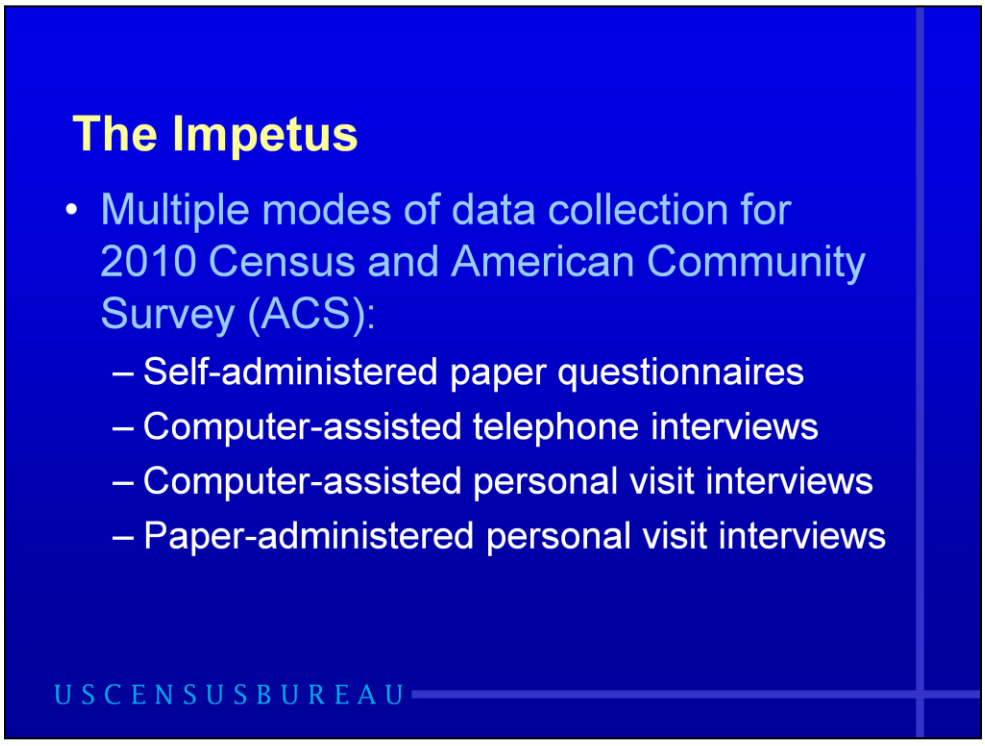
Mode Consistency Guidelines: Asking Questions in Multimode Surveys

Paper Presented at 2007 FedCASIC

Jennifer Hunter Childs,
Theresa DeMaio, Eleanor Gerber,
Joan Hill, Elizabeth Martin,
and Courtney Reiser

U.S. Census Bureau

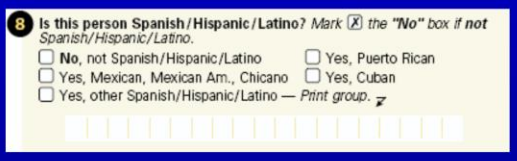
USCENSUSBUREAU



In conjunction with experts in the field of survey methodology, the U.S. Census Bureau developed guidelines for survey questionnaire development that aim at maintaining consistent data quality across multimode surveys. The impetus for this effort was the multiple modes of data collection that will be used for the 2010 Census and the American Community Survey (ACS), including self-administered questionnaires and telephone and personal visit interviews conducted using both paper and electronic instruments. Review of the various questionnaires being developed during the census planning cycle revealed an extreme amount of variation between instruments for the same questions. All aspects of questionnaire design differed -- sometimes the question wording was changed, sometimes different punctuation was used, sometimes the layout was different. No one anticipated the amount of variation that would be found across the various instruments and modes. In some instances, questions or response categories were changed so substantially when they were adapted for use in another mode, that they were no longer really the same question. Such differences may give rise to large differences in the data for the same question (see, e.g., Bennett and Griffin, 2002; Martin and Gerber, 2003).

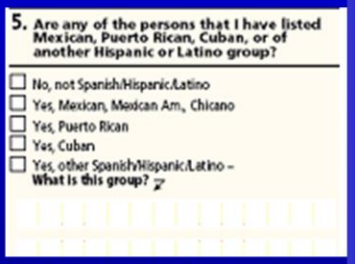
The Impetus (cont.)

- Variation between past instruments in:
  - Question wording
  - Response Task
  - Layout



8 **Is this person Spanish / Hispanic / Latino?** *Mark* ☒ *the* ***"No"*** *box if* ***not*** *Spanish/Hispanic/Latino.*

☐ **No**, not Spanish/Hispanic/Latino
☐ Yes, Puerto Rican
☐ Yes, Mexican, Mexican Am., Chicano
☐ Yes, Cuban
☐ Yes, other Spanish/Hispanic/Latino — *Print group.*



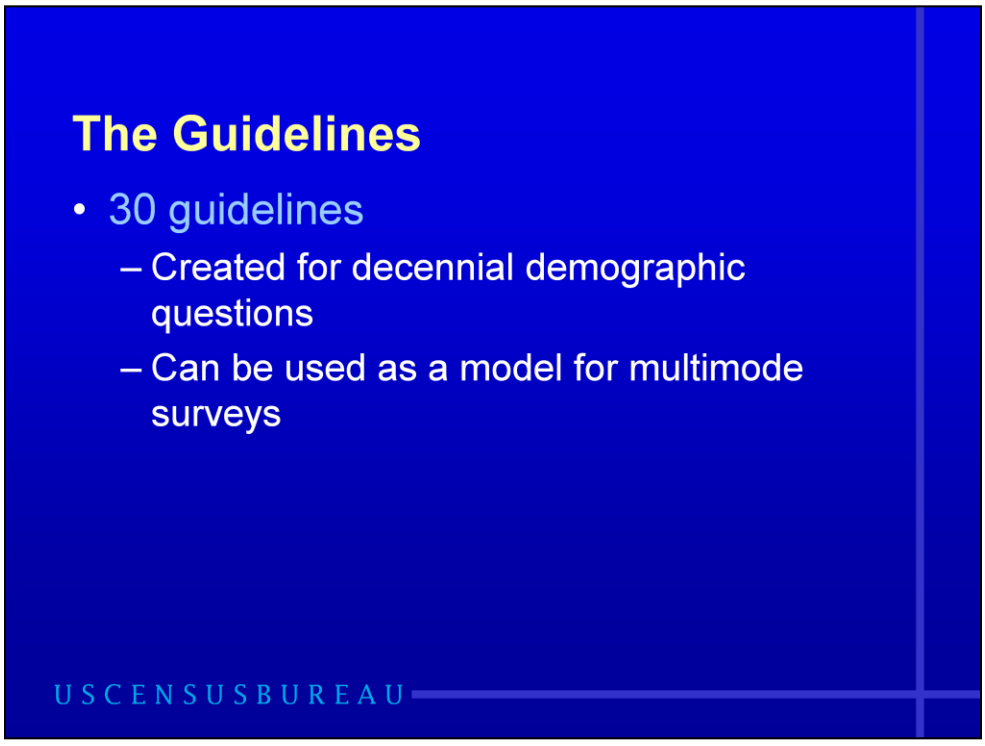
5. **Are any of the persons that I have listed Mexican, Puerto Rican, Cuban, or of another Hispanic or Latino group?**

☐ No, not Spanish/Hispanic/Latino
☐ Yes, Mexican, Mexican Am., Chicano
☐ Yes, Puerto Rican
☐ Yes, Cuban
☐ Yes, other Spanish/Hispanic/Latino – **What is this group?**

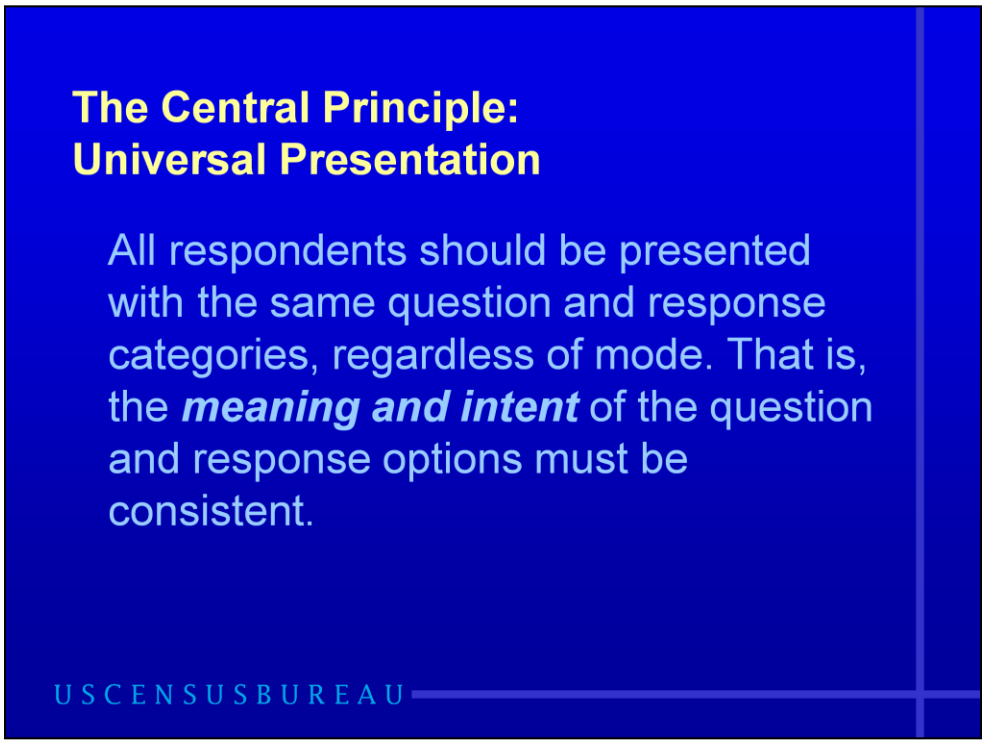
The Working Group

- Census Bureau Staff
- Outside Experts

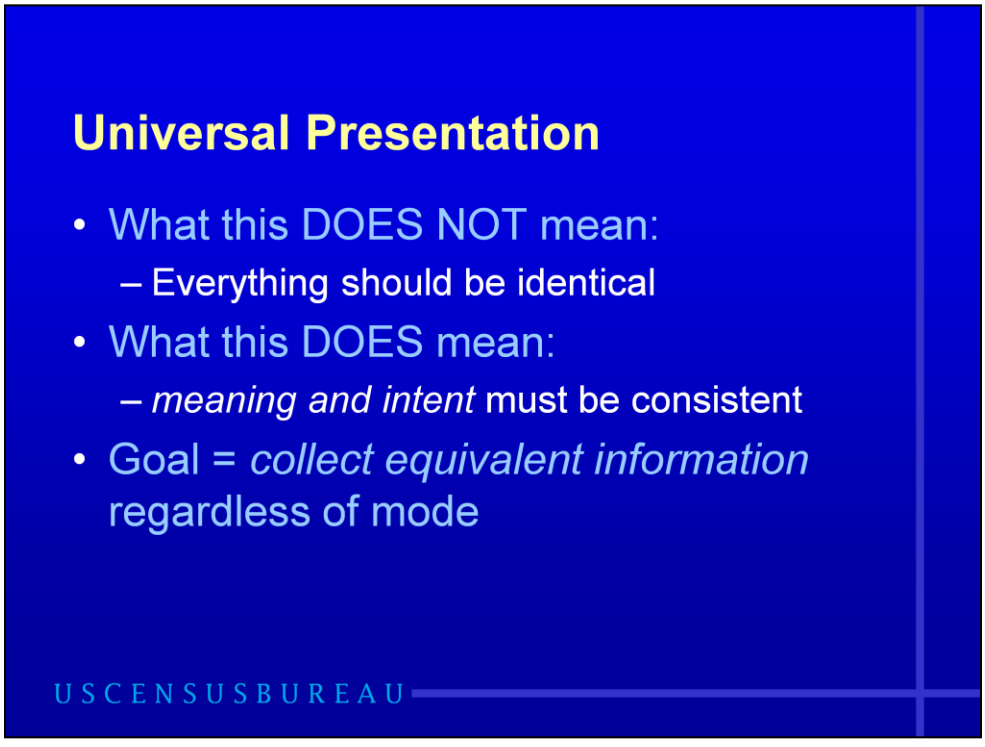
USCENSUSBUREAU



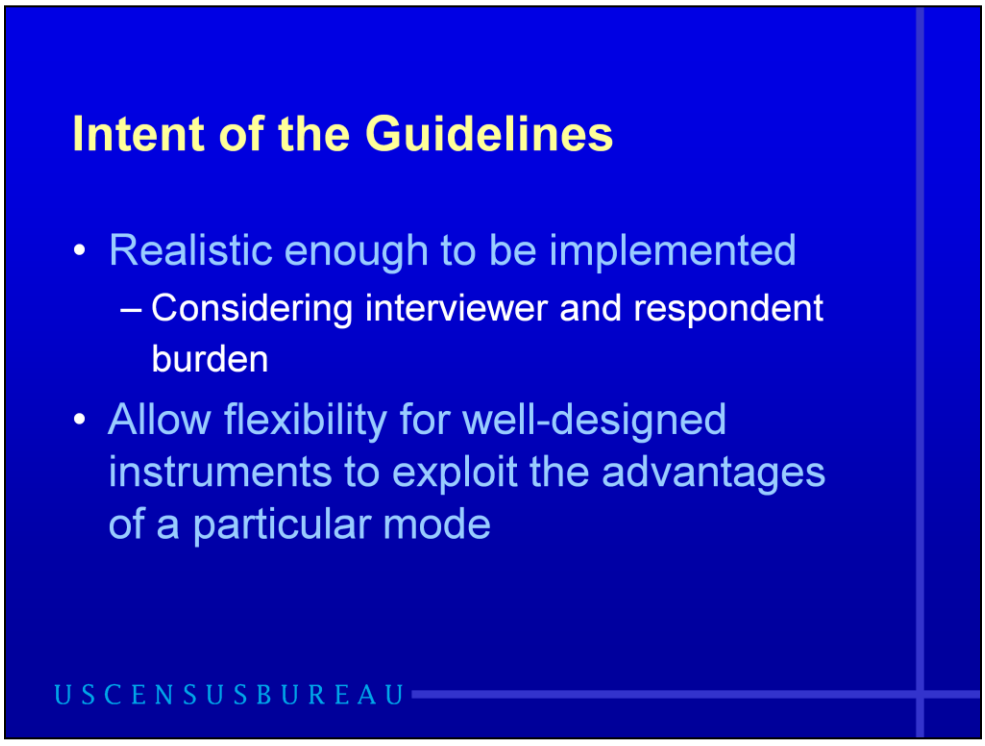
An interdivisional working group charged with investigating mode differences in the census and ACS questionnaires developed 30 guidelines that, while created with the decennial “short form” demographic questions in mind, can be used across organizations as a model for developing multimode surveys. The guidelines introduce the principle of Universal Presentation and apply it to several major aspects of instrument design including question wording and instructions, visual design elements, flashcards, edit messages and help. Other Guidelines outline apparent exceptions to this principle. These are situations in which a change in the question wording, order, instructions, or other features is essential for operational reasons, or better preserves the question than would asking exactly the same question in an identical way in different modes. This paper/presentation describes the guiding principle of these guidelines – that of universal presentation – and presents several guidelines that apply to the major aspects of questionnaire design previously mentioned.



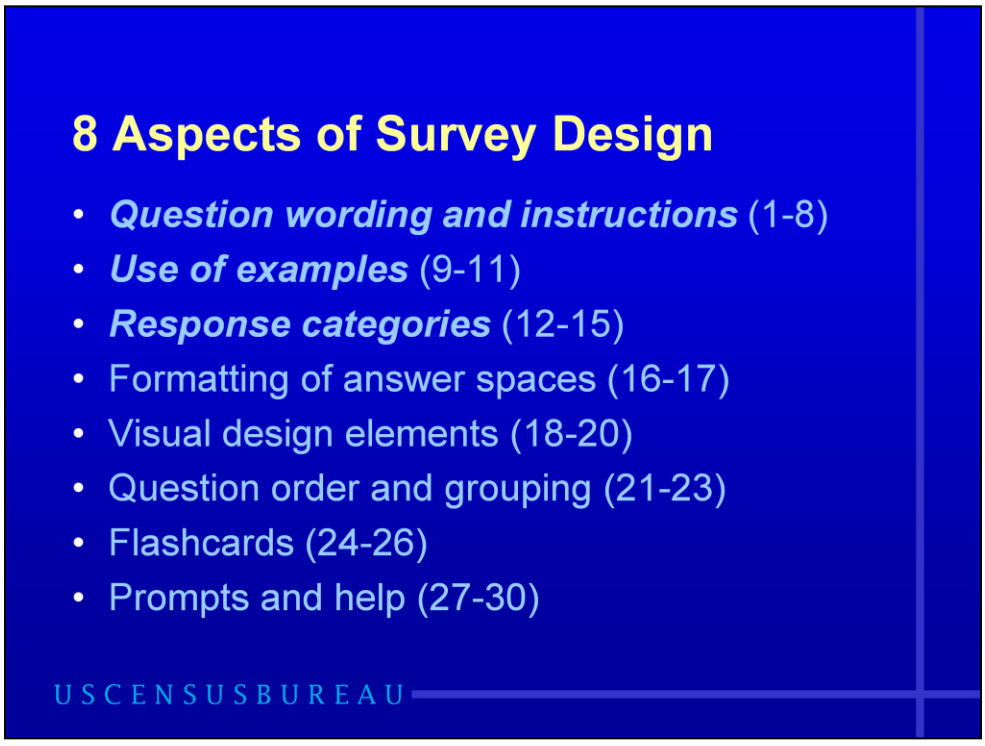
The underlying principle for these guidelines is ***Universal Presentation***: All respondents should be presented with the same question and response categories, regardless of mode. That is, the meaning and intent of the question and response options must be consistent.



While one might assume that this principle requires questions, categories, instructions, etc., to be identical across modes, this assumption turns out to be neither feasible nor desirable. Rote repetition would result in awkward and difficult-to-administer instruments that are unlikely to achieve consistent response data. Rather, Universal Presentation says that the meaning and intent of the question and response options must be consistent. In some cases, questions or instructions need to be modified so they can be communicated to, attended to, and understood by respondents the same way in different modes. The goal is that instruments collect equivalent information regardless of mode. By equivalent, we mean that the same respondent would give the same substantive answer to a question regardless of the mode of administration.



We attempt to balance the operational demands of the census with the requirement of standardization. For example, it is neither reasonable nor realistic to require that very long lists of response categories or examples be repeated exactly as worded for every person in tens of millions of U.S. households by quickly-trained census enumerators. Evidence from behavior coding in census tests shows that enumerators do not read long lists of response categories. These Guidelines try to set forth a realistic expectation of field implementation, and attempt to allow enough flexibility to permit development of well-designed instruments that exploit the advantages of a particular mode while maintaining essential consistency. The Guidelines are intended to support census enumerators and ACS interviewers in carrying out their jobs well, rather than making interviews more difficult and time-consuming, or by forcing them to take shortcuts that undermine data consistency.

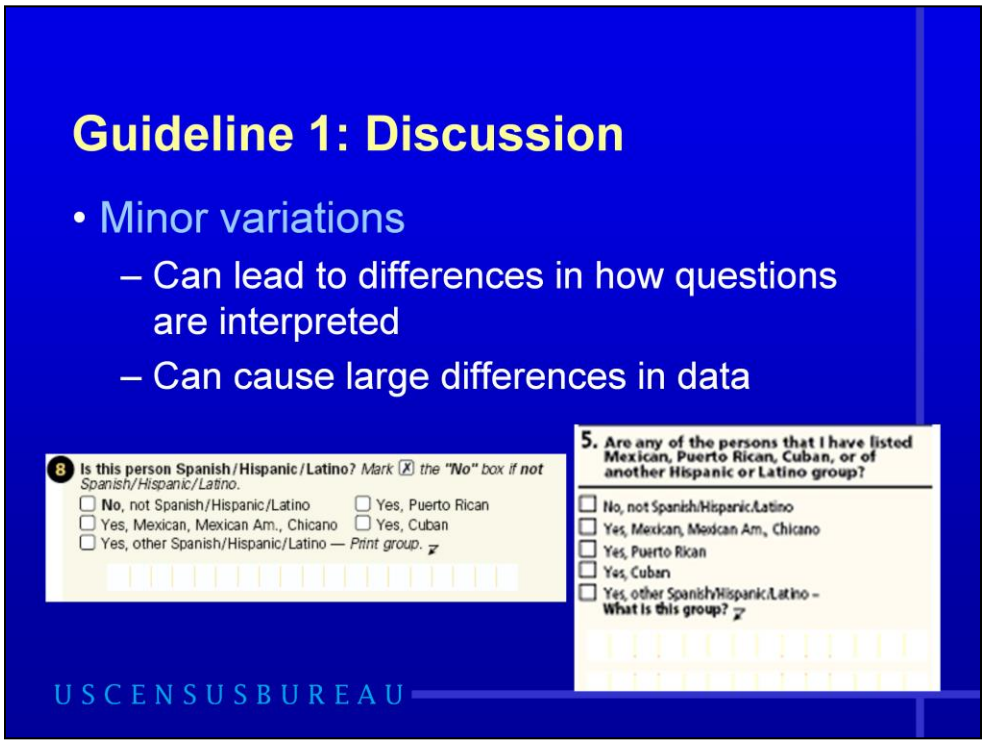


These Guidelines apply the principle of Universal Presentation to eight major aspects of instrument design: question wording and instructions, examples, response categories, formatting of answer spaces, visual design elements, question order and grouping, flashcards, and prompts and help. Some of these Guidelines spell out apparent exceptions to the principle of Universal Presentation. These are situations in which a change in the question wording, order, instructions, or other features is essential for operational reasons, or better preserves the question than would asking exactly the same question in an identical way in different modes.

Question Wording and Instructions Guideline 1 – Part 1

Every effort should be made to maintain the same wording of questions across survey modes.

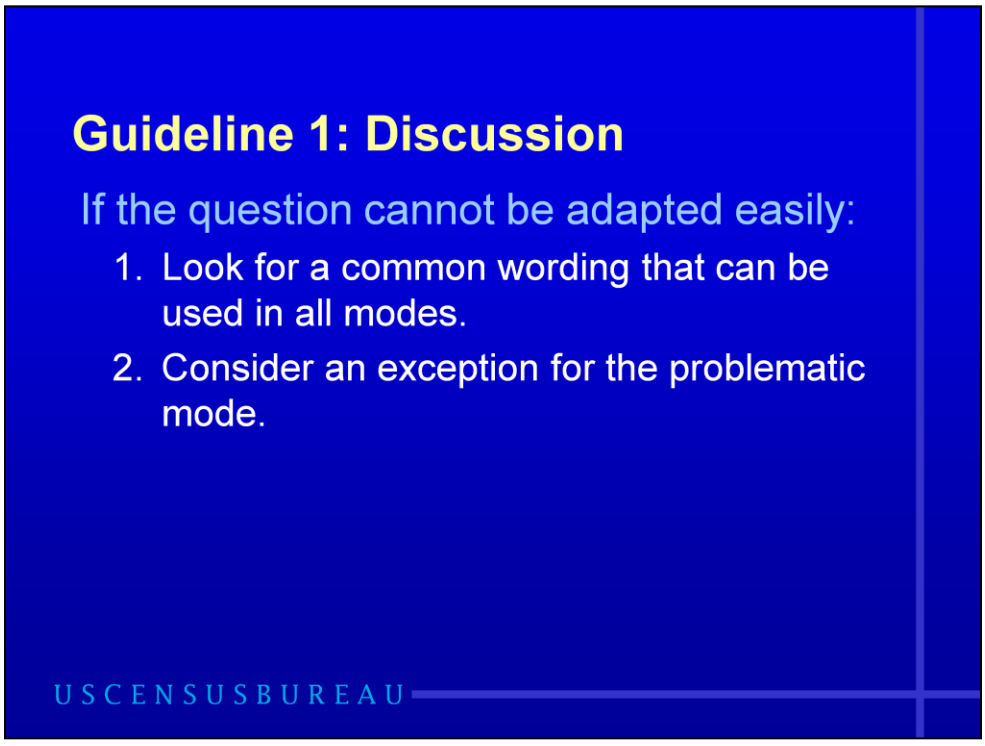
USCENSUSBUREAU



Repeated tests conducted by the Census Bureau and by survey methodologists for nearly 100 years show that seemingly minor variations in question wording can lead to unanticipated differences in how questions are interpreted and answered, resulting in large differences in data for the same question (see, e.g., Bennett and Griffin, 2002; Martin and Gerber, 2003). Thus, it is essential that the same question meaning be conveyed in all modes.

Guideline 1 – Part 2

When variant wording for a particular mode is proposed, the decision to change wording should be based upon evidence that a difference in wording is more likely than the same wording to produce consistent meaning and equivalent answers.

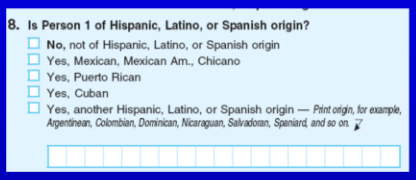


If attempts to use identical wording seem especially troublesome to a particular mode, then two courses of action seem appropriate. One is to look for a common wording that can be used in all modes. When the original wording is so complex or wordy that it cannot be administered the same way in various modes, then the original wording should be simplified to achieve standardization across modes. Often, this may imply simplifying the wording used in the original form. Of course, such revisions should be evaluated, preferably in field tests.

The other possibility is to consider an exception for the problematic mode, but this requires supporting evidence. Sometimes, a change in wording is needed to communicate the intent of a question in different modes and produce equivalent responses in each. An example is the instruction in the race question to, “Mark one or more…”. Cognitive testing for a 2001 census evaluation showed that redundancy was needed for respondents to absorb the one-or-more option in telephone administration (Davis et al., 2001). A slightly reworded question and instruction produced more common understanding of the intent of the question in telephone interviews than would have resulted from asking the identical question that appeared in the mail questionnaire (Martin and Gerber, 2004).

Guideline 1: Example

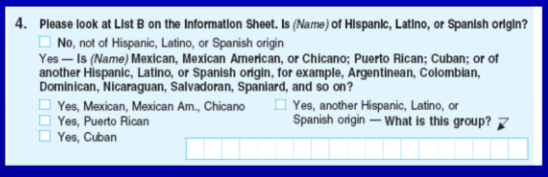
- Self-Administered



8. Is Person 1 of Hispanic, Latino, or Spanish origin?

- ☐ **No**, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

- Interviewer-Administered

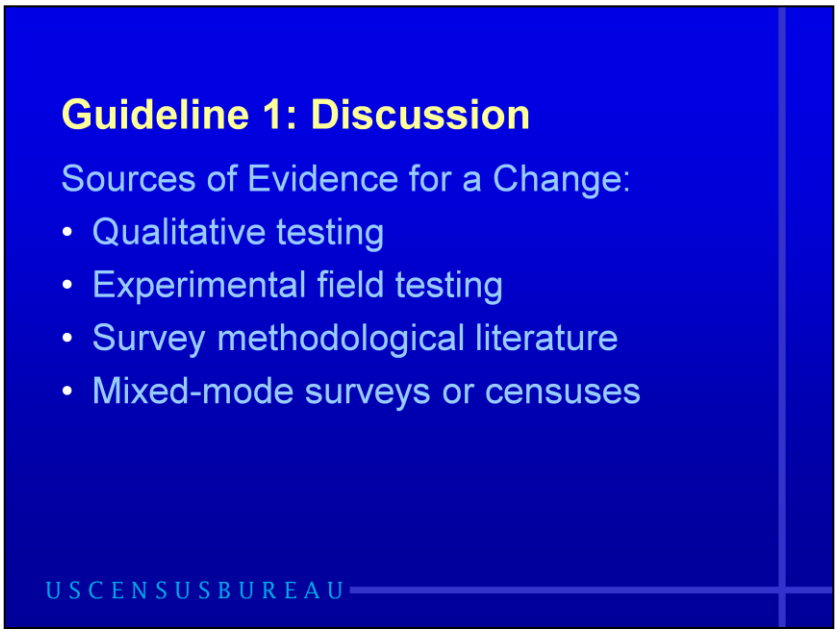


4. Please look at List B on the Information Sheet. Is *(Name)* of Hispanic, Latino, or Spanish origin?

- ☐ **No**, not of Hispanic, Latino, or Spanish origin

Yes — Is *(Name)* **Mexican, Mexican American, or Chicano; Puerto Rican; Cuban; or of another Hispanic, Latino, or Spanish origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on?**

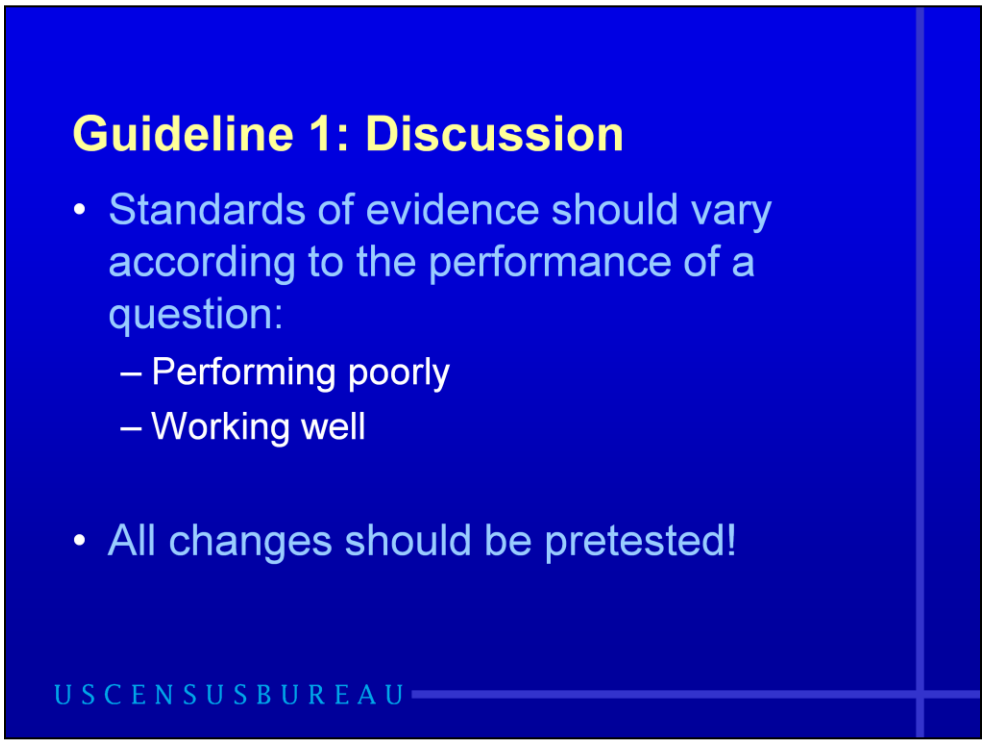
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin — **What is this group?**



The evidence needed to support an exception may come from qualitative testing, field testing, or from survey methodological literature that shows that the same wording or way of asking a question does not produce comparable results in different modes. Ideally, evidence that the same question does not produce the same results across modes would be based on a split-panel field experiment in which respondents were randomly assigned to different modes, but were asked an identically worded question. Lack of data comparability would be indicated by differences in response distributions, for example. Similarly, evidence that different question wordings produce comparable results in different modes would be based on a field experiment in which (for example) treatment 1 involved asking wording #1 in mode #1, and treatment 2 involved asking wording #2 in mode #2, with respondents assigned randomly to one of the two treatments. Use of different wordings in the two modes would be supported if the variant produced the same response distribution and data of similar or higher quality. Ideally, such an experiment would also randomize question wording within each mode, which would be useful for diagnosing problems and sources of mode discrepancies. Practically speaking, such evidence is unavailable, because studies that experimentally assign mode are expensive and rarely done.

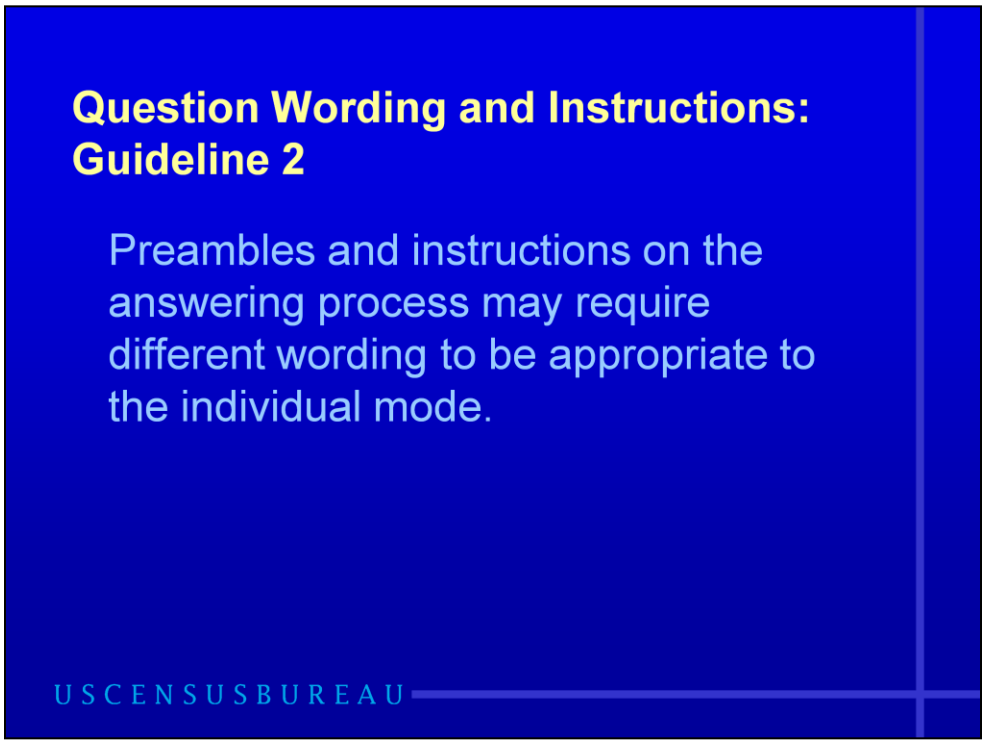
Suggestive, but ambiguous, evidence may come from mixed-mode surveys or censuses. Such surveys frequently produce large differences in data for respondents interviewed using different modes (see, for example, Bennett and Griffin, 2002). However, such differences cannot be unambiguously attributed to the effects of mode. They may instead be due to selection bias, since

respondents select the mode in which they respond.

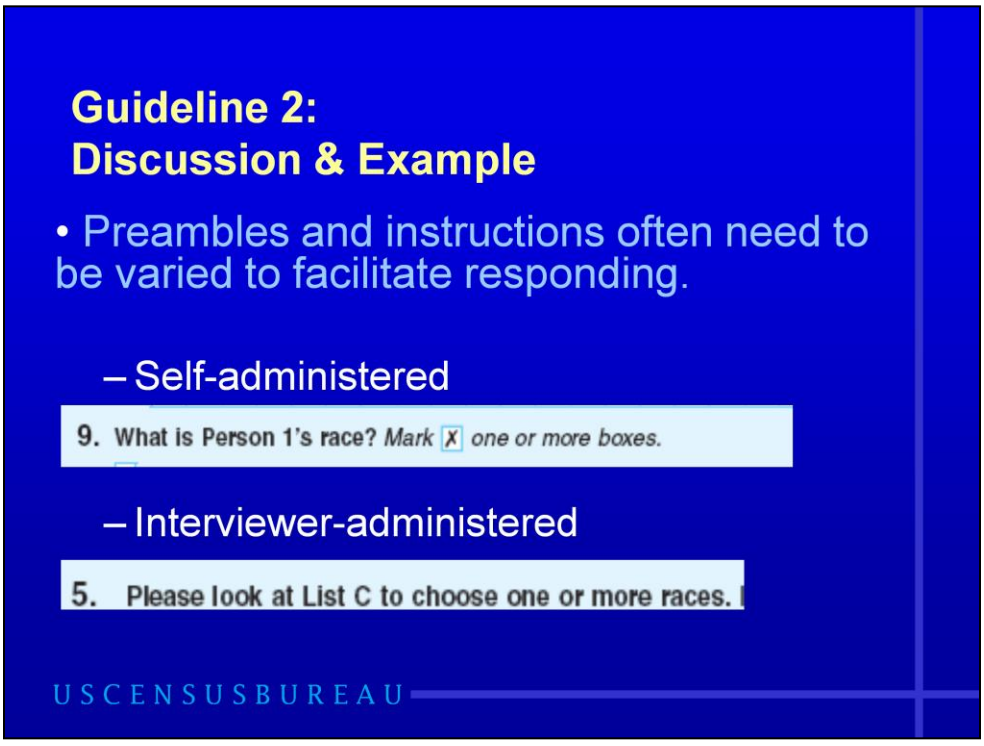


Practically speaking, we think that the standards of evidence required to support an exception to Guideline 1 should vary according to the performance of a question. For example, if testing shows that a question performs poorly in a given mode, then an adaptation may be called for even in the absence of evidence demonstrating that an adapted version is superior and works well. On the other hand, if a question is working well, then it should not be reworded without supporting evidence to show that the revised version is superior or at least equivalent. In other words, "if it's not broke, you have to prove you can do better before you fix it"—but if a question appears broken when administered in a given mode, some tinkering may be needed even when there is not as much evidence as we would like to support the change.

All changes to questions should be tested in some way, in accordance with the Census Bureau's pretesting standard (U.S. Census Bureau, 2003). Substantive changes to questions and categories should be evaluated in a field experiment before being implemented, if possible.



Guidelines 2, 3, and 4 are acceptable exceptions to Guideline 1's requirement for the same question wording across modes.

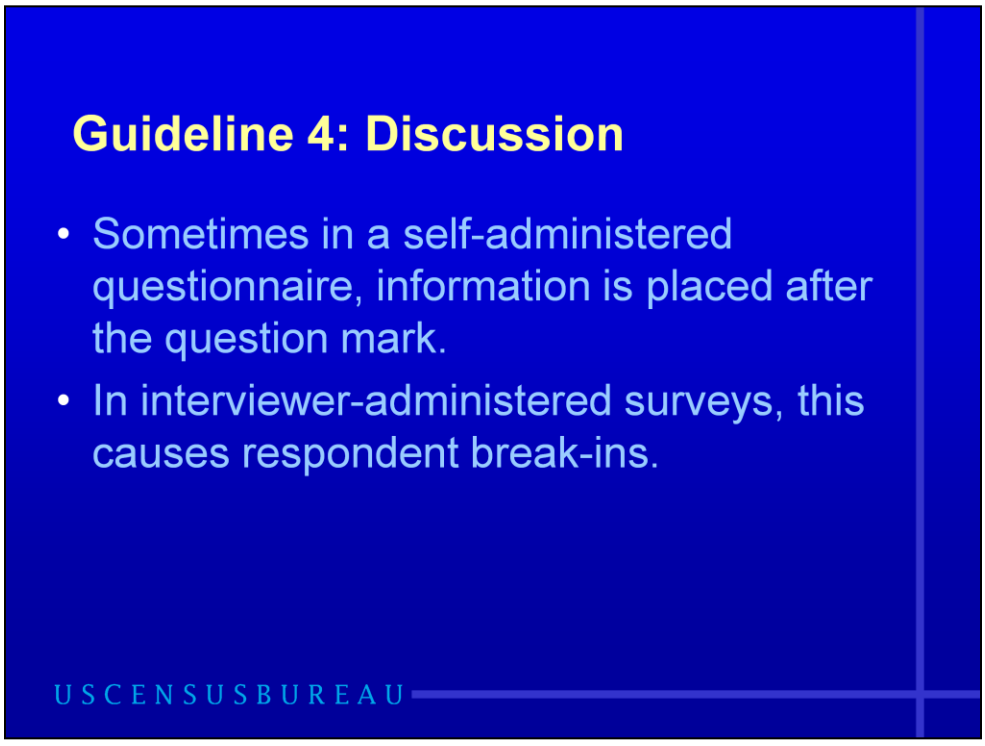


A preamble is an introduction to a question that may include instructions on how to answer that question. Although often it is advisable to strictly follow the same wording of the question across modes, the preamble and other accompanying instructions may, and often should, be varied to facilitate responding in different modes. For example, the preamble to the race question in the CATI mode ("I am going to read a list of race categories…") helps telephone respondents understand they should wait to hear the categories, but is unnecessary in a visual mode when respondents fill out a mail questionnaire or are shown a flashcard.

Similarly, different wording may be needed to facilitate actions that are a part of the answering process. For example, it makes no sense to tell paper questionnaire respondents to click on their answers, to tell web respondents to circle the appropriate answers, or perhaps to give any special directions at all to telephone respondents on how to designate their answers to a telephone interviewer.

Question Wording and Instructions Guideline 4

The elements of a question may be reordered if necessary to ensure that respondents are exposed to all essential information in interviewer-administered modes.



Sometimes in self-administered questionnaires, information (e.g., instructions, examples, response categories) is placed after the question mark of a question. This is not terribly problematic in self-administered surveys because respondents can see that additional text follows the question. However, in interviewer-administered surveys, respondent break-ins occur when additional instructions or examples follow the question proper (Oksenberg, Cannell, and Kalton, 1989). The universal cue that a question has ended and an answer is expected is the rising inflection indicated by a question mark.

In interviewer-administered modes, essential information needs to be placed before the end of a question, or presented another way (such as a flashcard). In interviewer-administered modes, response options, examples or instructions may be incorporated as part of the question, to ensure that respondents are exposed to them before they begin to answer the question. Alternatively response options or examples may be presented on a flashcard.

An example is the overcount question (“Does this person sometimes live or stay somewhere else?”), which is very vague without its response categories (e.g., to attend college, while in the military, in a nursing home). Respondents try to answer before the categories are presented, or interviewers skip some or all of the numerous categories. A possible solution would be to reorder the elements of the question to incorporate some or all categories as part of the question.

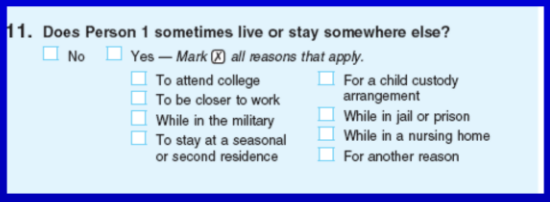
Reordered versions of a question should be tested in the intended mode to ensure they have the same meaning as the original.

Guideline 4: Discussion

- Response options, examples and/or instructions:
  - May be incorporated as part of the question
  - May be presented on a flashcard
- All changes should be pretested!

Guideline 4: Example

Self-Administered



11. Does Person 1 sometimes live or stay somewhere else?

☐ No ☐ Yes — *Mark (X) all reasons that apply.*

- ☐ To attend college
- ☐ To be closer to work
- ☐ While in the military
- ☐ To stay at a seasonal or second residence
- ☐ For a child custody arrangement
- ☐ While in jail or prison
- ☐ While in a nursing home
- ☐ For another reason

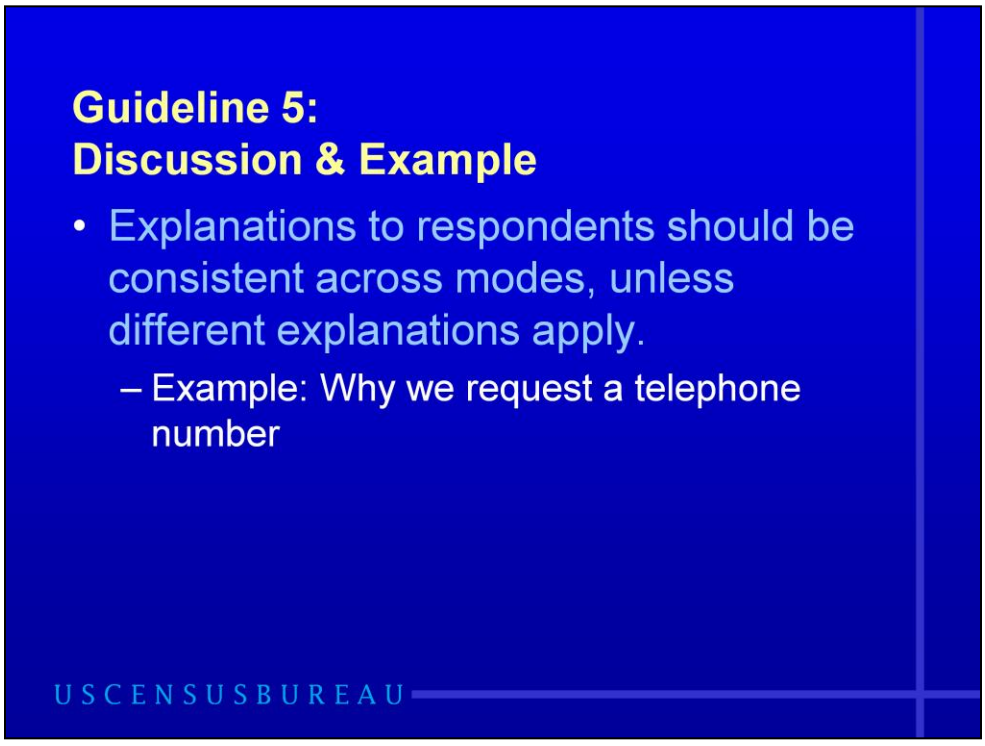
Interviewer-Administered

- *Does Jane ever live or stay someplace else to attend college, to be closer to work, while in the military. . .*

Question Wording and Instructions: Guideline 5

Substantive instructions and explanations should be consistent across modes.

USCENSUSBUREAU



Explanations to respondents (e.g., why we request a telephone number) should be consistent across modes, unless different explanations apply (for example, phone numbers are requested for a different purpose). A substantive preamble (e.g., "We need to count every person living in the United States on April 1, 2010") should be used consistently across modes.

Procedural instructions and preambles do not need to be consistent across modes; see Guideline 2.

Guideline 5: Discussion & Example

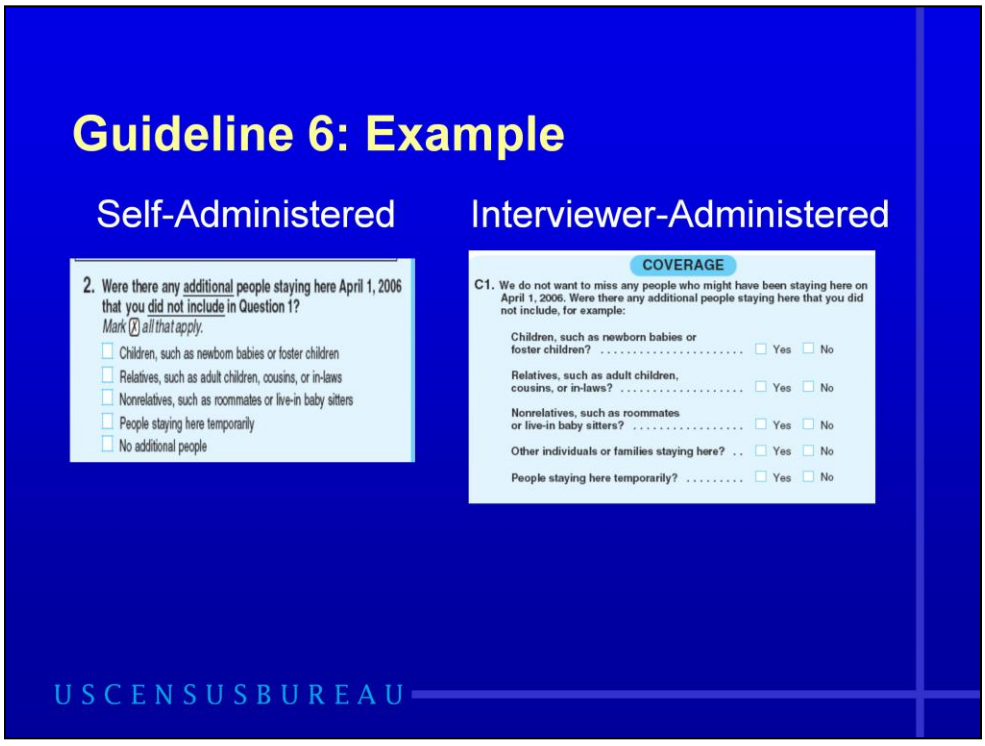
- Substantive preambles should be conveyed consistently across modes.
  - Example - “We need to count every person living in the United States on April 1, 2010.”

Question Wording and Instructions Guideline 6

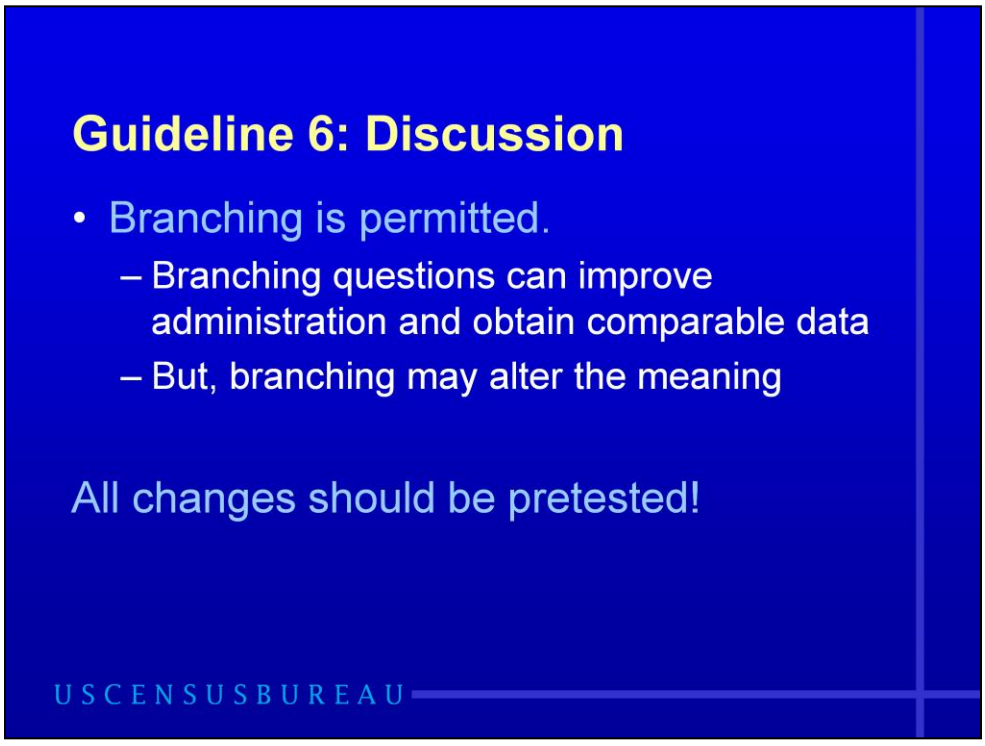
The underlying response task posed by a question should be consistent across modes, even if the implementation differs.

Guideline 6: Discussion

- Asking respondents to perform the same task across modes maximizes the consistency of responses.



Providing respondents with the same wording, and asking them to perform the same task (e.g., select one, mark all that apply) maximizes the consistency of responses to a question across modes. The response task should not be altered as the question is adapted for various modes. Providing a list of categories and asking respondents to "mark all that apply" in a self administered mode is not consistent with a series of "yes/no" questions in interviewer-administered surveys. This alters the task the respondent is asked to perform, and the two tasks do not produce comparable results. Rasinsky, Mingay, and Bradburn (1994) and Smyth et al. (2006) show that reporting is more complete when a series of individual questions ("yes/no") are asked. These two question formats are unlikely to yield consistent data.

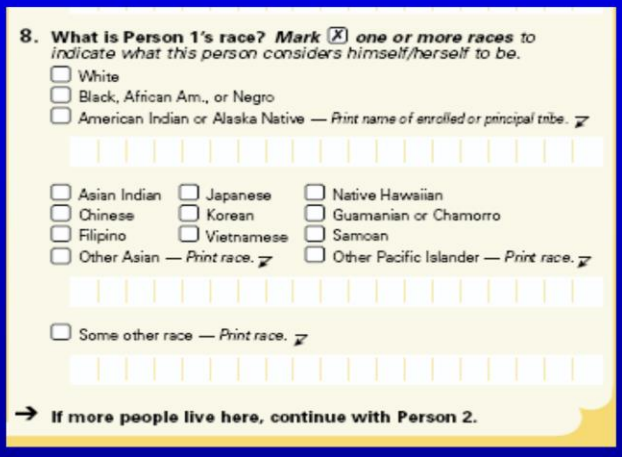


It is permissible under this guideline to create a branching structure for a question with a long list of response categories. The use of branching questions receives support from the survey methodological literature. For example, Krosnick and Berent (1993) find that branching questions obtain more reliable data and are quicker to administer than non-branching versions of questions about political party identification and policy attitudes.

A branching question may be appropriate for the Census Bureau's traditional race question, which, with its 15 categories, is difficult to administer on the telephone. Transforming it into a series of branching questions makes it easier to ask and answer and appears to obtain comparable data (Bentley et al., 2003). However, a branching question may alter the meaning of the question or its categories. An example is the relationship question. In one decennial instrument, respondents were first asked, "Is NAME related to NAME?"; "yes" answers branched to a set of related categories, while "no" branched to nonrelated categories. The problem was that respondents did not define "related" in a uniform way (some included only blood relations,while others included relation by marriage). They needed the response categories to understand the distinction that the Census Bureau intended, so in this case the branching question performed poorly (Hunter and DeMaio, 2004, and Hunter, 2005).

The comparability of a branching question should be pretested before it is implemented, if possible.

Guideline 6: Branching Example



8. **What is Person 1's race?** ***Mark*** ☒ ***one or more races*** *to indicate what this person considers himself/herself to be.*

- ☐ White
- ☐ Black, African Am., or Negro
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

- ☐ Asian Indian
- ☐ Chinese
- ☐ Filipino
- ☐ Japanese
- ☐ Korean
- ☐ Vietnamese
- ☐ Other Asian — *Print race.*
- ☐ Native Hawaiian
- ☐ Guamanian or Chamorro
- ☐ Samoan
- ☐ Other Pacific Islander — *Print race.*

- ☐ Some other race — *Print race.*

→ **If more people live here, continue with Person 2.**

USCENSUSBUREAU

Guideline 6: Branching Example

Please choose one or more categories that best describe your race.

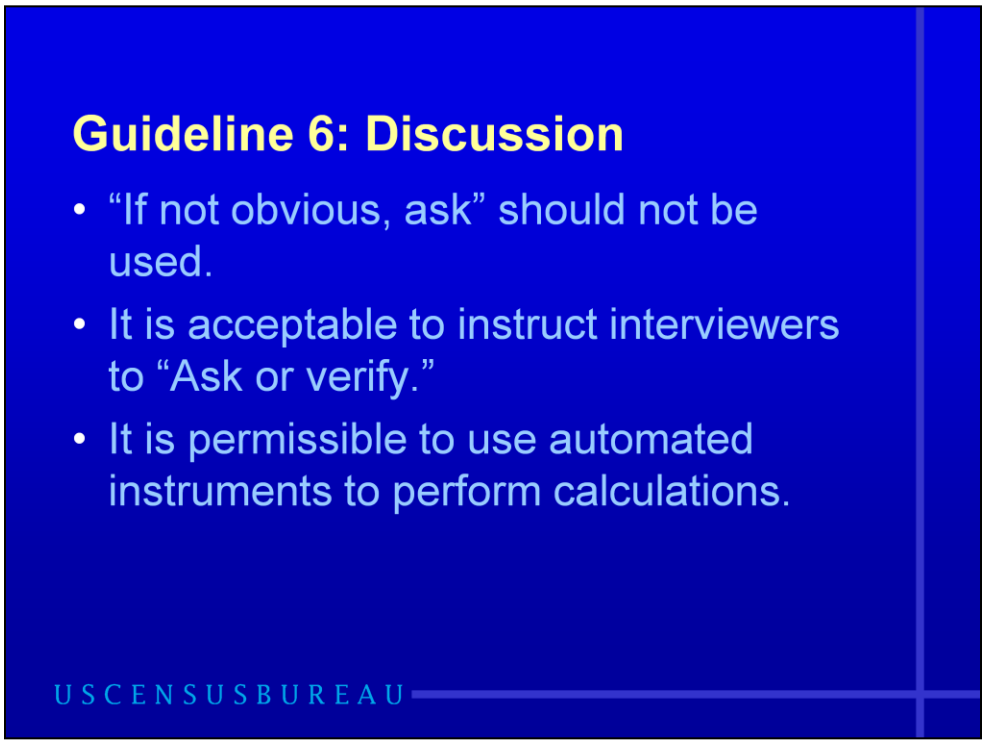
[1] White
[2] Black, African-American, or Negro
[3] American Indian or Alaska Native
[4] Asian*
[5] Native Hawaiian
[6] Other Pacific Islander**
[7] Or, some other race

*To what Asian group do you belong? Are you:

[1] Asian Indian
[2] Chinese
[3] Filipino
[4] Japanese
[5] Korean
[6] Vietnamese
[7]Or, some other Asian group?

**To what Pacific Islander group do you belong? Are you:

[1] Guamanian or Chamorro
[2] Samoan
[3] Or, some other Pacific Islander group?

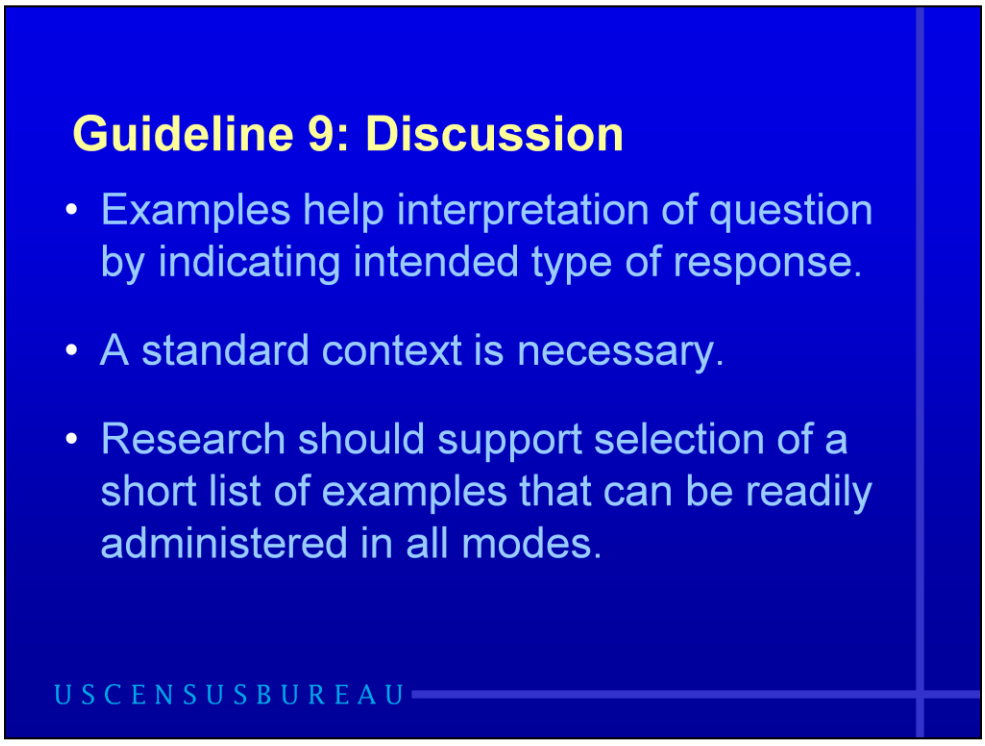


This guideline implies that interviewer instructions of the form, "If not obvious, ask" should not be used. This instruction transforms the item from a question to be answered by respondents into one based on observation by interviewers, which alters the response task in an unacceptable way, and (especially on the telephone) may lead to errors. No question should be reported by the interviewer by observation or inferred through response to another item. However, it is acceptable to instruct interviewers to "Ask or verify." This means that interviewers are allowed to confirm with respondents that (for example) they are male or female, but are not allowed to record an answer based solely on observation. (An interviewer might say, for example, "I'm putting you down as male, is that correct?")

It is permissible under this Guideline to use automated instruments to perform calculations (e.g., calculating age based on date of birth) when that results in better data.

Use of Examples
Guideline 9

The same examples, in the same order, should be presented in a question in all modes.



Examples may help respondents interpret a question—for example, by indicating the type of response that is intended—and they also help respondents find a response that fits them.

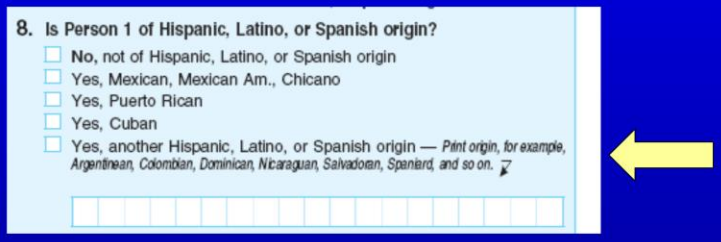
For example, improved reporting of ethnicities when examples are given occurs on the conceptual level, primarily by indicating to respondents the intended specificity of the response. Since it is important to ensure common question interpretation in all modes, the same examples should be presented to respondents in all modes.

Without a visual aid, such as a flashcard, it is difficult to convey a long list of examples in an interviewer-administered mode. When interviewers are required to read a long list of examples, it leads respondents to break in before the list is finished, or interviewers to present only a few examples from the list (Hunter and Landreth, 2005). Thus, in practice, verbal communication of a long list of examples is unlikely to be consistent with visual presentation.

Ideally, choice of a short list of examples should take into account ease of administration in all modes before a final common list is selected. Research should be conducted to support selection of a set of examples that can be readily administered in all modes, and that aid correct interpretation of the question, without distorting responses. Field experiments are needed to show that examples do not bias the data by suggesting particular responses.

Providing a standard context by using the same examples across modes is critical. It is not advisable to have examples in one mode that do not exist in another, unless there is clear evidence of overall improvement in data quality.

Guideline 9: Examples

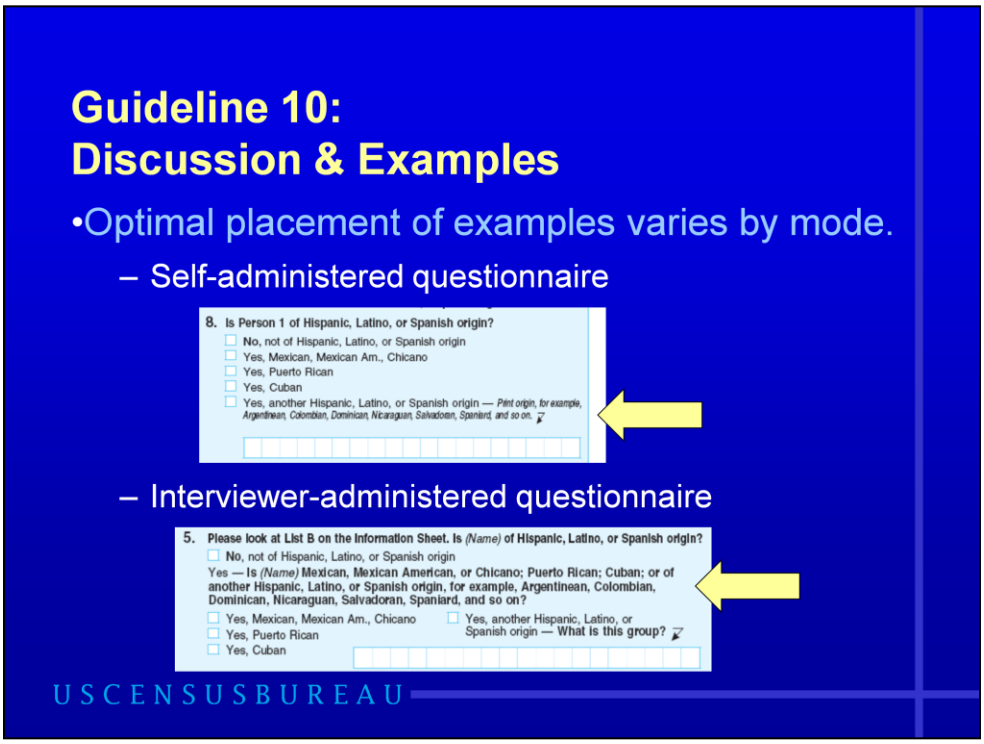


8. Is Person 1 of Hispanic, Latino, or Spanish origin?

- ☐ No, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

Use of Examples: Guideline 10

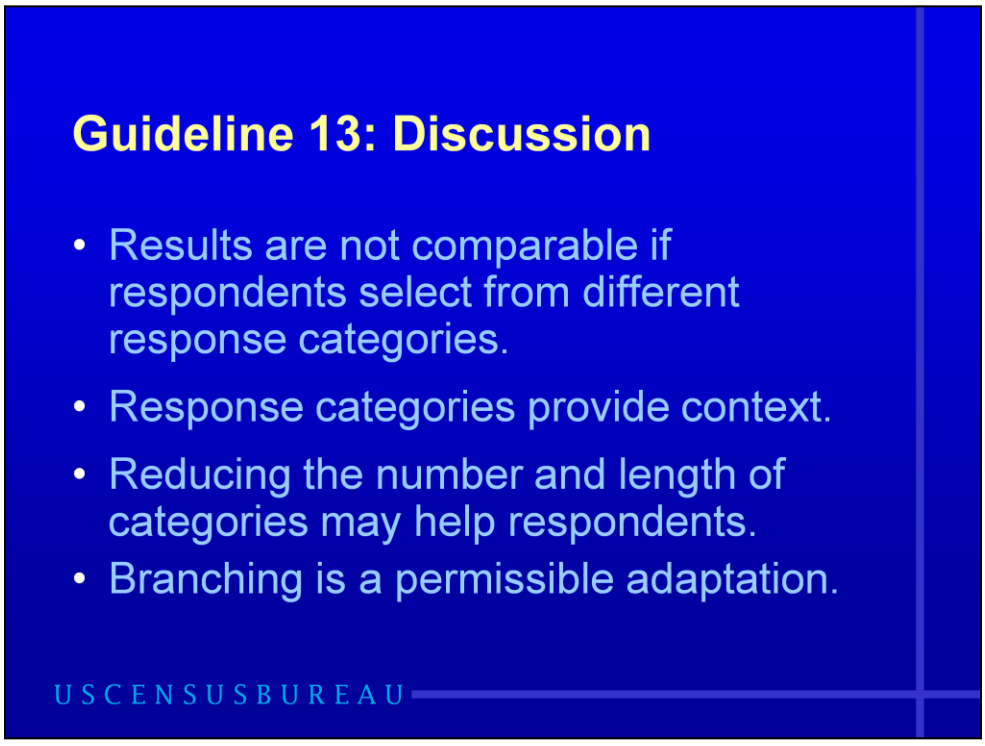
Placement and presentation of examples may vary by mode, consistent with the principle of universal presentation.



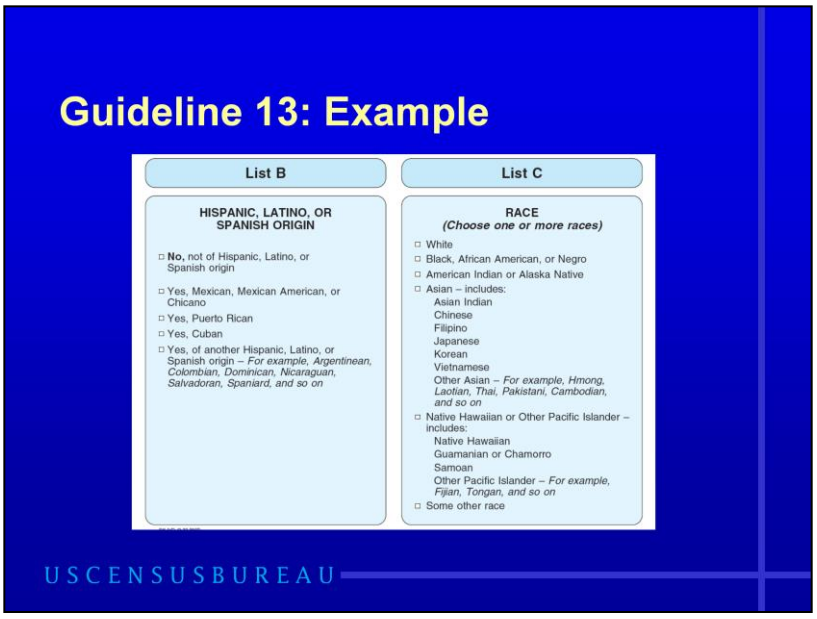
Optimal placement and presentation of examples vary by mode. In a self-administered questionnaire, respondents can read them whether they appear before, after or during the question. In an interviewer administered questionnaire, respondents generally provide an answer as soon as they hear the question. As a result, if examples are placed after the question, respondents may never hear them. It is preferable to incorporate examples in the question itself for interviewer administration, or attach them to a preamble if it is present, or present them on a flashcard if the interview is in person. Choice among these methods of presentation might depend on the number of examples needed to communicate the concept.

Response Categories: Guideline 13

For closed questions, instruments in all modes should be designed and administered to expose respondents to the response categories, in the same order, and to present response categories with the question.



Results are not comparable if respondents in different modes select from different response categories. In interviewer-administered modes, categories should be written the same way they are to be read, and read the same way they are written. Branching questions are a permissible adaptation.



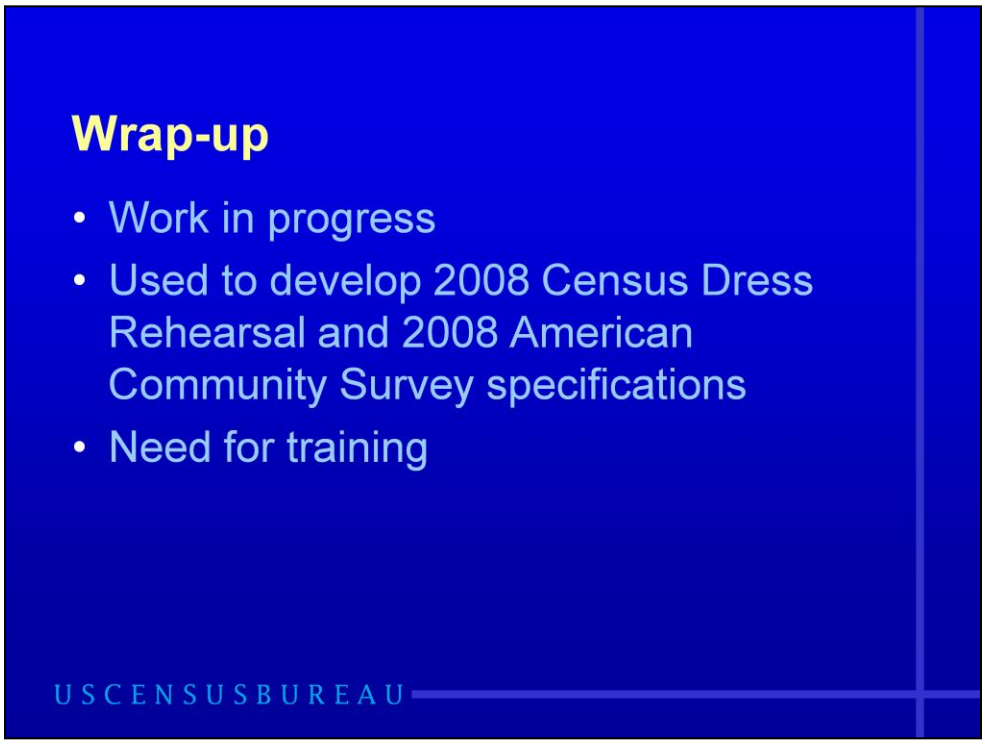
Response categories provide necessary context for respondents to understand the intent of a question. Presenting questions and categories together supports the respondents' use of the response categories to interpret the question, and helps respondents keep the question in mind as they select a category (Couper et al., 2004).

Respondents in self-administered questionnaires may not read all information that is presented, and in aural mode may not retain all response categories in short-term memory. The aim of the designer is to increase the chances that interviewers present the categories to respondents, and that respondents read or hear them and retain them in working memory while formulating a response.

Behavior coding evaluations for the 2004 census test (Hunter and Landreth, 2005) show that enumerators too often (1) fail to present any of the response categories (e.g., overcount question), (2) rephrase or eliminate some categories to shorten a long question (e.g., tenure), or (3) offer additional categories that are not part of the question (e.g., race). These problems reflect the need for enumerator training, but they also reflect flaws in the design of the instruments.

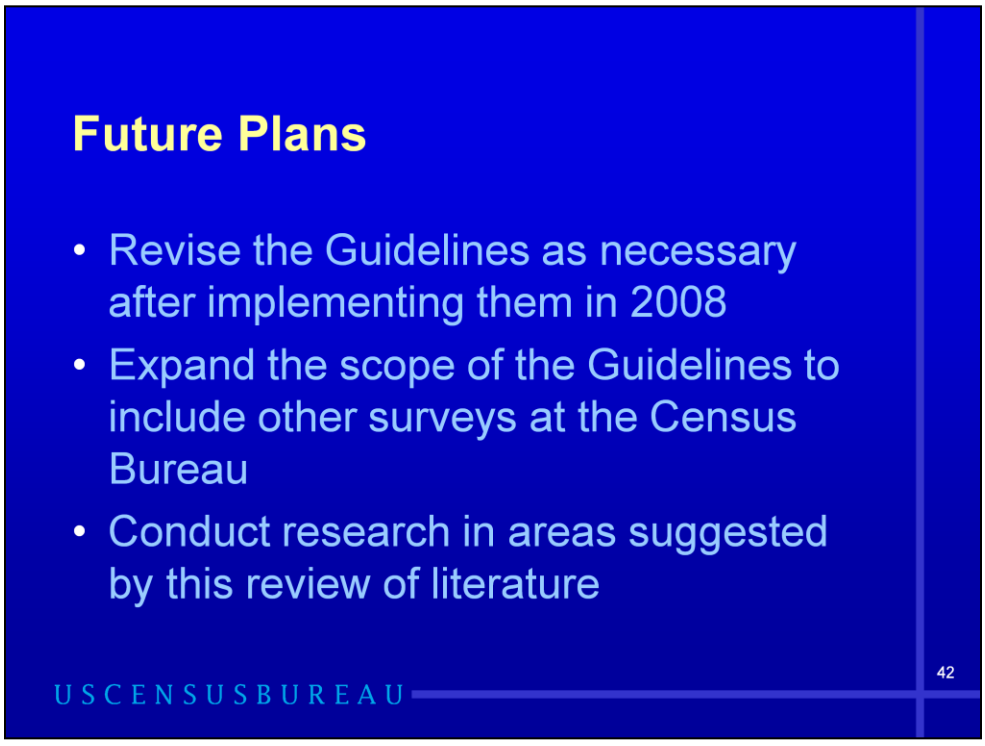
Comparability across modes, and performance in any given mode, may be improved by reducing the number and length of categories, so that respondents may adequately hear (or read) and retain them in all modes.

It is permissible under this guideline to use branching or unfolding questions, which should be considered when the list of categories is long and visual aids cannot be used. (See also Guideline 26.) More research is needed to establish the comparability of data from branched and unbranched versions of the same question.



The compilation of these Guidelines was a momentous task to which quite a number of people across the Census Bureau contributed. Since this type of task had not been completed before, we view them as a work in progress, not as a rigid set of Guidelines that are set in stone. The Census Bureau has been using these Guidelines to develop the final versions of the questionnaires for the 2008 Dress Rehearsal and the 2008 American Community Survey. It has become evident that the staff responsible for developing the questions need to become familiar with the entire set of Guidelines, because issues related to a specific question are addressed in more than one guideline. We plan to conduct training sessions with questionnaire developers to address issues that have surfaced in trying to use the Guidelines. This will serve two purposes: (1) address confusion about how to apply the guidelines; and (2) identify issues that may need to be clarified in future iterations of the guidelines.

In general, these Guidelines have become institutionalized with the staff working on decennial census and ACS development. To this degree, the guidelines have been very successful.



Another portion of our Guidelines publication contains recommended research that, if conducted, will help improve these guidelines Regular review and revision of the guidelines will be critical to keep them up-to-date with findings from new methodological studies.