3mc2008 – Proceedings ©

# DESIGN AND ANALYSIS OF COGNITIVE INTERVIEWS FOR CROSS-NATIONAL TESTING

Kristen Miller (NCHS), Rory Fitzgerald (City U), Rachel Caspar (RTI), Martin Dimov (CSD), Michelle Gray (NatCen), Cátia Nunes (ICS), José-Luis Padilla (U of Granada), Peter Pruefer (ZUMA), Nicole Schoebi (SIDOS), Alisú Schoua-Glusberg (RSS), Sally Widdop (City U), Stephanie Willson (NCHS)

This report is a product of a multi-national testing project conducted by the Comparative Cognitive Test Workgroup. The workgroup is a coordinated, international effort organized to develop and conduct an evidence-based methodology for testing survey questions within cross-cultural or multinational contexts. For this project, the coalition consisted of representatives from 7 different nations and incorporates 6 different languages: the US (in English and Spanish), UK, Bulgaria, Portugal, Switzerland (in French), Germany, and Spain. The group is coordinated by Kristen Miller at NCHS and Rory Fitzgerald from the European Social Survey. In Fall 2007, workgroup members met in London to discuss project goals and to determine the process and protocol for conducting the study. In the next 5 months, 135 cognitive interviews were conducted by participating countries. Interviews were structured around questions provided by The Budapest Initiative (a UNECE/Eurostat task for to develop global measures for health states) and the European Social Survey (a biannual cross-national attitude survey conducted in over 30 European countries). In February, 2008, the group held a joint analysis meeting whereby a process was developed to conduct a systematic, comparative analysis of those interviews. Through this process, the group was able to identify various interpretive patterns resulting from socio-cultural and language-related differences among countries as well as other patterns of error that could undermine the comparability of survey data. This report summarizes the process and findings of the Budapest Initiative component.

This paper will layout:

I. The types of analyses that are possible with this data
II. The types of findings that can be discovered
III. What should be done differently to improve the method

## BACKGROUND

### Description of Cognitive Testing

The aim of cognitive testing is to investigate how well survey questions perform when asked of respondents, that is, if respondents understand the questions according to their intended design and if they can provide accurate answers based on that intent. As a qualitative method, the primary benefit of cognitive interviewing is

that it provides rich, contextual insight into the ways in which respondents 1) interpret a question, 2) consider and weigh out relevant aspects of their lives and, finally, 3) formulate a response based on that consideration. As such, cognitive interviewing provides in-depth understanding of the ways in which a question operates, the kind of phenomena that it captures, and how it ultimately serves (or fails) the scientific goal. Findings from a cognitive interviewing project typically lead to recommendations for improving a survey question, or results can be used in post-survey analysis to assist in data interpretation.

Traditionally, cognitive testing is performed by conducting in-depth, semi-structured interviews with a small sample of approximately fifteen to thirty respondents. The typical interview structure consists of respondents first answering the evaluated question and then answering a series of follow-up probe questions that reveal what respondents were thinking and their rationale for that specific response. In this regard, cognitive interviews unfold within a narrative format and are often personal and, in comparison to traditional survey interviews, are particularly unique to each respondent. Through this semi-structured design, various types of question-response problems, such as interpretive errors or recall accuracy, are uncovered—problems that often go unnoticed in traditional survey interviews. By asking respondents to provide textual verification and the process by which they formulated their answer, elusive errors (what DeMaio and Rothgeb have termed "silent misunderstandings")[1] are revealed.

As a qualitative method, the sample selection for a cognitive testing project is purposive. Respondents are not selected through a random process, but rather are selected for specific characteristics such as gender or race or some other attribute that is relevant to the type of questions being examined. When studying questions designed to identify persons with disabilities, for example, the test sample would likely consist of respondents with a previously known disability and, to discover potential causes of false positive reporting, some respondents with no known disability. Because of the small sample size, not all social and demographic groups are represented. Analysis of cognitive interviews does not produce generalizable findings, but rather, provides an explicit exploration of response processes including patterns of interpretation which could lead to response error.

Analysis of cognitive interviews can be conducted from transcribed interviews or, as is often the case, from interviewer notes. The texts of the interviews (either transcribed materials or interviewer notes) are collated by question so that comparisons can be made systematically across all respondents. Several levels of analysis can typically be performed. First, distinct occurrences in which respondents experience difficulty or confusion while answering are identified. Additionally, specific instances or patterns of error are also noted and, most importantly, the particular causes of those errors are identified. In addition to response errors, analysis of cognitive interviews can be conducted to reveal patterns of question interpretation. By comparing each respondent's interpretation to a particular question, patterns can

[1]

be identified and then examined for consistency and degree of variation among respondents. This type of interpretive analysis does not necessarily illustrate overt response errors, but rather provides deeper insight into the substance or the actual meaning that constitutes the survey data.

Application for Cross-National or Cross-Cultural Surveys

Socio-cultural differences among respondents can generate question response differences, not just in terms of differing with the scientific intent, but with the way that other respondents may view or process a question. For example, American Indian respondents who use tobacco in sacred rituals may be confused whether they should count ritual-use for general smoking questions, or, if they assume that the question pertains to ritual, may take offense to the sacred character of the question.[2] Similarly, respondents from particular cultural regions may be less inclined to report having a physical limitation because disability is considered a stigmatized condition.[3] Uncovering these types of socio-cultural differences are particularly important to identify prior to fielding a survey. Rather than interpreting the differences in the survey data as bias in the response process, they can be construed as "real" differences and reported as study findings.

By conducting a comparative analysis of cognitive interviews, it is possible to identify patterns of error and patterns of interpretation across groups of people. For example, a cognitive testing study conducted in both rural Mississippi and the metropolitan DC area illustrated that those rural respondents with limited access to health care were more likely to experience problems understanding questions that contained technical wording, such as "mammogram" and "PSA test."[4] Additionally, by conducting a comparative analysis, cognitive interviews can identify problems in questions that have been incorrectly translated or that convey even subtle meaning differences in other languages. As such, the method can provide insight into whether a particular error pattern or interpretive pattern might be idiosyncratic or could produce systematic bias in the survey data.

With cross-national surveys there is an additional element that cognitive testing can help to examine. It is arguable that the impact of the context in which the questions is asked is significantly amplified on a cross-national survey. This is certainly true for The European Social Survey. For example whilst welcoming its wide coverage of a diverse range of cultures and political systems, it also poses in questionnaire design. On the one hand, the larger the number of countries in a cross-national study, so the greater is the analytical potential of the data with a wider range of national contexts available as independent variables. On the other hand, it is arguable that the larger the number and the more diverse the range of countries, so the more difficult it is for the study to produce equivalence. For instance, the entrance of Turkey into the ESS in Round 2 as its first Muslim country raised immediate issues about the Judaic-Christian assumptions behind the existing set of ESS questions on

[2]

[3]

[4] Miller, American Journal of Health Behavior

religion. Similarly the imported ESS questions on 'democracy' cued in quite different issues within the 'new' democracies of East and Central Europe from those they cued in within Western Europe. In the new democracies, the word referred more to free elections, while in the older democracies it referred to civil rights and liberties. Naturally we annotated the source questionnaire prior to its translation to convey to translators which of the two different connotations we were looking for, and then simply hoping that a form of words is available in every language to cue in the equivalent connotation[5]. Cognitive interviewing can help to identify such difficulties prior to fielding allowing researchers the opportunity to consider methods to achieve equivalence or accept the limitations of the method.

Limitations of Comparative Cognitive Testing

Because cognitive testing directly examines the thought processes that respondents use to answer survey questions, the method holds much promise for uncovering cultural or language-related problems in question design. Nevertheless, because of specific characteristics inherent to the methodology, the method itself has potential limitations for conducting comparative analyses and, ultimately, for making socio-cultural based conclusions. Those characteristics include:

- Small sample size
- Non-representative sample
- Non-standardized interviewing protocol
- Requirement of trained interviewers
- Under-developed literature and practice regarding rigor of analysis
- Lack of standardized criteria for what constitutes a cognitive interview finding

To conduct a successful comparative analysis, these characteristics must be included as an integral component in the design of the interviewing protocol as well as in the actual method of analysis.

Although it can identify particular problems, because of the small, purposive sample, the method provides little insight into the actual prevalence or the magnitude of impact that the particular problem may have on the survey data. The strength of the method is that it reveals interpretive patterns as well as the contextual frame of the question-response process—it does not provide a specific threshold by which a question fails or passes. Whether only one cognitive test respondent experienced the problem does not imply that the problem should be characterized as a fluke. Nor should one case of a problem be seen as a serious flaw. In order to determine whether or not a particular problem discovered is serious enough to attempt "fixing," the researcher must take into account various pieces of information such as the nature of the problem, whether or not the problem is tied to specific characteristics of respondents or possible experiences, and weigh out how the survey data may ultimately be affected by the flaw. This level of insight can only come from a systematic analysis across all of the cognitive interviews. As a qualitative study, cognitive test findings provide pieces of insight from various perspectives that, when

[5] Fitzgerald and Jowell (2008)

brought together, can assist the question design analyst in assessing the quality of the question as it pertains to the type of phenomena that it should capture. In this regard, the usefulness of findings is tied to the type of analysis that is actually performed as well as the completeness of that analysis.

On a practical level, the non-standardized interviewing protocol, which is critical for fully exploring how each respondent interprets and formulates a response to a question, makes comparative analyses between multiple sites difficult. Unless analyses across the test sites are coordinated and conducted in concert, cognitive interviews themselves may not be comparable. Additionally, in conducting a comparative analysis, it is important to consider how data were collected (e.g. with trained or inexperienced interviewers, from thinking aloud narratives or pre-scripted follow-up probes), how the interviews were recorded (e.g. interviewer notes or transcriptions), as well as how the cognitive interview data may be limited or even flawed. That is, in conducting a comparative analysis, it is critical to consider the validity of the cognitive interview data itself and how the data quality might vary across the different interviewing sites. Without taking this necessary step, it will be difficult to distinguish an "actual" comparative finding from artifacts of the cognitive interviewing process, particularly if that process involves numerous locations with different interviewers conducting interviews in multiple languages.

## METHODS

The primary objective of the Comparative Workgroup project was to develop and conduct a protocol that would assess each question's performance as well as to make an evaluative statement regarding their comparability across multiple countries and differing languages. In developing the test protocol, the workgroup set out to answer the following comparative questions:

1. Do the survey questions work consistently across all countries and subgroups?
2. Do respondents interpret questions consistently regardless of country, language, or demographic?
3. Do respondents use the same thought processes to answer questions?
4. If not, then, why are there differences? What about the countries, languages or demographic subgroups generate different response processes?
5. How can we "fix" or manage these differences through question design?

To fully answer these questions, the workgroup would need to identify and address the aspects of cognitive testing that undermine comparability across test sites.

To begin the project, a meeting of workgroup members was held in London to lay out the parameters of the project and to establish the testing protocol. Aspects of traditional cognitive testing were discussed and then incorporated into the overall design. Those issues included:

1. Sample composition, selection and recruitment

2. Language equivalence and translation procedures
3. Use of non-standardized probing techniques, the impact on data quality and comparability, and establishment of a semi-structured interview guide
4. Differing skill levels of interviewers, impact on data quality and comparability, and interviewer training
5. Cognitive interview documentation, what constitutes a finding, and data processing and organization

Importantly, plans were made to ensure communication and coordination across the multiple interviewing locations. Specifically, weekly conference calls were scheduled, and time-lines were established for making translations and conducting interviews. Additionally, the ESS created a workgroup website so that common documents (e.g. the interviewing guide, sample requirements, translation procedures) could be easily accessed, and members could pose questions and have discussions with group members. Lastly, a final workgroup meeting was scheduled after all interviews were conducted to analyze interview data through a systematic group process. That joint analysis took place in Washington DC, in February 2008.

The following sections detail the design and implementation of that process.

**Sampling**

Countries were asked to conduct a minimum of 10 interviews and, if possible, were urged to conduct more interviews. It was determined that differences in sample size, while not ideal, would not bias the analysis as it would in a quantitative study. The greatest disadvantage would be that, for those countries with smaller samples, the possibility of an incomplete data set would be greater, that is, it does not fully capture the range of question response processes as it would for a larger sample. The key problem at the analysis stage is instances where one finds a problem in a country with a large sample that cannot also be found in a country with a smaller sample. One cannot always know if the problem exists in both countries or not making the decision about whether to 'fix' the question more difficult. Fixing the question in one country might create new difficulties in countries where the original problem does not exist. Where possible therefore the sample sizes should ensure equivalent coverage in each country. Take for example a test of questions for a nationally representative sample survey. A homogeneous country would probably require a smaller sample size than a one with a more heterogeneous one. But this should be taken into account in the design.

Samples were to be diverse in age, gender and socio-economic status. Additionally, to adequately test the health state questions, at least half of respondents were to have either a hearing, visibility, mobility or cognitive functioning problem. This was an inevitable compromise between the sampling needs of the BI where health conditions needed to be over represented and the ESS that was looking for a sample broadly reflecting the general population according to key characteristics that were likely to affect comprehension and processing of attitude questions. Since the

sample was purposive and based on specific requirements, countries were able to recruit by whatever means was most efficient for them, for example, by placing an advertisement, handing out fliers, or through existing networks of respondents. All countries except Bulgaria provided respondent remuneration (approx. $40USD).

The charts below depict the number, demographic profile and health state of the final sample for each country.

**Respondent Demographics by Country**

| | Total | Gender | | Age (in years) | | | Education | |
|---|---|---|---|---|---|---|---|---|
| | | Men | Women | 18 – 29 | 30–69 | 70+ | < HS degree | HS degree + |
| Bulgaria | 10 | 5 | 5 | 2 | 4 | 4 | 4 | 6 |
| Germany | 10 | 5 | 5 | 2 | 4 | 4 | 4 | 6 |
| Great Britain | 29 | 15 | 14 | 8 | 9 | 12 | 9 | 20 |
| Portugal | 8 | 3 | 5 | 3 | 3 | 2 | 3 | 5 |
| Spain | 18 | 10 | 8 | 6 | 6 | 6 | 9 | 9 |
| Switzerland-French | 17 | 9 | 8 | 7 | 4 | 6 | 2 | 12 |
| United States-English | 30 | 11 | 19 | 3 | 19 | 8 | 14 | 16 |
| United States-Spanish | 13 | 3 | 10 | 1 | 9 | 3 | 6 | 7 |
| **Total** | 135 | 61 | 74 | 32 | 58 | 45 | 54 | 81 |

**Kommentar [L1]:** I am not sure it was always degree that was used as the cutting point. Please check this.

**Respondent Health Problems by Country**

| | Mobility | Hearing | Cognitive | Mental Health |
|---|---|---|---|---|
| Bulgaria | 3 | 2 | 1 | 1 |
| Germany | 2 | 2 | 1 | 0 |
| Great Britain | 5 | 8 | 3 | 2 |
| Portugal | 3 | 0 | 1 | 0 |
| Spain | 3 | 4 | 3 | 0 |
| Switzerland-French | 2 | 3 | 0 | 1 |
| United States-English | 14 | 4 | 3 | 4 |
| United States-Spanish | 3 | 2 | 3 | 5 |
| **Total** | 35 | 23 | 15 | 13 |

## Data Collection

The interviewing protocol consisted of two sections: a BI component and an ESS component. It was semi-structured, consisting of the test questions followed by a few general pre-scripted probe questions. Interviewers were instructed to spend 30 minutes on each section regardless of whether or not that component was completed. Additionally, interviewers were instructed to begin half of their interviews with the

**Formatiert:** Einzug: Erste Zeile: 0 cm

BI component and the other half with the ESS questions. The protocol was written in English. (See Appendix A). Countries that were conducting interviews in languages other than English were responsible for producing a translated protocol. Countries were required to produce translations using the committee approach. (See Appendix B). Translation is of course a critical element in the process of developing an equivalent questionnaire. Where possible countries were asked to use the ESS committee TRAPD approach to translation of the protocol[6]. This occurred in Switzerland, Spain, Germany and with the US Spanish questionnaire. This technique avoids back translation and instead uses a team approach to develop an optimally equivalent translation In Bulgaria and Portugal it was not possible to implement this procedure in full but back translation was still avoided. In any event it is essential that where possible all countries use the same translation procedure and that the procedure used prior to cognitive testing is identical to that which will be used before the ultimate field implementation. This aids later evaluation of the 'source' of problems with particular questions.

The written probes were intended to serve only as a guide for interviewers (particularly those inexperienced in cognitive interviewing) to illicit how respondents understood the question as well as how they formulated their answer; the prescripted probes were not intended to be used rigidly. During the interview, respondents were asked each survey item and were then probed to explain their answer. Each interview varied depending on whether the respondents had a physical or mental health problem. Typical follow-up questions included, "How so?" and "Why do you say that?"

**Kommentar [L2]:** I am not sure how so was typical – the 2 open probes on the protocol were typical from what I gathered at the joint analysis meerting

Interviewers ranged in their cognitive interviewing experience. Specifically, interviewers for the US, Spain, Germany and the GB were very experienced and regularly conducted cognitive tests. On the other hand, cognitive interviewing was new to those interviewing for Bulgaria, Portugal and Switzerland. To compensate for the lack of experience, a training session was held at the London meeting. Additionally, particular effort was given to communicate with those newer interviewers throughout the project. In retrospect it would have been optimal for all those conducing interviews to have attended an interviewing style harmonisation meeting. It is likely that different institutions train differing style of probing, note taking and analysis. Making some attempt at harmonising these for cross-national projects would be useful in future.

All interviews were audio-recorded except for those conducted in Spain and the US-English, which were video-recorded. From these recordings, interviewers wrote detailed sets of notes which were then compiled by question. Interviewers then charted their data in tables formatted so they would be easily accessible for a thorough joint analysis. Notes documents were written in the language of the interview, however, charts were translated into English so that all workgroup members could understand and analyze data across all countries.

**Kommentar [L3]:** I think some countries tries went straight from tapes to charts eg Switzerland

[6] Harkness (2007)

**Formatiert:** Englisch (Großbritannien)

The main problem conveyed by interviewers was that the protocol was too long; there was not enough time to adequately cover all of the questions. Although interviewers attempted to prioritize questions that were not covered in previous interviews, some sections had incomplete data. This was particularly the case for the hearing question that was placed at the end of the BI component. For the ESS questions some of the items at the end on age stereotypes only received basic attention and some of the other age items with serious problems were not covered in detail because major revisions were needed. Consequently, results will not be presented for these items.

**Method of Analysis**

For cross-national or cross-subgroup comparative analyses, the analysis itself should be conceptualized in three distinct layers. The first and simplest level of analysis occurs within the interview, specifically, as the interviewer attempts to understand how one respondent has come to understand, process and then answer a survey question. The interviewer must act as analyst during the interview, evaluating the information that the respondent describes and following up with additional questions if there are gaps, incongruencies or disjunctures in the explanation. From this vantage point (i.e. within a single cognitive interview) basic response errors, such as recall trouble or misinterpretation, can be identified.

The second layer of analysis occurs through a systematic examination of all interviews together. Specifically, interviews should be examined to identify patterns in the way respondents interpret and process the question. By making comparisons across all of the interviews, patterns can be identified and then examined for consistency and degree of variation among respondents. Inconsistencies in the way respondents interpret questions may not necessarily mean misinterpretation, but can illustrate even the subtle interpretation differences that respondents use as they consider the question in relationship to their own life circumstance. From this vantage point is it possible to identify the phenomena that is captured by the particular survey question which, in the end, illustrates the substantive meaning behind the statistic. Additionally, from this layer of analysis, it is possible to identify patterns of calculation across respondents. This is particularly useful for example in understanding how qualifying clauses such as, “in the past 2 weeks” or “on average” impact the way respondents form their answer and whether respondents consistently use the clause in their calculation.

The last level, the heart of the cross-cultural analysis, occurs through an examination of the patterns across sub-groups, identifying whether particular groups of respondents interpret or process a question differently. This level of analysis (i.e., identifying patterned differences among subgroups) is particularly important because this is where potential for bias would occur. A key sub-group in cross-national questionnaire development is country since this represents a key source of likely differences in the way respondents process the question.

To implement these layers of analysis for this Comparative Workgroup project, cognitive interview data was charted, allowing for a systematic analysis across all interviews. (See Appendix 3). At the workgroup meeting, analysis consisted of a lead researcher guiding the workgroup through the multiple levels of analysis—first identifying basic errors, then, determining whether those errors occurred in patterns across interviews. Secondly, the interview data was examined to identify patterns of interpretation and patterns of calculation. Finally, the patterns were further investigated to determine whether they occurred within in a specific subgroup. Because the charts were organized by country, subgroup comparisons focused primarily on country and language differences. Charts were used as the primary source of data, but interviewer notes were also referenced when clarification was needed. For a few instances were even further clarification was required, workgroup members reviewed recordings of the interviews—though this review occurred after the analysis meeting. Because of limited time, analysis of the BI questions could not be completed in the joint analysis meeting. The remaining analysis occurred after the meeting, with one researcher analyzing the charts and then following up with group members for clarification as required.

**Kommentar [L4]:** The charts can't be published because of data privacy issues. What will be in Appendix 3? I assume blank charts with the headings. We should (for berline0 discuss the differing charting techniques that we have used.

**Kommentar [k5]:** I'm planning on putting in a "fake"chart and noting that it is fake because of confidentiality. Hwoever, I think it's important to illustrate what the charts really looked like because it's a key piece of the method

**Kommentar [L6]:** I think we need to distill this further

## RESULTS

Prior to the meeting a scheme to identify the sources of error in cross-national social surveys was developed based upon experience from questionnaire design in the European Social Survey (Fitzgerald, 2008). It was hoped that cognitive interviewing would enable these sources to be identified in turn aiding researchers in their efforts to fix problems. The error sources were as follows:

**Kommentar [L7]:** It is essential to me that we stress that this scheme was based on theory and empirical evidence. My working paper on the background to the scheme will be available prior to Berlin. Since I developed the scheme I would ask that this is cited

1. Socio-cultural differences: given respondents' socio-cultural context, the question is attempting to measure a concept that either does not exist or takes on a different of meanings that are not comparable.

2. Translation error – the translation of the item produced a question in the target language which was not functionally equivalent to that in the source questionnaire.

3. Interaction between source question and translation – the question appears to work well in British English (or the source language being used) but has features in its design which make translation difficult. Examples include the use of idioms, colloquial language, scales with vague quantifiers.

Outside of these comparative problems, some questions in the protocol were categorized by a fourth category that would apply in all cognitive interviewing projects …

4. Poor source question: all or part of a question is poorly designed such that the question (even in the source language) does not measure the phenomena as intended.

**Kommentar [L8]:** Please switch these around and introduce the source problems first

**Kommentar [k9]:** I'm not sure why. To me if flows better this way

For the BI questions, the majority of comparative problem types were related to translation or the interaction between the source question and the translation.

EXAMPLES to illustrate:

1. the 4 types of findings
2. Would be nice to additionally illustrate the myriad of types of analyses that can be conducted.

Possible Examples to Pull From:

**Walking**

Short Distance

**How much difficulty do you have walking 100 yards on level ground that would be _____? Would you say: no difficulty, a little difficulty, a lot of difficulty, or are you unable to do this?**

*If aid:* **How much difficulty do you have walking 100 yards on level ground that would be _____ without using your _____** [mention aid(s) in W1b]? **Would you say: no difficulty, a little difficulty, a lot of difficulty, or are you unable to do this?**

The following table depicts how respondents answered the BI question.

**Responses by Country**

| Country | None | Little | A Lot | Unable | Cannot Answer | Total |
|---|---|---|---|---|---|---|
| US English | 14 | 10 | 2 | 3 | 0 | 29 |
| US Spanish | 8 | 3 | 0 | 0 | 0 | 11 |
| Switzerland | 13 | 1 | 1 | 0 | 0 | 15 |
| Spain | 15 | 3 | 0 | 0 | 0 | 18 |
| Portugal | 7 | 1 | 0 | 0 | 0 | 8 |
| Germany | 8 | 1 | 0 | 1 | 0 | 10 |
| Bulgaria | 6 | 1 | 3 | 0 | 0 | 10 |
| Great Britain | 20 | 5 | 1 | 2 | 0 | 28 |
| | 91 (69.4%) | 27 (20.6%) | 7 (5%) | 6 (4.6%) | 0 | 131 |

In explaining the basis of their answers, respondents primarily described day-to-day experiences walking, for example, “going to the store,” “exercising on a treadmill,” “walking the dog,” and “walking from the mountain into town.” Among the Spanish interviews, however, there were a few cases that extended outside the

**Kommentar [L10]:** The focus of the Berlin paper should really be on the comparative angle. If other lessons about cognitive interviewing and analysis were learnt then there is another paper in it that I think you should write. But for 3MC we should keep the focus on cross-national comparisons

**Kommentar [k11]:** I agree for the most part—though we need to be illustrate through this discussion that this is a much more sophisticated data set than what is normally attributed to cognitive testing studies. It is a huge advantage, and I think it is important to illuatrate this asset—if not making it explict. Once the ESS work is done, we need to work together to figure out what examples to include

**Kommentar [L12]:** I think it is OK to show the numbers but the proportions should be removed because they have little meaning

**Kommentar [k13]:** I disagree—I they add—to the point that readers are going to mentally calculate then anyway.

action of walking, such as “climbing stairs,” “gardening,” and “daily activities.” At this time, it is not clear why there is a difference among Spanish respondents; an explanation will require additional analysis as well a comparison of the US-Spanish and Spain translations.

Almost half of the French-speaking respondents from Switzerland misunderstood the question as asking about running instead of walking. This error was found to be related to the translation.Rather than understanding the phrase "parcourir 100 mètres" (which, in English, means "to cover the distance of 100 meters"), some respondents understood the word as "courir" (which means “to run”).

**Kommentar [L14]:** If this is a translation issue why did some respondents understand correctly?

Regarding the distance, each country was asked to use whatever examples that they deemed appropriate for their country. Those examples were:

| **Country** | **Example of 100 meters** |
|---|---|
| US English | “the length of a football field” |
| US Spanish | “the length of a football field” |
| Switzerland | “the length of a football field” |
| Spain | “the length of a football field” |
| Portugal | “one lap of a running track” |
| Germany | “the length of a football field” |
| Bulgaria | Interviewer described the distance in the interview |
| Great Britain | “the length of a football field” |

Relatively consistent across the countries and languages, many respondents stated that the examples helped them to define 100 meters/yards; they would not have known how to define that distance without examples. However, some respondents stated that, even with the examples, they were not able conceptualize 100 meters/yards. These were primarily women or other respondents who were not familiar with sports-related references. While it was difficult for some of these respondents to form an answer, all respondents were able to speculate the distance (with varying degrees of accuracy) in order to provide a response. It is important to note that, in many cases, it is difficult to assess whether or not respondents’ conceptions of 100 meters/yards is truly accurate. To explain their conceptualization, respondents could only describe specific landmarks in which only they were familiar (e.g., the distance from their house to the bus stop, or from their house to school). However, consensus among workgroup members was that, in most cases, estimates were likely to be accurate, and the group concluded that the examples were beneficial and did not contribute additional problems.

It was clear, however, that in a few cases, respondents incorrectly overestimated the distance (for example, thinking it was equivalent to 2 kilometers) and then answered incorrectly—because they did not believe they could walk that exaggerated length. There were no cases in which a respondent underestimated the length, thereby, reporting that they would have no difficulty when, in reality, they would have difficulty.

Approximately one in five respondents stated that they had mobility problems to the extent that they required the use of an assistive device, such as a cane, walker or wheelchair. The chart below illustrates the sample with a breakdown of assistive device use.

**Use of Assistive Device by Country**

| Country | Device | Total |
|---|---|---|
| US English | 14 | 29 |
| US Spanish | 2 | 11 |
| Switzerland | 2 | 15 |
| Spain | 0 | 18 |
| Portugal | 0 | 8 |
| Germany | 2 | 10 |
| Bulgaria | 4 | 10 |
| Great Britain | 4 | 28 |
| | 29 (22.1%) | 131 |

For those respondents requiring assistive devices, there were no outstanding problems regarding the device clause. Specifically, no respondent had difficulty understanding and then reporting their ability to walk without the use of their device. Only a couple respondents, acknowledged some confusion, but had no difficulty once the interviewer clarified the clause or simply repeated the question.

In a few cases across each country, respondents had difficulty answering the question because their particular type of walking problem is not always constant. Instead, their problems varied along the basis of a chronic condition (e.g., osteoporosis, arthritis) or environmental conditions (the weather, ground cover such as cobblestones or grass). In these cases, respondents were apprehensive about providing an answer that was rooted within an amount or magnitude of difficulty, but were more inclined to answer with frequency, such as "sometimes."

Long Distance

*If no aid:* **How much difficulty do you have walking 500 yards on level ground that would be about _____? Would you say: no difficulty, a little difficulty, a lot of difficulty, or are you unable to do this?**

*If aid:* **How much difficulty do you have walking 500 yards on level ground that would be about _____ without using your _____** [mention aid(s) in W1b]? **Would you say: no difficulty, a little difficulty, a lot of difficulty, or are you unable to do this?**

The table below shows respondents answers to the question:

**Responses by Country**

| Country | None | Little | A Lot | Unable | Cannot Answer | Total |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| US English | 13 | 3 | 4 | 5 | 0 | 25 |
| US Spanish | 2 | 6 | 1 | 0 | 3 | 12 |
| Switzerland | 10 | 2 | 2 | 0 | 1 | 15 |
| Spain | 14 | 2 | 2 | 0 | 0 | 18 |
| Portugal | 7 | 0 | 1 | 0 | 0 | 8 |
| Germany | 7 | 1 | 1 | 1 | 0 | 10 |
| Bulgaria | 6 | 1 | 1 | 1 | 1 | 10 |
| Great Britain | 17 | 4 | 3 | 2 | 1 | 27 |
| | 76 (60.8%) | 19 (15.2%) | 15 (12%) | 9 (7.2%) | 6 (4.8%) | 125 |

For most aspects, the long distance question operated in the same manner as the short distance question. Like the previous question, there were no difficulties regarding the assistive device clause; respondents had no outstanding difficulty understanding or reporting their ability without the use of their aid. And, like the previous question, a few Swiss respondents understood the question as asking about running (even though it was previously established in the short-distance question that the verb was "parcourir" and not "courir").

However, respondents' conceptualization of the longer distance became more of a problem than in the previous question. Unlike the examples in the short-distance question, these examples were less tangible and, in some cases, too abstract for respondents to imagine. The examples were:

| **Country** | **Example of 500 meters** |
|---|---|
| US English | Washington DC: "1/3 of a mile"<br>North Carolina: Interviewer used example of the road that all respondents traveled to get to the site of the interview |
| US Spanish | "the length of 5 football fields" |
| Switzerland | "the length of 5 football fields" |
| Spain | "the length of 5 football fields" |
| Portugal | "a bit more than a running track" |
| Germany | "the length of 5 football fields" |
| Bulgaria | (Interviewer described the distance in the interview) |
| Great Britain | "the length of 5 football fields" |

Consequently, a few respondents (unlike in the previous question) were unable to speculate the distance and provide an answer. Some other respondents did provide an answer, however, it was based on a grossly overestimated conceptualization of 500 meters. For example, one Spanish respondent imagined five football pitches to be "very far" and, consequently, reported that she would have "a lot" of difficulty. Similarly, in thinking that the distance must be extremely far, a US Spanish speaking respondent could not answer the question, stating, "I don't know, I have never done it."

## Cognition

Cognition 1

**Because of a physical, mental or emotional problem, do you have difficulty concentrating, remembering or making decisions? Would you say: no difficulty, a little difficulty, a lot of difficulty, or are you unable to do these things?**

In the joint analysis meeting, no overt problems were identified. Initially, it was suggested that the question might be double-barrelled, but upon examination there was no evidence to substantiate the problem. A more subtle problem, however, some respondents (for whatever reason) were not viewing the question as health question, one intended to measure cognitive impairment. More specifically, workgroup members suggested that some respondents who answered "a little" might have answered thinking of relatively trivial problems and not because of a true cognitive functioning problem. Without tallied data, however, it was impossible to fully explore this concern in the meeting.

Now, in looking at the totality of responses, it is possible to see whether or not and the extent that an interpretation problem exists within the sample. The table below shows those results by country.

**Responses by Country**

| Country | None | Little | A Lot | Unable | Total |
|---|---|---|---|---|---|
| US English | 9 | 15 | 5 | 0 | 29 |
| US Spanish | 5 | 7 | 1 | 0 | 13 |
| Switzerland | 5 | 8 | 2 | 0 | 15 |
| Spain | 11 | 7 | 0 | 0 | 18 |
| Portugal | 4 | 4 | 0 | 0 | 8 |
| Germany | 5 | 4 | 1 | 0 | 10 |
| Bulgaria | 2 | 6 | 2 | 0 | 10 |
| Great Britain | 13 | 13 | 3 | 0 | 29 |
| | 54 (40.9%) | 63 (47.7%) | 14 (10.6%) | 0 | 132 |

Of primary concern, more than half of all respondents reported a cognitive functioning problem, and, given the sample selection criteria, it is implausible that all of these respondents would have a true problem. This indicates the likelihood of an interpretation problem. That is, as suspected in the joint analysis, it appears that some respondents are not viewing the question as one that captures functioning problems. Looking across the countries, however, there does not appear to be any country that stands out as different; if there is an interpretation problem at least the problem appears to be across the board.

Given the specific nature of the problem, it is possible that this problem could easily be resolved by counting the "little" reports as "none"—essentially turning the variable into 3 categories None/A lot/Unable. However, this is only a viable solution if those who interpreted the question as "normal" problems answered "a little," and those who interpreted it as a mental health question answered "a lot." This, however,

requires determining what types of cognitive problems might be excluded by combing those respondents. To explore the viability of this solution, the cognitive interview data would need to be more closely examined.

To address this issue, interview data was examined to determine what respondents considered when answering "little"—specifically, to discern between those cases of solid cognitive functioning problems and those who reported trivial, questionable problems. Those cases that were classified as *questionable* were those that (from data in the charts) explained that it was "not really a problem," mentioned that the problem was of "no significance" or "no impact on their life," that the type of problem is "common" or "usual," or that they were not concerned about the problem. Those cases that were deemed *more solid* were those that (from data in the charts) indicated that the problem was due to an emotional, mental or health problem, such as a stroke, ADD or depression. In the end, approximately two-thirds of all respondents answering "little" appeared to have solid cognitive functioning problems. A full one-third, however, described their problem as normal–not interpreting the question as a question about cognitive functioning. The table below illustrates the break down by country.

| Country | Respondents Reporting "Little" | | Total Reports of "Little" |
|---|---|---|---|
| | Questionable | More solid | |
| US English | 6 | 9 | 15 |
| US Spanish | 3 | 4 | 7 |
| Switzerland | 5 | 3 | 8 |
| Spain | 1 | 6 | 7 |
| Portugal | 2 | 2 | 4 |
| Germany | 1 | 3 | 4 |
| Bulgaria | 1 | 5 | 6 |
| Great Britain | 3 | 9 | 12 |
| | 22 (34.9%) | 41 (65%) | 63 (47.7) |
| | (17.7% of all 132 cases) | (31% of all 132 cases) | |

It is important to note that, to a certain extent, the distinction between the two forms of interpretation are unclear—particularly because the phenomena itself is subject to interpretation. For example, it is not clear whether cognitive problems due to temporary depression or grief should be counted as a true cognitive problem, even though it may manifest as a true problem in a respondent's life. Additionally, some respondents answered affirmatively thinking of forgetting seemingly trivial items such as forgetting names or birthdays. However, to them—especially if their memory problem occurs daily—the problem is not trivial. Not wanting to override respondents' judgements, these vaguer cases were left in the *more solid* category. Further, and most importantly, the categorizations are only as good as the description provided in the charts.

Nevertheless, it can be safely concluded that while those responses in the "little" category contain an element based on a "normal" interpretation, it also contains an element of true cognitive impairment—an element that would not want to be lost by

collapsing the category "little" with "none." Consequently, it may be prudent to consider other ways of asking about cognitive functioning in order to more accurately capture variation in functioning abilities.

Cognition 2

**How much difficulty do you have remembering important things? Would you say: no difficulty, a little difficulty, a lot of difficulty, or are you unable to do these things?**

With a cursory look at the initial responses, this second cognition question appears better at capturing the health interpretation than did the first cognition question. See the table below. Only 37% of respondents answered affirmatively to this second question. Additionally, while half of all respondents answered "a little" in the previous question, only a third answered "a little" to this question—essentially the same amount of the *more solid* responses from the previous question.

**Responses by Country**

| Country | None | Little | A lot | Unable | Total |
|---|---|---|---|---|---|
| US English | 10 | 6 | 3 | 0 | 19 |
| US Spanish | 8 | 6 | 0 | 0 | 14 |
| Switzerland | 12 | 2 | 1 | 0 | 15 |
| Spain | 13 | 4 | 1 | 0 | 18 |
| Portugal | 5 | 3 | 0 | 0 | 8 |
| Germany | 7 | 3 | 0 | 0 | 10 |
| Bulgaria | 6 | 3 | 1 | 0 | 10 |
| Great Britain | 15 | 11 | 2 | 0 | 28 |
| | 76 (62.3%) | 38 (31.1%) | 7 (5.7%) | 0 | 122 |

At first glance, it might appear that those "little" responses of Cognition 2 might be capturing the *more solid* responses of Cognition 1. If this is true, then the finding would lead to the conclusion that Cognition 2 is the better question (in that it more accurately distinguishes those respondents with truer cognitive functioning problems). The cross analysis of the two questions (see chart below) depicts those respondents answering "a little" to either cognition question, thereby characterizing the extent to which the two questions overlap and separating out the truer cases of functioning problems.

| | | **Cognition 1: How much difficulty do you have concentrating, remembering, or making decisions?** | | | |
|---|---|---|---|---|---|
| | | **None** | **A little** | | **A lot** |
| **Cognition 2** | **None** | | **Questionable** | **More Certain** | |

| | | | | | |
|---|---|---|---|---|---|
| | | ---------- | USE12 USE17<br>USE10 USE25<br>USE28<br>USS2 USS8<br>Sw3 Sw14 Sw20<br>Sw24 Sp05<br>P2 Sp14<br>G6 GB12 GB46<br><br>**(13.1% of cases)** | USE19 USE24 Sw21<br>Sw22 Sw23<br>Sp10 Sp11 Sp16<br>P6<br>G2 G7<br>B1 B2 B3 B6 GB24<br>GB31 GB33 GB37<br>GB43<br><br>**(17.2% of cases)** | ---------- |
| | **A little** | USE11 USE31<br>USS13<br>Sp3 Sp17<br>P1 P5<br>G5 G8<br>GB17 GB28 GB36<br>GB41<br><br>**(10.6% of cases)** | B7<br>USE22<br>USS11<br>GB47<br>P4<br><br>**(4% of cases)** | USE18 USE26<br>USE27 USE29<br>USE38 USE33<br>USE36<br>USS1 USS12 USS7<br>USS6 Sp7 Sp18 P8<br>G9 B5 GB13 GB26<br>GB34 GB35 GB44<br><br>**(17.2% of cases)** | Sw34<br>Sw35<br>USS5 B9<br>GB11<br>GB47<br><br>**(4.9%)** |
| | **A lot** | ------ | Sw1<br><br>**(.8% of cases)** | | ---------- |

* the letter-number combinations appearing in the cells are identifications for individual respondents within countries: USE-United States English, USS-United States Spanish, Sw-Switzerland, P-Portugal, G-Germany, B-Bulgaria, GB- Great Britain, S-Spain.

If the above hypothesis is correct (i.e., that the Cognition 2 question more accurately sorts out the trivial problems), then the majority of cases would be located in the Bright Green (as opposed to the Red and Pink) area of the chart. (And, in the Bright Green area as opposed to the Light Green area if the Questionable/More Certain categorizations are correct.) The fact that there are so many cases in the Bright Red areas suggest that this hypothesis is not correct and that, while there is some overlap, the two questions appear to capture some relatively different ideas. To better understand the extent of the incongruity between the two questions, the qualitative data of individual cases was examined to determine why these respondents answered "a little" to one of the cognition questions, but "none" to the other. This level of analysis could illustrate how and why each question performed differently, as well as which question better captured the phenomena intended by the Budapest Initiative.

Firstly, analysis was conducted to explain those cases that are captured by Cognition 1, but not by Cognition 2 (those falling in the pink and red areas of the chart). Of all respondents, 37 (30.3%) answered "a little" to Cognitive 1 (Concentrating, Remembering, Making Decision) and "none" to Cognitive 2 (Remembering Important Things). The following 3 themes explain the incongruity between the two questions (note, that because they are not mutually exclusive, some cases appear in more than one theme):

1) 13 of those respondents based their answer on a concentration or decision making problem; they did not have a memory problem so answered "none" to Cognitive 2. This is important because it illustrates that Cognitive 1 is picking up this dimension of functioning ability that Cognitive 2 is not picking up.

Those cases include: USE28, Sw21, P2, P6, G6, G7, B1, B3, GB12, GB24, GB31, GB37, GB43

2) 25 of those respondents were impacted by the word "important" in the Cognitive 2 question, such that it raised the criterion from the first to the second question to a more serious level: these respondents would answer yes, "a little" to the Cognitive 1 question, but "none" to the more serious Cognitive 2 question because of the word "important." If the q*uestionable/more solid* categories are correct, we should see most of these cases in the pink area of the chart. While there are many, there are still a fair amount in the Red area. This conclusion would suggest that Cognitive 2 might be a better question—at least in sorting out the trivial problems. However, a critical qualitative finding is that respondents broadly varied in their interpretation of "important things," from "remembering relatives birthdays" to "paying medical bills." Consequently, those respondents with more severe interpretations of "important things" were inappropriately sifted out of the Cognitive 1 Question, while those with less severe interpretations were included. This conclusion suggests that Cognition 2 is not the better question because, while it does pare down respondents reporting problems, it does not necessarily pare down the correct respondents.

Those cases include: USE12, USE17, USE10, USE25, USS2, USS8, Sw3, Sw14, Sw20, Sw21, Sw23, Sw24, Sp05, Sp11, Sp14, Sp16, G7, P2, P6, B2, B3, B6, GB12, GB46, GB33

3) 6 of the respondents answered none to Cognition 2 because they have specifically developed strategies to not forget those "important" things. This is critical because the intent of the question is to measure health state; it is not intended to pick up adaptive strategies.

Those cases include: USE19, USE24, Sp5, P2, G6, GB31

For 4 of the cases, it was difficult to make sense of the discrepancy; there was not enough detailed information to explain the discrepancy.

Those cases include: Sw22, Sp10, G2, GB24

Finally, to explain those cases that are captured by Cognition 2 (Important Things), but not by Cognition 1 (Concentrating, Remembering, Making Decisions), specifically, those cases in the blue area of the chart. Of the entire sample, 13 (10.6%) respondents answered "a little" to important things but "none" to concentrating, remembering, deciding. Of those cases, explanations for only two of those cases could be determined. First, one respondent (P5) did not consider memory in the concentrating, remembering and making decisions, and then said no—but did have a problem with remembering, which he did in the important things question.

Second, another respondent (P1) answered "none" to the first question, but answered the second question "little" because she has forgotten some birthdays of family members, which "are important!" Hearing the word "important" in the second question changed the types of things that she would include in her answer.

Without being able to determine the incongruity among the other cases, it is impossible to determine if these two cases represent a common theme. It is possible, however, to speculate. For example, as was for P1, the word "important" in the Cognitive 2 question might have operated in the opposite direction for some respondents than intended. Additionally, it may be possible that some respondents did not consider the word "important" when forming an answer. Both of these explanations (should they prove true) would further indicate that Cognitive 2 is a weaker question.

Conclusion: The Cognition 1 question appears to capture more of the intended phenomena than the Cognition 2 question. The second question is more likely to miss those with concentrating and making decision problems, as well as those who have adapted life strategies to compensate for their cognitive functioning problem. Further Cognition 2 is subject to a broad range of interpretation because of the word "important." While this question is able to pare down respondents (more than Cognition 1) because of the word "important," it does not consistently and equally do so across all respondents and so does not necessarily sort out the correct respondents. From this analysis, the interpretive variation does not appear to be systematic across any one country or language. At the same time, the Cognition 1 question does appear to be capturing some respondents that do not have true cognitive functioning impairment, and consequently, could be improved.

## Affect

### Affect 1

**Overall, during the past week, how worried, nervous, or anxious did you feel? Would you say: not at all, slightly, moderately, a lot, or extremely worried, nervous, or anxious?**

The table below shows respondents answers to the question:

**Responses by Country**

| Country | Not at all | Slightly | Moderately | A lot | Extremely | Total |
|---|---|---|---|---|---|---|
| US English | 10 | 6 | 4 | 3 | 1 | 24 |
| US Spanish | 3 | 1 | 3 | 2 | 4 | 13 |
| Switzerland | 0 | 2 | 9 | 3 | 1 | 15 |
| Spain | 7 | 6 | 4 | 0 | 1 | 18 |
| Portugal | 1 | 1 | 5 | 1 | 0 | 8 |
| Germany | 1 | 5 | 3 | 0 | 0 | 9 |
| Bulgaria | 4 | 2 | 2 | 0 | 2 | 10 |
| Great Britain | 5 | 8 | 11 | 1 | 3 | 28 |
| | 31 (24.8%) | 31 (24.8%) | 41 (32.8%) | 10 (8%) | 12 (9.6%) | 125 |

As in the Cognition 1 question, workgroup members suggested that this question could be double-barrelled. However, while some respondents stated that their answer varied for each of the three feeling statements, when presented with response categories, they were able to formulate one response to the question. Only one German respondent refused to answer Affect 1, stating that his answer would differ along the three different feelings.

For the most part, respondents in all of the countries thought specifically about worrying. In only a few cases did the respondent think outside this interpretation. For example, one US Spanish-speaking respondent answered in regards to his clinically-diagnosed depression as opposed to anxiety. Another US Spanish-speaking respondent answered "moderately" thinking of anxiety as the happiness and anticipation of taking a trip to visit her relatives.

In the joint analysis, the group identified two elemental themes by which respondents based their answers: 1) specific experiences or episodes in the past week or 2) a state of being, such as a characteristic of their personality or a more static condition like being unemployed.

Those that based their answer upon a state of being considered such things as 1) a health problem that has them concerned (Spain7, Bulgaria10), 2) worry about economic insecurity (Spain8, Portugal3) and 3) recognition that they are "worriers by nature" (e.g., GB1 described constantly clenching her teeth). Those that based their answer more upon state of being were not as likely to focus on the "past week" time reference posed in the question. It is not clear if these respondents ignored the time frame because it was not relevant to their conceptualization, or whether they formed their conceptualization specifically because they did not hear the time frame.

Those that considered the time frame were inclined to consider specific incidents or experiences within the past week. Examples include:

US17: got a speeding ticket so was worried that wife would be angry
US18: was worried about speaking Spanish in Intro to Spanish class
Swiss 11: needed to find people for his shooting society
Swiss 20: was worried about permission from army about weekend leave
Swiss 24: was taking an apprenticeship class
Swiss 35: was preparing the Christmas meal and wanted it to be nice

Germany 1: was worried when the cat was sick
Germany 10: had work due at the university
Bulgaria 8: felt nervous about a test
Great Britain 3: was going away for the weekend but hadn't heard confirmation from the hotel

In forming their answers, respondents who considered a specific incident tended to evaluate that specific incident and did not average out the amount of worry across the full seven days. That is, while the question asked respondents to consider the seriousness or magnitude of their anxiety and then average it across the week, most were inclined to simply rate the magnitude or seriousness (as they perceived it) of the one or two particular incidents. This explains the rather large amount of respondents (a full three-fourths of the sample) reporting an anxiety problem—with half reporting a least a moderate problem. It should be noted that only 13 respondents in the entire study were screened in with a mental health problem

In this way, it appears that the time frame might also undermine the reliability of the question. In the joint analysis, each country identified which cases had incongruent answers to the open and closed versions of the question. In some of those cases, the shifting of answers was due to the fact that respondents, by the time the second question was asked, thought of another incident that rated differently.

It should be noted as well, that in some of these shifts, specifically the "none" to the "slightly" responses, occurred because respondents interpreted the word "slightly" as "close to none." At first, with the open-ended version, respondents did not believe that the question was asking about such insignificant worries, but then picked up this connotation when the response categories contained "slightly."

Additionally, error was identified in some respondents' consideration of "the past week." Specifically, a few respondents considered the past couple months, another the past couple weeks, and still another the past day (interestingly, this respondent changed her answer when she considered the entire week because she remembered another episode of worry.)

In sum, the interpretations were relatively consistent in that most considered worrying. However, the specific basis of the responses varied across the two patterns: incidents vs. state of being. The seven day time frame generates another variable.

The need for additional analysis was also identified. For example, a more careful analysis could be conducted by comparing the different subgroups (i.e. country and language) base their answer—state of being vs. episode. However, this would require the charts to be more consistent and specific than they currently are in order to categorize and count each case. Additionally, in the workgroup analysis meeting, some discussion arose about potentially different meanings conveyed in the translated versions of the response categories. Because the response categories consisted of 5

levels of vague quantifiers (not at all, slightly, moderately, a lot , extremely), there is a greater potential for loss of comparability across the languages.

Affect 2

**Overall, during the past week, how sad, low, or depressed did you feel? Would you say: not at all, slightly, moderately, a lot, or extremely sad, low, or depressed?**

The same themes from the first affect question extend to the second—that respondents base their answer on either a specific episode in the past week, or they consider their the personality or state of being due to a relatively static situation such as their poor health or unemployment. However, for this question, more respondents tended to base their answer on a state of being. This explains the drop in rates from Affect 1 to Affect 2 (only half as opposed to three-fourths the sample reported a problem, with only a third as opposed to half reported at least a moderate problem) The table below shows respondents answers to the question:

**Responses by Country**

| Country | Not at all | Slightly | Moderately | A lot | Extremely | Total |
|---|---|---|---|---|---|---|
| US English | 10 | 3 | 1 | 4 | 0 | 18 |
| US Spanish | 3 | 2 | 3 | 3 | 2 | 13 |
| Switzerland | 6 | 2 | 3 | 3 | 0 | 14 |
| Spain | 9 | 6 | 1 | 1 | 1 | 18 |
| Portugal | 2 | 0 | 6 | 0 | 0 | 8 |
| Germany | 6 | 1 | 3 | 0 | 0 | 10 |
| Bulgaria | 3 | 2 | 2 | 3 | 0 | 10 |
| Great Britain | 15 | 7 | 3 | 2 | 1 | 28 |
| | 54 (45.4%) | 23 (19.3%) | 22 (18.5%) | 16 (13.4%) | 4 (3.3%) | 119 |

As in the previous question, “state-of-being” respondents tended to ignore the past week clause and focused on the current situation (however long) that they associated with their sadness. For example, Bulgaria3 answered a lot because she was “old and alone;” she wasn’t thinking of any particular time frame.

Conclusion: This question appears to be more solid than Affect 1 simply because people are more likely to evaluate their state of being as opposed to a particular experience in the past week. In this question, some respondents still base their answer on an episode (the Swiss woman cooking the Christmas meal in the Affect 1 question, was still thinking about the meal in Affect 2 Question—she wasn’t sure that everyone would appreciate her meal). Perhaps the number of “episode-based” respondents would be reduced if there were no time-frame presented in the question. It’s not clear how the past week clause is impacts the question response process for those who are thinking about the state of being, as it appears that these cases tend to answer according to their current state and time period (however long) that coincides with that state.

## Pain

Pain 1

**Overall, during the past week, how much physical pain or discomfort did you have? Would you say: none at all, a little, moderate, a lot, or extreme physical pain or physical discomfort?**

Similarly in each country, respondents included a diverse range of causes, including arthritis, a bad fall, a root canal, tinnitus, sore muscles from exercise, a pierced tongue, a cataract operation, tingling in the hands, stomach pain, headaches, swelled feet, and a cut finger—essentially including any incident or episode that (to them) caused pain. The table below shows respondents answers to the question by country:

| Country | Not at all | A little | Moderate | A lot | Extremely | Total |
|---|---|---|---|---|---|---|
| US English | 4 | 7 | 4 | 3 | 0 | 18 |
| US Spanish | 6 | 3 | 2 | 1 | 0 | 12 |
| Switzerland | 4 | 3 | 5 | 2 | 1 | 15 |
| Spain | 3 | 10 | 4 | 0 | 0 | 17 |
| Portugal | 4 | 1 | 2 | 1 | 0 | 8 |
| Germany | 4 | 5 | 1 | 0 | 0 | 10 |
| Bulgaria | 3 | 3 | 2 | 1 | 1 | 10 |
| Great Britain | 4 | 10 | 8 | 5 | 1 | 28 |
| | 32 | 42 | 28 | 13 | 3 | 118 |

Discussion from the joint analysis meeting revealed that a translation issue regarding the word *discomfort* created a potential comparability problem. Depending on the word chosen for the translation, *discomfort* could mean a lower threshold of pain (which is the intended interpretation) or a general sense of uncomfortable-ness. For example, one Bulgarian respondent stated that *discomfort* occurs after eating or drinking too much and getting no sleep, whereas, *pain* is a much graver situation. Similarly, another Bulgarian respondent answered affirmatively because she had the flu; she had a runny nose and couldn't breathe. Another US Spanish-speaking respondent answered *a lot* because she did a lot of work and was feeling tired. After the meeting, with further examination of the charts, it was discovered that some British respondents also interpreted *discomfort* as being uncomfortable and reported being tired or having stress. US English-speaking respondents, however, understood *discomfort* to mean low-level pain and did not relate the concept to sickness or fatigue.

Regarding magnitude, it was difficult for respondents to explain in detail how they arrived at their answer. Other than simply describing their pain in terms of "it was a lot" or "it was very bad," respondents were limited in their ability to describe the amount or the intensity of their pain. Instead, respondents tended to explain their answer in the following ways:

1. the impact of the pain on their lives, specifically, whether or not (and the extent to which) they could overcome the pain. For example, some respondents described their mental stamina and how they simply would not let pain interfere with their daily activities. Others described how pain medication allowed them to carry on in their usual way.
2. the amount of concern they had about the pain, particularly, if they were concerned that the pain was indicative of a more serious condition. For example, one US respondent who answered *a lot*, explained her answer stating that, even though her doctor told her that her toe pain was from arthritis, she was not convinced and is worried there is another problem.
3. the frequency or time-span of the pain and the equation they used to average across a time period. As the question asks, some respondents specifically calculated across the past week. For example, one Portuguese respondent answered *a little* because her knee (which she hurt in a fall the previous week) is not a continuous pain; it only hurts with particular motions, for example, when she exercises. However, other respondents did not average across the entire week. For example, one Swiss respondent who fell while skiing answered *moderate*, explaining that for 20 minutes the pain was very sharp. He did not answer *extreme* because there was no physical damage and the pain went away. However, he did not answer *a little* because, at the time, it was very painful.

Because pain threshold is a uniquely subjective phenomena, it was impossible in the interview to investigate the validity of each respondent's answer—particularly in the way respondents referenced the magnitude or intensity of the pain. Even by examining the way respondents justified their answer, it was impossible to determine the correctness of their response. Some respondents, for example, reported *mild* or *moderate* because they had to take a pain medicine to alleviate the symptom. However, some others in the same situation answered *none* because the pain went away.

Instead of considering perceptions of pain and the inevitable variability of those perceptions, the central concern in measuring pain is in understanding how respondents arrive at their answer, specifically, the calculations that they perform as well as the various factors that are considered. To be sure, the most identifiable variability occurred in the way that respondents calculated their answer. Because this question contains multiple factors (frequency, intensity and time period), respondents have many paths by which they could formulate an answer. To get the most comparable pain reports, then, it makes sense to separate the concepts, asking them separately.

Pain 2

**How many days during the past week did you have physical pain or discomfort?**

*Record number of days*:

_____

Problems:

- For those with short, non-serious episodes, the day time-frame was a problem. For example, one Portuguese respondent cut herself with a knife and, on another day, had a stomach ache. She did not feel as though it was accurate to report two days, instead, she said it was really "two moments."

- a Spanish speaker didn't understand the phrase, "in the past week"

In forming answers, respondents either:
1. Counted the specific days or nights—to help them count, Some respondents thought of the particular activities that the did in the past week, for example gardening or shopping, that they remembered doing with pain.

   This is difficult for some because it is hard to say when it discretely ended because the nature of pain is that it gradually abates

   For those that did not have serious pain, for example, a simple headache, it was also hard for them to recollect.

2. Had constant pain, so answered "everyday"

3. Estimated (i.e. did not count) because they were not cued into the fact from the question before that an actual count was requested

Recommend rewording to "in the past 7 days" because respondents may be more likely to count

Pain 3

**During those times when you had physical pain or discomfort, how would you describe your level of physical pain or discomfort? Would you say it was mild, moderate, severe or extreme?**

- Question operated in the same way as the Pain 1—not clear if there are real any differences in the 2 questions
- Would be nice to compare respondents answers from Pain 1 to Pain 3 (easily done if charts were using different software)
- Same issues as with the first question regarding asking about magnitude
  - Subjective phenomena without ability to validate
  - Respondents use different factors to calculate, though less variability because they are asked to report specifically on the time period with pain (essentially to exclude the time that they were pain-free). There

was variation on this in Pain 1; it's not clear whether this was the case for Pain 3

- For those with more than one episode of pain or different levels of pain throughout the week, it was more difficult and still required some calculation that varied across respondents
    - Some respondents took the median, thinking that for one day it was extreme, but then for the last day it was mild, so they answered moderate.
    - Some other respondents answered providing a more exact average across the number of days that they had pain.
    - Several other respondents answered thinking of the most extreme period instead of taking an average.

Pain 4

**Thinking about the last time you had physical pain or discomfort: On a scale from 1 to 100 how intense was the pain: 0 is no pain or discomfort and 100 is the worst pain or discomfort imaginable.**

Scale: 0 ---------- 100

*Record response*: _____

- Difficult question for respondents because the end points (though seemingly demarcated) are vague and difficult to imagine
- As in the previous questions about intensity, respondents had to pull together various aspects of their pain (i.e. impact on life, concern, frequency) to have something concrete to report
- Because the task was so difficult, many respondents gave random or thoughtless answer, numerous reports of 50. For example, one respondent said, "About half and half. It wasn't too extreme and it wasn't too bad."

## Fatigue

Fatigue 1

**During the past week, how many days have you felt tired or had little energy?**

- Diverse range of what respondents included:

- Feeling sleepy, a question about motivation and feelings of depression, being tired from having asthma, feeling tired after having a big meal (this respondent reported 3 days), tired from working hard, sleepy and tired from being overweight, depression, feeling sick, tired from exercising, being sick, having low blood sugar, mental tiredness from an emotional problem, not sleeping well, having hang over, recovering from an operation, driving for a long time
- Problem of counting days if it is only part of a day
- Problem with the vague quantifiers and not intuiting the amount from the order (e.g. thinking moderate is less than mild)
- German translation issue was noted that it is difficult to translate feeling tired

Fatigue 2

**On those days, how much of the day did you feel tired or have little energy? Would you say all day, most of the day, about half of the day, or only for a few hours?**

- Primarily worked well
- Problem if the tiredness was more than one day and if it varied on those days

Fatigue 3

**During those times when you felt tired or had little energy, how would you describe your level of tiredness or loss of energy? Would you say it was mild, moderate, severe or extreme?**

Same issues as with Pain 1 and Pain 3

Respondents had to rework the question so that it was answerable—and did so in the following ways:

- The amount that the tiredness impacted their life
- The frequency and the duration of the fatigue
- Their ability to control the level of fatigue (e.g. one respondent answered moderate, explaining that because she had the option of controlling it, that she could have slept, but if she were sick and can't even consider it, that's severe)

Fatigue 4

**How much of a problem did you have with feeling tired or having little energy? Would you say none, a little, some or a lot?**

- Because this was at the end of the interview, there were not many cases and respondents were fatigued.
- However, no problems were identified
- Respondents answered based on the impact that the tiredness had on their life; however, because this was at the end, it is hard to say if respondents were thinking this way because of the previous questions.

## CONCLUSION

1) This is an important step in comparative survey research
2) Significant in that it was a coordinated effort that took on the hard issues: what do we want to know in comparative cog testing, how does the method serve and how does it fail
3) Describe all the benefits
   a) For example: This type of analysis is advantageous because it allows for a comparison of at least two questions, but it doesn't hold one as a gold standard. Comparative analyses of questions can be extremely informative if the method of analysis is not just about the strength of the relationship between the two questions—specifically because this often sets one as the standard. In this way, The Washington group analysis is like this analysis—the use of the various patterns allows for comparisons between questions without holding one as a standard; and in the analysis we are able to learn as much about the follow up probes as we do the Washington Group test question. I know that there is more to this and it would be good to tease this out because, in the end, I think this is the crux of the argument for IRT and other quantitative modeling for question evaluation.
4) How to improve
   a) Better coordination for the charting
   b) Have people submit chart data interview by interview—not wait until all of the interviews are collected
   c) Use Access or some other database tool to allow for more complex analyses
   d) Have the an initial first wave of analysis using the entire dataset
   e) Use the joint analysis meeting to discuss what was found in that first wave of analysis