. A., Schoebi, N., Joye, D., Mohler, P. P., & Behr, D. (2007). Oral translation in telephone surveys. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. De Leeuw, L. Japec, P. J. Lavrakas, M. W. Link & R. L. Sangster (Eds.), *Telephone survey methodology* (pp. 231-249). Hoboken, NJ: Wiley.

Harkness, J. A., Villar, A., Kruse, Y., Branden, L., Edwards, B., Steele, C., et al. (2009). *Using interpreters in telephone surveys.* Paper presented at the American Association for Public Opinion Research, Hollywood, FL.

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. . *International Journal of Cross-Cultural Management, 6*, 243-266.

Hofstede, G. (2001). *Culture's consequences. Comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Holbrook, A. L., Johnson, T. P., & Cho, Y. I. (2006). *Extreme response style: Style or substance*. Paper presented at the American Association for Public Opinion Research, Montreal, Canada.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (Eds.). (2004). *Culture, leadership, and organizations. The GLOBE study of 62 societies* Thousand Oaks, CA: Sage.

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly, 49*, 253-260.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*, 296-309.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*, 243-252.

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology, 70*, 172-194.

Javeline, D. (1999). Response effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly, 63*, 1-28.

Ji, L.-J., Schwarz, N., & Nisbett, R. E. (2000). Culture, autobiographical memory, and behavioral frequency reports: Measurement issues in cross-cultural studies. *Personality and Social Psychology Bulletin, 26*, 585-593.

Johnson, T. P., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles. Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*, 264-277.

Johnson, T. P., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R., & Lacey, L. (1997). Social cognition and responses to survey questions among culturally diverse populations. In L. E. Lyberg, P. P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 87-113). New York: Wiley.

Jones, E. L. (1963). The courtesy bias in South-East Asian surveys. *International Social Science Journal, 15*, 70-76.

Jowell, R. (1998). How comparative is comparative research? *American Behavioral Scientist, 42*, 168-177.

Kalgraff Skjåk, K. (in press). The International Social Survey Programme: Annual cross-national social surveys since 1985. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B.-E. Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts*. Hoboken, NJ: Wiley.

Kane, E., & Macaulay, L. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly, 57*, 1-28.

Knaüper, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly, 63*, 347-370.

Knowles, E. S., & Nathan, K. T. (1997). Acquiescent responding in self-reports: Cognitive style or social concern? *Journal of Research in Personality, 31*, 293-301.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Krosnick, J. A., & Berent, M. K. (1990). *The impact of verbal labeling of response alternatives and branching on attitude measurement reliability in surveys*. Paper presented at the American Association for Public Opinion Research Lancaster, Pennsylvania.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. E. Lyberg, P. P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141-164). New York: Wiley.

Kuz'min, V. B. (1981). A parametric approach to description of linguistic values of variables and hedges. *Fuzzy Sets and Systems, 6*, 27'41.

Landsberger, H. A., & Saavedra, A. (1967). Response set in developing countries. *Public Opinion Quarterly, 31*, 214-229.

Lenski, G. E., & Leggett, J. C. (1960). Caste, class, and deference in the research interview. *The American Journal of Sociology, 65*, 463-467.

Lentz, T. F. (1938). Acquiescence as a factor in the measurement of personality. *35*, 659.

Lewis, N. A., & Taylor, J. A. (1955). Anxiety and extreme response preferences. *Educational and Psychological Measurement, 15*, 111-116.

Light, C. S., Zax, M., & Gardiner, D. H. (1965). Relationship of age, sex and intelligence level to extreme response style. *Journal of Personality and Social Pscyhology, 2*, 907-909.

Lynn, P., Lyberg, L. E., & Japec, L. (2006). What's special about cross-national surveys? In J. A. Harkness (Ed.), *Conducting cross-national and cross-cultural surveys*. Mannheim, Germany: German Social Science Infrastructure Services.

Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics. The role of acculturation and education. *Journal of Cross-Cultural Psychology, 23*, 498-509.

Messick, S. (1968). Response sets. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (pp. 492-496). New York: Macmillan Company & The Free Press.

Messick, S. (1991). Psychology and methodology of response styles. In R. Snow & D. Wiley (Eds.), *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach* (pp. 161-200). Hillsdale, New Jersey: Lawrence Erlbaum

Messick, S., & Jackson, D. N. (1961). Acquiescence and the factorial interpretation of the MMPI. *Psychological Bulletin, 58*, 299-304.

Mohler, P. P., Hansen, S. E., Pennell, B.-E., Thomas, W., Wackerow, J., & Hubbard, F. (in press). A survey process quality perspective on documentation. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B.-E. Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multiregional and multicultural contexts*. Hoboken, NJ: Wiley.

Mohler, P. P., Smith, T. W., & Harkness, J. A. (1998). Respondents' ratings of expressions from response scales: A two-country, two-language investigation on equivalence and translation. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence. ZUMA-Nachrichten Spezial No. 3* (pp. 159-184). Mannheim, Germany: ZUMA.

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach: Sociodemographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality & Quantity, 37*, 277-302.

Moors, G. (2004). Facts and artifacts in the comparison of attitudes among ethnic minorities: A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review, 20*, 303-320.

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity, 42*, 779-794.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*, 58-88.

O'Neill, H. W. (1967). Response style influence in public opinion surveys. *Public Opinion Quarterly, 31*, 95-102.

Oldendick, R. W. (2008). Rating. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods* (1 ed., Vol. 2, pp. 691-692). Los Angeles, CA: Sage.

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin, 128*, 3-72.

Oyserman, D., & Lee, S. W. S. (2007). Culture: Effects of priming cultural syndromes on cognition and motivation. In R. M. Sorrentino, Susumu (Ed.), *Handbook of motivation and cognition across cultures*. New York, NY: Elsevier.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407-418.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/disapreferred turn shapes. In J. Atkinson & J. Heritage (Eds.), *Structure of Social Action* (pp. 57-101). Cambridge: Cambridge University Press.

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin, 63*, 129-156.

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior, 25*, 189-197.

Sacks, H. (1987). On the preferences for agreement and contiguity in sequences in conversation. In G. Button & J. R. E. Lee (Eds.), *Talk and social organization*. Clevedon, England: Multilingual Matters.

Sapin, M., Pollien, A., Joye, D., Leuenberger-Zanetta, S., & Schoebi, N. (2008). *Effect of different translation of answer scales.* Paper presented at the International Conference on Survey Methods in multinational, multiregional, and multicultural context (3MC), Berlin, Germany.

Saris, W. E., Krosnick, J. A., & Shaeffer, E. M. (2005). *Comparing questions with agree/disagree response options to questions with construct-specific response options*. Retrieved from http://communication.stanford.edu/faculty/krosnick/docs/Saris%20Paper%20-%20New%202005.pdf

Schaeffer, N. C. (1991). Conversation with a purpose - or conversation? Interaction in the standardized interview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 367-391). New York: John Wiley and Sons.

Scheuch, E. K. (1968). The cross-cultural use of sample surveys: Problems of comparability. In S. Rokkan (Ed.), *Comparative research across cultures and nations* (pp. 176-209). Paris: Mouton.

Scholz, E. (2005). Harmonization of survey data in the International Social Survey Programme (ISSP). In J. H. P. Hoffmeyer-Zlotnik & J. A. Harkness (Eds.), *Methodological aspects in cross-national research. ZUMA Nachrichten Special 11*. Mannheim, Germany: ZUMA.

Schuman, H., & Presser, S. (1981). *Questions and Answers in the Attitude Surveys: Experiments on Question Form, Wording, and Context.* New York: Harcourt Brace Jovanovich.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. Hillsdale, NJ: Erlbaum.

Schwarz, N. (2003). Culture-sensitive context effects: A challenge for cross-cultural surveys. In J. A. Harkness, F. J. R. Van de Vijver & P. P. Mohler (Eds.), *Crross-cultural survey methods*. Hoboken, NJ: Wiley.

Schwarz, N., Knauper, B., Hippler, H.-J., & Noelle-Neumann, E. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570-582.

Schwarz, N., Oyserman, D., & Peytcheva, E. (in press). Cognition, communication, and culture: Implications for the survey response process. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B.-E. Pennell & T. W. Smith (Eds.), *Survey*

methods in multinational, multiregional, and multicultural contexts. Hoboken, NJ: Wiley.

Shapiro, A. H., Rosenblood, L., Berlyne, G. M., & Finberg, J. (1976). The relationship of test familiarity to extreme response styles in Bedouin and Moroccan boys. *Journal of Cross-Cultural Psychology, 7*, 357-364.

Siegel, P., Martin, E., & Bruno, R. (2001). Language use and linguistic isolation: Historical data and methodological issues *Statistical Policy Working Paper 32: 2000 Seminar on Integrating Federal Statistical Information and Processes* (pp. 167-190). Washington DC: Federal Committee on Statistical Methodology, Office of Management and Budget.

Skevington, S. M., Sartorius, N., Amir, M., & The WHOQOL group. (2004). Developing methods for assessing quality of life in different cultural settings: The history of the WHOQOL instruments. *Social Psychiatry and Psychiatric Epidemiology, 39*, 1-8. doi: 10.1007/s00127-004-0700-5

Skevington, S. M., & Tucker, C. (1999). Designing response scales for cross-cultural use in health care: Data from the development of the UK WHOQOL. *British Journal of Medical Psychology, 72*, 51-61.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35*, 50-61.

Smith, T. W. (1995). Little things matter: A sampler of how differences in questionnaire format can affect survey responses. In A. S. Association (Ed.), *Proceedings of the Joint Statistical Meeting, Section on Survey Research Methods* (pp. 1046-1051). Alexandria, VA: AMSTAT.

Smith, T. W. (2003). Developing comparable questions in cross-national surveys. In J. A. Harkness, F. J. R. van de Vijver & P. P. Mohler (Eds.), *Cross-cultural Survey Methods*. New York: Wiley.

Smith, T. W. (2004). Developing and evaluating cross-national survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaire* (pp. 431-452). New York: Wiley.

Smith, T. W. (2009). *A translation experiment on the 2008 General Social Survey*. Paper presented at the International Workshop of Comparative Research Design and Implementation, Ann Arbor, MI.

Smith, T. W., Mohler, P. P., Harkness, J. A., & Onodera, N. (2009). Methods for assessing and calibrating response scales across countries and languages. In M. Sasaki (Ed.), *New frontiers in comparative sociology* (pp. 45-96). Boston, MA: Brill.

Smyth, J. D., Dillman, D. A., & Christian, L., M. (2007). Context effects in internet surveys. New issues and evidence. In A. Joinson, K. McKenna, T. Postmes & U.-D. Reips (Eds.), *Oxford handbook of internet psychology* (pp. 429-445). Oxford: Oxford University Press.

Stening, B. W., & Everett, J. E. (1984). Response styles in a cross-cultural managerial study. *Journal of Social Psychology, 122*, 151-156.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Struwig, J., & Roberts, B. (2006). *Data collection methodology in South Africa, Sixth ZUMA Symposium of Cross-Cultural Survey Methodology*. Paper presented at the ZUMA Symposium of Cross-Cultural Survey Methodology, Mannheim, Germany. www.hsrc.ac.za/Research_Publication-6459.phtml

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Szabo, S., Orley, J., & Saxena, S. (1997). An approach to response scale development for cross-cultural questionnaires. *European Psychologist, 2*, 270-276.

Tanzer, N. K. (1995). Cross-cultural bias in Likert-type inventories: Perfect matching factor structures and still biased? *European Journal of Psychological Assessment*, *11*, 194-201.

Thomas, R., Bremer, J., Terhanian, G., & Smith, R. (2006, March 21-22). *Scale usage differences across ethnicities and countries: Myth or reality?* Paper presented at the 8th General Online Research conference, Bielefeld, Germany.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, D.C.: National Academic Press.

Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly, 71*, 91-112.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Triandis, H. C., Marín, G., Lisansky, J., & Betancourt, H. (1984). *Simpatía* as a cultural script for Hispanics. *Journal of Personality and Social Psychology, 47*, 1363-1375.

Triandis, H. C., & Triandis, L. M. (1962). A cross-cultural study of social distance. *Psychological Monograph: general and applied, 76*.

U.S. Census Bureau. (1991). 1990 Census of population and housing, summary tape file 1 (Nebraska) Retrieved March 2009, from http://factfinder.census.gov

U.S. Census Bureau. (2001). Census 2000 summary file 1 (Nebraska) Retrieved March 2009, from http://factfinder.census.gov

U.S. Census Bureau. (2009). American Community Survey 3-Year estimates, 2005-2007 summary tables Retrieved March 2009, from http://factfinder.census.gov

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousands Oaks, CA: Sage.

van de Vijver, F. J. R., Ploubidis, G., & van Hemert, D. A. (2004). Toward an understanding of cross-cultural differences in acquiescence and extremity scoring Retrieved December 2006, from http://www.srl.uic.edu/shethsudman/presentations/vandevijver.pdf

van Hemert, D. A., van de Vijver, F. J. R., Poortinga, Y. H., & Georgas, J. (2002). Structural and functional equivalence of the Eysenck Personality Questionnaire within and between countries. *Personality and Individual Differences, 33*, 1229-1249.

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*, 346-360.

Velez, P., & Ashworth, S. D. (2007). The impact of item readability on the endorsement of the midpoint response in surveys. *Survey Research Methods, 1*, 69-74.

Verba, S. (1971). Cross-national survey research: The problem of credibility. In I. Vallier (Ed.), *Comparative methodology on sociology: Essays on trends and applications* (pp. 309-356). Berkeley, CA: University of California Press.

Villar, A. (2006). *Agreement answer scales and their impact on response styles across cultures.* Paper presented at the Annual Conference of the Midwest Association for Public Opinion Research (MAPOR), Chicago, Illinois.

Watkins, M. L. (1992). The implications of extreme response style (ERS) for cross-cultural and comparative research in South-Africa. *Journal of Industrial Psychology, 18*, 13-19.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*, 409-422.

Welkenhuysen-Gybels, J., Billiet, J., & Cambre, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology, 34*, 702-722.

Whitworth, E. M. (2001). *How selected language groups coped with Census 2000.* Paper presented at the Annual Meeting of the American Statistical Association, Atlanta, GA.

Wivagg, J., & Santos, R. (2006). *Bilingual interviewing: Contact effort, call outcomes, cooperation, and survey results in Spanish-language versus English-language telephone*

interviewing. Paper presented at the American Association of Public Opinion Research, Montreal, Canada.

Yang, Y., Harkness, J. A., Chun, T.-Y., & Villar, A. (in press). Response styles and culture: towards a conceptual framework. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B.-E. Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts*. Hoboken, NJ: Wiley.

Zax, M., & Takahashi, S. (1967). Cultural influences on response style: comparisons of Japanese and American college students. *The Journal of Social Psychology, 71*, 3-10.

Zhou, B., & McClendon, M. J. (1999). Cognitive ability and acquiescence. In American Statistical Association (Ed.), *Proceedings of the Joint Statistical Meeting, 54th Annual Conference of the American Association for Public Opinion Research* (pp. 1003-1012). Alexandria, VA: AMSTAT.