

This article was downloaded by: [The New School]

On: 28 August 2008

Access details: Access Details: [subscription number 776107991]

Publisher Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Discourse Processes

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t775653637>

Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces

Patrick Ehlen ^a; Michael F. Schober ^b; Frederick G. Conrad ^c

^a Center for the Study of Language and Information, Stanford University. ^b Department of Psychology, The New School for Social Research. ^c Institute for Social Research, University of Michigan.

Online Publication Date: 01 September 2007

To cite this Article Ehlen, Patrick, Schober, Michael F. and Conrad, Frederick G. (2007) 'Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces', *Discourse Processes*, 44:3, 245 – 265

To link to this Article: DOI: 10.1080/01638530701600839

URL: <http://dx.doi.org/10.1080/01638530701600839>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces

Patrick Ehlen

*Center for the Study of Language and Information
Stanford University*

Michael F. Schober

*Department of Psychology
The New School for Social Research*

Frederick G. Conrad

*Institute for Social Research
University of Michigan*

Computer-based interviewing systems could use models of respondent disfluency behaviors to predict a need for clarification of terms in survey questions. This study compares simulated speech interfaces that use two such models—a generic model and a stereotyped model that distinguishes between the speech of younger and older speakers—to several non-modeling speech interfaces in a task where respondents provided answers to survey questions from fictional scenarios. The modeling procedure found that the best predictor of conceptual misalignment was a critical *Goldilocks range* for response latency—that is, a response time that is neither too slow nor too fast—outside of which responses are more likely to be conceptually misaligned. Different Goldilocks ranges are effective for younger and older speakers.

Let us say you are cooking spaghetti in the kitchen and the phone rings, and before you know it you have agreed to answer some questions in a survey interview. The interviewer asks you, “How many hours per week do you usually work at your job?” As a corporate lawyer who aspires to be a professional chef, your answer is

Correspondence should be addressed to Patrick Ehlen, Center for the Study of Language and Information, Stanford University, Cordura Hall, 210 Panama St., Stanford, CA 94305. E-mail: ehlen@stanford.edu

potentially elaborate. You get paid for making calls and eating lunch and playing golf, and you might say you are working even now, practicing for your next career. However, instead of going into all that, you find yourself saying, “Well ... uh ... usually seventy.”

What are you up to? You have made a guess about what you think the interviewer is getting at, while recognizing that your own concept of “work” or “job” or “usually” might be different; that is, you have recognized the possibility of *conceptual misalignment* (Schober, 1998). Instead of giving an elaborate answer that would defy Grice’s (1975) cooperative principle—which sees to it that we do not say much more about things than we think we need to for the task at hand—you answer based on that guess. However, at the same time, you have laid your own awareness of possible conceptual misalignment out on the table for the interviewer to note, using subtle signals that indicate things may not be so straightforward: You hedge (“well”), you pause (“...”), and you insert a filler (“uh”) before you answer, all of which are disfluent behaviors that can signal when a speaker has a problem (Brennan & Williams, 1995; Schober & Bloom, 2004; Smith & Clark, 1993). We will call these signals *misalignment cues*.

Communication happens on two planes: On one plane, you answer the question you think is being asked, providing information that attends to the task at hand; on another plane, you send out some misalignment cues that manage possible subtasks of the dialogue, signaling misalignment information that may or may not be so important but is subtly offered in case it proves relevant (Bangerter & Clark, 2003; Clark, 1996). Now it is up to the interviewer to catch on to those signals and decide if he or she wants to pursue that misalignment problem further.

However, the interviewer’s recognition of your misalignment problem is by no means assured, partly because people speak disfluently for all kinds of reasons, and partly because he or she brings his or her own expectations to the dialogue about what that disfluency might mean. If you are very far into the interview and you have said “well” and “uh” in every answer, by now the interviewer may have decided you are just a person who says “well” and “uh” a lot and may not infer any misalignment at all—that is, he or she creates an *individualized model* of your disfluency behaviors and may only infer misalignment when those behaviors deviate from his or her expectations according to that model. Or, even if this is the first question in the interview, he or she might bring some established assumptions—a *stereotyped model*—that shape his or her expectations about your speech. Did we mention yet that you are 76 years old? Because older speakers are known to be more disfluent (Schow, Christensen, Hutchinson, & Nerbonne, 1978; Shrivastav, Hollien, Brown, Rothman, & Harnsberger, 2003; Yairi & Clifton, 1972), and interviewers act differently with older respondents compared to younger ones (Bradburn et al., 1979), she may be working from a model that tells her that your hedging and pausing and fillering do not indicate any misalignment problem at all, unless your use of those cues deviates from the stereotyped model that shapes her idea of how 70-something-year-old people should speak.

In short, we can speculate that misalignment cues are overdetermined and that, like words in general, their interpretation requires a dynamic and interactive process that depends on the expectations—or speaker models (Schober & Brennan, 2003)—that people bring to a dialogue.

If the interpretation of misalignment cues requires an interactive process, this poses a problem for anyone who wants to use them to automatically diagnose conceptual misalignment problems in survey interviews. Could a machine algorithm listen for certain misalignment cues and then offer clarification of keywords like “work” or “job” when it hears those cues, no matter who is talking to it (i.e., using a simple generic model of misalignment cues)? Or, would it be better if that machine were also tuned in to other aspects of the interaction that people often pay attention to, such as the age of the person it is talking with (using a stereotyped model of misalignment cues) and changed its expectations accordingly?

After all, telephone interviews may one day consist of machine speech interfaces that recognize respondents’ unconstrained speech (see Conrad & Schober, 2007), and working dialogue systems that conduct interviews are already in use (Blyth, 1997; Cole et al., 1994). However, if they are to be as effective as human interviewers, these machines would need to detect and predict cases of possible conceptual misalignment, where respondents need clarification of concepts in a question. Clarification has been shown to improve comprehension—and thus data quality—in survey interviews (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad, & Fricker, 2004). Ideally, that clarification would be targeted and offered only when necessary, without being an intrusive burden, perhaps by modeling expectations about the speaker’s needs. Our purpose here is to explore how that process of modeling and targeting might work in automated speech survey interfaces, comparing the effectiveness of a generic model of misalignment cues (that treats all respondents the same) to a stereotyped model (that distinguishes respondents by age group), and also comparing these modeling procedures to interfaces that rely on other criteria to address problems of conceptual misalignment.

To identify possible misalignment cues in answers to survey questions and derive models that could then be tested for their effectiveness at identifying cases of conceptual misalignment, this study was divided into two experiments: (a) a *model derivation experiment* that collected data on responses to survey questions that were used to derive and validate generic and stereotyped models, and (b) a *model implementation experiment* that tested implementations of the models derived from the first experiment on a new set of respondents.

EXPERIMENT 1: MODEL DERIVATION

We first wanted to know how misalignment cues are used by two different age groups when responding to survey questions about conceptually ambiguous content. Our goal was to derive two separate models based on that usage: a generic

response model that assumes all participants will show similar response behaviors regardless of their ages, and a stereotyped response model that takes advantage of any differences we find in the response behaviors of the two groups. These models were implemented and tested in Experiment 2. To derive these models, we first ran two conditions that provided behavior data for the models. A third condition served as a control condition and also as a testing data set, used to run a preliminary validation of the models before implementing them in Experiment 2.

Method

Participants. Sixty participants were recruited by advertisement in the New York City area and assigned to one of two age groups distinguished by age: One group of 30 participants ranging from ages 65 to 80 constituted an “older” group; the other group of 30, ranging from 21 to 39, constituted a “younger” group. These were randomly assigned across three conditions, with 10 participants per group per condition. Genders were balanced across conditions, with 5 men and 5 women from each age group participating in each condition. All were native English speakers. Participants were asked if they considered themselves to have “average sight ability” and “average hearing ability.” In addition, at the outset of the experiment they were asked to read some numbers and instructions from the front page of a packet to test their ability to hear and comprehend the artificial voice (explained later) and to read the materials.

Materials. Participants completed a referential communication survey task similar to the design used by Schober and Conrad (1997), answering a selected set of 12 survey questions from existing, ongoing U.S. government surveys. Rather than responding about their own lives, they responded about fictional scenarios so the accuracy of their answers could be determined. Rather than being interviewed by a person, participants answered questions presented by a computer over a standard telephone. All participants were asked the same questions, although scenarios varied according to how well potentially ambiguous *key concepts* in the scenarios mapped onto the questions asked.

The method by which participants received clarification about those potentially ambiguous key concepts also varied and served as the independent variable that distinguished the three model derivation conditions: (a) a baseline, *no clarification* condition provided no clarification; (b) a *respondent-initiated clarification* condition provided clarification only when explicitly requested by the participant; and (c) a *required clarification* control and validation condition provided clarification for every question.

Questions and scenarios. Each participant was asked to answer 12 survey questions taken from ongoing survey interviews conducted by the U.S. Bureau of

Labor Statistics (BLS) in three different domains: 4 questions about housing came from the Consumer Price Housing Index survey (e.g., “How many people live in this house?”), 4 questions about work situations came from the Current Population Survey (e.g., “How many hours per week do you usually work at your job?”), and 4 questions about purchases came from the Current Point of Purchase Survey (e.g., “Have you purchased or had expenses for household furniture?”), all of which have been used in prior studies (e.g., Schober & Conrad, 1997). For each question, survey designers at BLS had already developed official definitions for key concepts designed to clarify whether, for example, a “floor lamp” should be counted as a piece of “household furniture,” which helped us to infer whether a participant’s understanding of the key concept in a question matched the intended definition. In other words, it allowed us to measure conceptual alignment between the participant and the question designer.

The official definitions sounded something like this:

Let me give you our definition of household furniture. Include tables, chairs, footstools, sofas, china cabinets, utility carts, bars, room dividers, bookcases, desks, beds, mattresses, box springs, chests of drawers, night tables, wardrobes, and unfinished furniture. Do not include TV, radio, and other sound equipment, lamps and lighting fixtures, outdoor furniture, infants’ furniture, or appliances.

Participants answered these questions while looking at fictional scenarios in which the key concepts were manipulated, so the response accuracy of participants’ answers could be used to infer conceptual alignment. Each question had two alternate scenarios: either a *straightforward mapping*, where the description in the scenario mapped onto the key concept of the question in a clear and simple way; or a *complicated mapping*, where the description in the scenario mapped onto the question in a way that was more open to interpretation, making it hard to answer correctly without information about the official definition of the key concept. For example, a participant who is asked, “Has Kelly purchased or had expenses for household furniture?” might see a straightforward scenario that shows Kelly’s receipt for the purchase of an end table—which most people would interpret as household furniture—or instead might see a complicated scenario with a receipt for a floor lamp, which some may interpret as household furniture, although the official definition states that floor lamps should not be counted.

One half of the scenarios seen by each participant were straightforward mappings, and one half were complicated mappings.

Procedure. Participants sat at a desk in a laboratory with a packet of scenarios and a regular Bell-style telephone and were asked to dial a number when they felt familiar enough with the scenarios to answer questions about them. Although participants were told they would be interviewed by a computer, the number they di-

aled was actually answered by a computer in the next room that was controlled by an experimenter, who used a Wizard-of-Oz telephony interface to interact with the participant. Using this interface, the experimenter would present the questions and reply to the participants' answers using audio files that were prerecorded using a synthesized text-to-speech voice. The experimenter interface was created in Visual Basic and used shortcut keys to trigger responses. The experimenter, after some practice, could listen to a participant's answers and respond quickly. A counter on the interface displayed the number of milliseconds that passed after each question was asked.

In the *no clarification* condition, participants heard a question, provided an answer, and immediately moved on to the next question. In the *respondent-initiated clarification* condition, participants were told they could ask for clarification of terms if they felt like they needed it. If they asked for clarification, the synthesized voice read the official definition of the key concept, and then asked the question again. In the *required clarification* condition, clarification was provided for every question. After participants answered a question they automatically heard the scripted definition and were presented with the question a second time, allowing them to change their answers if they wished. This pattern of question–answer–clarification–requestion–reanswer, although unnatural as an example of dialogue, was done so we could know how each participant in the required clarification condition would answer a question both *without* and *with* the benefit of clarification, providing a platform to later test and validate the models derived from the data obtained in the no clarification and respondent-initiated clarification conditions on an independent data set.

After participants completed the interview, they filled out a paper-and-pencil questionnaire that collected demographic information and also asked about their experience during the interview. Some questions were qualitative, open-ended questions, and others asked the participant to mark satisfaction ratings on a 7-point scale. A composite satisfaction score was averaged from two of the scales that asked participants to rate the degree to which they found the clarification useful and the degree to which they found it desirable.

In sum, our model derivation experiment collected measures of potential misalignment cues taken from responses in the no clarification and respondent-initiated clarification conditions and used these to derive models of the behaviors that best predicted conceptual misalignment, inferred through response accuracy. Measures from responses collected in the required clarification condition were then used to validate those models

Predictors used. To derive the models, the following misalignment cues immediately following each question were pooled from responses in the no clarification and the respondent-initiated clarification conditions and tallied for use as po-

tential predictors of response accuracy: response latency, fillers, hedges, restarts, repeats, repairs, reports, and mutters. The criteria for coding these behaviors were adapted from Bortfeld, Leon, Bloom, Schober, and Brennan (2001). Response latency, in particular, was measured as the time in milliseconds from the end of the question to the beginning of the respondent's first utterance—whether an answer, non-answer, filler, or other verbalization, not including sighs, breathing, or other nonverbal mouth noises.

One other potential cue was added early in the experiment after participants were seen performing a consistent behavior we had not considered: Although most answers were straightforward and unelaborated (such as “yes” or “fifty”), some responses also repeated a word or words from the question that had just been asked that often included the key concept word(s). This behavior can be seen as a joint action (Clark, 1994) that “picks up” any potentially ambiguous words from a question and “keeps them in play” in the dialogue, allowing them to be confirmed and negotiated further by both parties, if necessary.

Consider the question, “How many hours per week does Mindy usually work at her job?” A response like “fifty” does not invite any further negotiation of the terms of the question and does not offer any recognition that the word “usually” is open to interpretation. However, a response like, “Usually, fifty” picks up the term “usually” as a way of keeping it in play so it can be confirmed or negotiated. Such behavior may show some awareness by the respondent that a concept is open to interpretation. We could call this behavior a *referential confirmation pick-up*; or, more simply, a *confirmation pick-up*.

The cues mentioned earlier (response latency, fillers, hedges, restarts, repeats, repairs, reports, mutters, and confirmation pick-ups) were all used as predictors to derive models of misalignment profiles to be used in the generic and the stereotyped modeling conditions in Experiment 2, using an ordinary least squares multiple regression to determine the best predictors of conceptual misalignment, inferred through response accuracy (A) as the criterion variable. The regression equation for each model began with factors for response latency (L), fillers (F), hedges (H), restarts (RS), repeats (RP), repairs (RA), reports (RO), clarification requests (CR), repeat requests (RR), confirmation pick-ups (CP), and mutters (M), creating the following initial least squares model:

$$A = \beta_1 L + \beta_2 F + \beta_3 H + \beta_4 RS + \beta_5 RP + \beta_6 RA + \beta_7 RO + \beta_8 CR + \beta_9 RR + \beta_{10} CP + \beta_{11} M + e$$

Coefficients (β) for each of these potential cues were determined that yielded the smallest residual constant (e), eliminating factors by backward variable elimination (p -in = .05, p -out = .10). This procedure allowed us to ascertain the most significant predictors of response accuracy.

Results

Let us first discuss whether we find any differences in misalignment cue behaviors between older and younger participants, and then examine the outcome of our modeling procedures. Additional results about overall response accuracy and time taken to answer each question in these conditions are discussed in conjunction with the results of Experiment 2, to facilitate comparison.

Age-group differences in disfluency rates. The first question we asked was whether there were indeed significant differences in disfluency rates between the two age groups. Observations from the no clarification and the respondent-initiated clarification conditions reveal that older participants were, in fact, more disfluent than younger ones, in keeping with previous findings (Bortfeld et al., 2001; Schow et al., 1978; Shrivastav et al., 2003; Yairi & Clifton, 1972). When compared to the younger group, the older group provided significantly more fillers, $F(1, 78) = 7.43, p = .008$; restarts, $F(1, 78) = 4.94, p = .029$; repeats, $F(1, 78) = 19.58, p < .001$; repairs, $F(1, 78) = 6.48, p = .013$; reports, $F(1, 78) = 9.95, p = .002$; repeat requests, $F(1, 78) = 9.31, p = .003$; and mutters, $F(1, 78) = 6.09, p = .016$.

Response latency and the Goldilocks ranges. Contrary to what we expected to find, the raw measure of response latency initially did not show any linear correlation with accuracy (e.g., longer latencies predicting greater inaccuracy), and it did not show significant differences between age groups. Thus, it was dropped from the initial least squares model during variable elimination in the multiple regression. This observation was puzzling because it stands to reason that less trouble should equal less thinking time, whereas more trouble should call for more time. When intuition runs counter to regression results, it falls under what Johnston (1972, p. 221) called “*a priori* information” that deserves further scrutiny, so we considered another possibility: Perhaps responses that are too fast are just as likely to be problematic as responses that are too slow, leaving a normal range of latency that is “just right” for predicting that people may provide an accurate response, but outside of which they are more likely to provide an inaccurate one. That range can be thought of as a *Goldilocks range* for response latency as a predictor of accuracy.

We tested for a Goldilocks range by making an educated guess about what the range might be, starting with 1 *SD* from either side of the mean response latency and replacing our response latency variable ($\beta_1 L$) with a Goldilocks variable that designates whether the latency of each case falls outside of that range ($\beta_1 G$). The fit of subsequent test models could then be evaluated and compared to the initial model by comparing the adjusted coefficient of determination (adjusted R^2) for each first-pass regression model using all predictors (Draper & Smith, 1981). When our 1 *SD* Goldilocks model showed a better fit than the initial model, we

solved for latency threshold values that maximized adjusted R^2 in each model.¹ These Goldilocks models provided better fitting first-pass models for predicting accuracy than the initial model (adjusted R^2 of $-.041$, $p = .610$) in both the generic (adjusted R^2 of $.003$, $p = .454$) and stereotyped (adjusted R^2 of $.212$, $p = .048$) cases.

The Goldilocks range determined from responses from all participants without regard to age group yielded a range between 2 and 7.2 sec, which is a range that can be used to help derive a generic model. Analysis of the two age groups using two independent Goldilocks factors (one for younger respondents and another for older) revealed a Goldilocks range of 4.6 to 10.2 sec for the younger group and 2.6 to 4.35 sec for the older group. Although these two stereotyped ranges bring to light a clear difference in how the two groups answer questions, they do not support the prediction that “older people take longer to answer” in general. Rather, the response latency range in which older people can be expected to provide a conceptually aligned answer is attenuated, and also shifted to a much *faster* response time than we see for the younger group (perhaps indicating that older respondents apply a different kind of knowledge or answering strategy).

Generic and stereotyped models. We employed an ordinary least squares multiple regression using backward variable elimination of all potential predictors to derive models for predicting response accuracy. These regressions revealed that, in fact, a participant’s failure to respond within the response latency Goldilocks range was the single remaining significant predictor of inaccurate responses for both the generic model and the stereotyped model. Confirmation pick-ups surfaced as the second-most enduring predictor in both regression models, although they predicted *accurate* responses rather than *inaccurate* ones. At $p = .15$ in the generic and $p = .23$ in the stereotyped model, confirmation pick-ups did not reach the criterion of significance to be included in the final regression, although a repeated measures ANOVA showed confirmation pick-ups as predictive of accurate answers for older respondents, $F(1, 38) = 5.28$, $p = .027$; and also predictive of complicated mappings for both age groups, $F(1, 38) = 4.90$, $p = .033$.

Therefore, our models of both the generic and the stereotyped cases used only the Goldilocks range factors to predict conceptual misalignment. The generic misalignment model used the generic Goldilocks range factor, whereas the stereotyped model used the two stereotyped Goldilocks range factors. Responses that fell within that range were more likely to be aligned and did not warrant offers of clarification, whereas responses falling outside the range were more likely to be misaligned, and therefore warranted clarification.

¹These maxima can be derived through the adroit touch of calculus or through the indelicate clobbering of an algorithm that tests small contiguous intervals within a reasonable range. We used the latter.

Validation of models. Because the required clarification condition solicited two responses from participants—one that came before the participant heard clarification and one that came after—we knew how respondents would have answered both with and without clarification. Any model could thus be tested against this independent data set by looking at cases where the model in question would have predicted a need for clarification and using the post-clarification response for those cases, while using the pre-clarification response for cases where the model did not predict a need for clarification (in essence, analyzing accuracy by treating the data as if clarification had not been provided in the instances where the model would not have predicted that clarification was necessary). Therefore, a single, unvarying data set could be used to test and compare the performance of any number of models at predicting accuracy.

By knowing how participants answer both with and without clarification, we can evaluate how well each model predicts actual *improvement* in respondents' answers. In other words, when participants changed their answers from an inaccurate to an accurate one after they received clarification, was that improvement in keeping with factors the model would have used to predict misalignment, or was that improvement associated with other factors that the model did not take into account? In a sense this analysis tells us more about predicting conceptual misalignment than by looking at just response accuracy because it informs us about the degree to which increases in accuracy in a modeling condition can actually be attributed to the modeling itself as opposed to whether any increase in accuracy is simply the result of other residual, non-modeled factors, such as participants receiving more clarification in the modeled condition than they would otherwise have received. Actual implementations of the models cannot provide such information.

For the generic model, when participants made an inaccurate response, the model predicted 53.3% of these as responses that came either too slow or too fast (a measure of model-predicted misalignment). For complicated mappings only, it predicted 54.2% of the inaccurate responses. When participants changed their answers on complicated mappings from inaccurate to accurate ones after hearing clarification, cases where the generic model predicted a need for clarification led to improvement (a measure of model-predicted improvement) 56.5% of the time. This can be opposed to cases where the model predicted *no* need for clarification and yet participants' answers improved anyhow (a measure of residual improvement), which occurred 43.1% of the time. To assess how well improvement can be attributed to the factors used in the model versus other non-modeled factors, we can compare these improvement percentages as population proportions; their differences are marginally significant at $p = .070$.

The stereotyped model predicted 83.3% of participants' inaccurate responses overall. For complicated mappings only, 85.5% were predicted by the stereotyped model. Model-predicted improvement of 53.5% was reliably higher than the residual improvement of 33.3%, $p = .038$.

TABLE 1
Validation Results for Stereotyped and Generic Models

Variable	All Mappings		Complicated Mappings	
	Generic	Stereotyped	Generic	Stereotyped
Alignment				
Model-predicted misalignment	53.3%	83.3%	54.2%	85.5%
Model prediction misses	55.4%	59.2%	45.8%	33.3%
Improvement				
Model-predicted improvement	27.3%	28.2%	56.5%	53.5%
Residual improvement	28.7%	26.3%	43.1%	33.3%
<i>p</i> of predicted versus residual	.408	.404	.070	.039

While these results, summarized in Table 1, show promising predictions of improvement on complicated mappings and provide some validation of our generic and stereotyped models, these data were collected in a different discourse context than what might occur under actual implementations of the models. That is to say, the discourse context is different because the required clarification condition data used for these validation studies required participants to hear clarification after every question, which could influence their answering behavior (by, for instance, implying that their initial answer was wrong when in fact it was not). Because these possible effects may not affect verbal behavior in the same way when participants are faced with an actual implementation of a generic or stereotyped model, we must ask if similar gains in improvement would also extend to implementations of the models, which can only be addressed with a second experiment.

EXPERIMENT 2: MODEL IMPLEMENTATION

The model implementation experiment implemented and tested the generic and stereotyped models derived in Experiment 1. It consisted of the *generic respondent model* and the *stereotyped respondent model* conditions, which respectively used the generic and stereotyped models to predict whether a response to a question may be conceptually misaligned and in need of clarification.

Method

Participants. Another 40 participants were recruited, and again assigned by age to one of the two age groups, ranging from 65 to 81 for the “older” group and from 21 to 39 for the “younger” group. They were randomly assigned across two

conditions, again with 10 participants per group per condition. Genders were balanced across conditions, and all were native English speakers.

Procedure. Materials and procedures were identical to those of Experiment 1, except in the method used to provide clarification in each of the two conditions. In these two conditions, the models derived from Experiment 1 determined whether the participant's behavior in answering each question warranted clarification. In both conditions, respondents listened to a question and provided an answer, and were then given clarification if they requested it or if their behavior met the criteria of the respondent models for that condition—that is, if their answers fell outside the Goldilocks range for that condition's model. An offer of clarification was followed by a chance to answer the question a second time.

Because a primary concern in these two conditions was to respond with clarification according to whether the onset of an answer fell within the Goldilocks range, we outfitted our experimenter Wizard-of-Oz interface with a “big light” that was pre-programmed with the latency ranges for each condition. The light shone green when response latency time passed within the range and shone red outside of it, making it easy to gauge where the participant's response fell in terms of response latency and to determine whether the experimenter should trigger clarification.

Results

Response accuracy. Do models that use generic or stereotyped Goldilocks ranges actually help to reduce conceptual misalignment and improve response accuracy when compared to the other non-modeled conditions? The short answer is that they do: Both modeling conditions result in significantly higher accuracy on complicated mappings than not modeling at all, with modeling conditions showing accuracy ratings that are not statistically different from providing clarification for every question.

Although overall mean accuracy for complicated mappings was only 52%, the differences in response accuracy for complicated mappings varied greatly by condition ($F_1(4, 95) = 35.87, p < .001$ and $F_2(4, 55) = 19.40, p < .001$), as shown in Figure 1. This overall effect was due primarily to a reliable difference between the respondent-initiated clarification and generic respondent model conditions. With no clarification at all, accuracy on complicated mappings reached only 20%. When participants were allowed to ask for clarification, accuracy rose, although not reliably, to 28%: $F_1(1, 95) = 1.46, ns$ and $F_2(1, 55) = 0.80, ns$.

However, for the generic modeling condition where all participants were offered clarification according to the same Goldilocks range, accuracy on complicated mappings reached a reliably higher 64%: $F_1(1, 90) = 35.03, p < .001$ and $F_2(1, 55) = 19.09, p < .001$. When different Goldilocks ranges were tailored for the respective younger and older groups, accuracy in the stereotyped condition

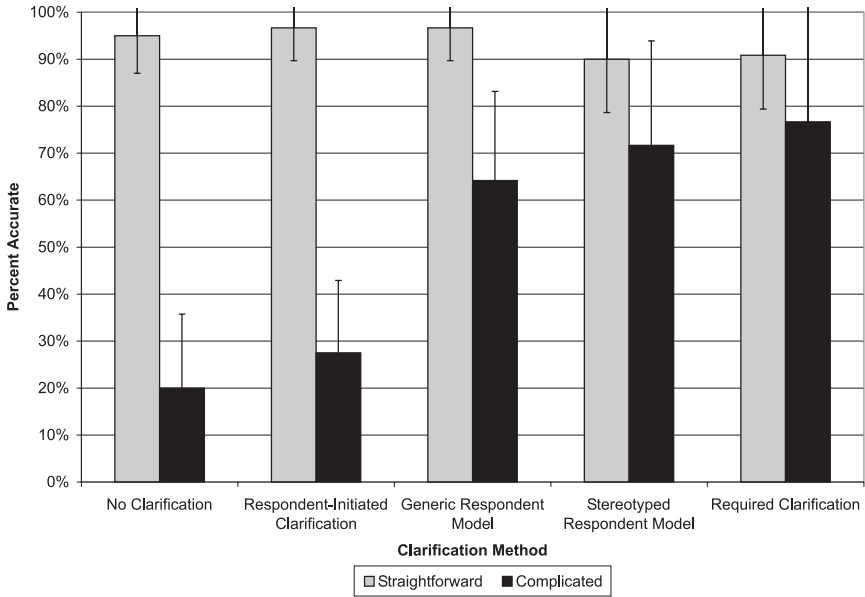


FIGURE 1 Response accuracy by condition for all ages.

reached 72%: $F_1(1, 90) = 1.46, ns$ and $F_2(1, 55) = 0.80, ns$. The highest accuracy on complicated mappings came when all participants heard clarification every time, at 77%, which was not reliably different from accuracy in the stereotyped condition: $F_1(1, 90) = 0.65, ns$ and $F_2(1, 55) = 0.35, ns$. Although these accuracy rates for the generic respondent model, the stereotyped respondent model, and the required clarification conditions do not differ significantly from one another, they do reflect a significant linear trend, $F_1(1, 90) = 128.10, p < .001$ and $F_2(1, 55) = 68.89, p < .001$, in which the higher rates of clarification afforded by increasingly fine-tuned modeling (or from providing clarification for every question) lead to higher accuracy.

Not surprisingly, and consistent with prior findings (Schober & Conrad, 1997; Schober et al., 2004), participants fared very well overall at answering questions to straightforward mapping scenarios (mean accuracy of 94%), which varied little across conditions and age groups.

Although stereotyped modeling did not lead to reliably better overall response accuracy on complicated mappings than generic modeling, there was a marginal difference by age group of participant, $F_1(1, 90) = 3.49, p = .065$ and $F_2(1, 110) = 2.58, ns$, where older respondents fared marginally better from stereotyped modeling (see Figure 2), showing an accuracy of 75% as compared to only 57% in the generic modeling condition, $F_1(1, 45) = 3.79, p = .058$ and $F_2(1,$

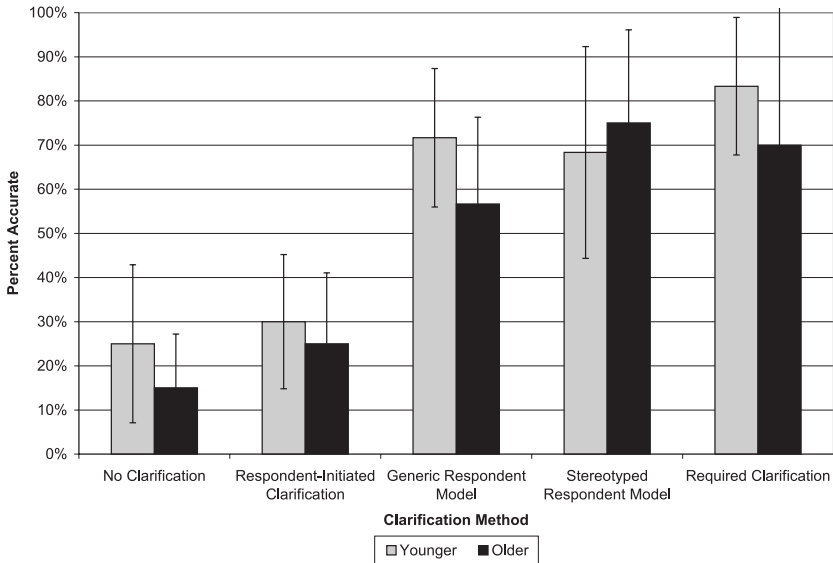


FIGURE 2 Response accuracy by age on complicated mappings.

55) = 3.79, $p = .057$. Younger participants, however, fared about the same with either type of modeling, reaching 68% accuracy with stereotyped modeling compared to 72% with generic modeling: $F_1(1, 45) = 0.17$, ns and $F_2(1, 45) = 0.09$, ns . There was no reliable interaction between age group and condition for all conditions: $F_1(4, 90) = 0.98$, ns and $F_2(4, 110) = 0.73$, ns . Older participants did not fare reliably better from receiving clarification every time (in fact, they appear slightly worse) than they did from stereotyped modeling, with an accuracy of 70%: $F_1(1, 45) = 0.28$, ns and $F_2(1, 55) = 0.28$, ns . Although younger participants fared best in the required clarification condition, with an accuracy of 83%, this increase in accuracy was not reliably better than accuracy in the generic respondent model, $F_1(1, 45) = 2.08$, ns and $F_2(1, 55) = 1.14$, ns ; or the stereotyped respondent model, $F_1(1, 45) = 3.43$, ns and $F_2(1, 55) = 1.89$, ns .

Amount of clarification offered. When it comes to complicated mappings between a concept in a question and the respondent's answer to that question, it is clear that providing more clarification leads to higher accuracy in the answers. However, one may wonder whether the higher accuracy seen in the two modeling conditions is simply due to higher rates of clarification offered overall, rather than actually tapping into conceptual misalignment at the right time.

One approach to investigating that possibility might be made by comparing the response data in which clarification is provided according to our models to response data in which the *same* amount of clarification is provided at random. That approach is problematic, however, because it puts the cart before the horse in using a measure that can only be obtained by the model (the number of clarifications the model determined was necessary) to test the model itself. Regardless, we ran another validation experiment on the model testing data we collected in the required clarification condition, using the same procedure we used in our initial model validation tests, but this time evaluating clarification offered at random. Because the stereotyped model provided the highest amount of clarification, we created a “randomized model” by taking the same number of clarifications participants received in the stereotyped condition and distributing them at random (randomized with an Excel script) among observations in the validation condition.

Taking the same number of clarification cases as the stereotyped model and distributing them at random, the cases where this randomized model predicted a need for clarification led to an improvement in accuracy 29.2% of the time, compared to 28.2% for the stereotyped model. Looking at only the complicated mappings, the randomized model’s predictions led to improvement 49.5% of the time, compared to 53.5% for the stereotyped model. For predicting improvement, the models look comparable so far.

However, when we examine cases where the model predicted people did *not* need help on complicated mappings and yet their accuracy improved after receiving clarification, the randomized model yields a high residual improvement rate of 53.8%, which at $p = .384$ shows no significant difference between improvement predicted by the randomized model and improvement that occurred for other reasons. In contrast, the stereotyped model showed model-predicted improvement of 53.5% versus 33.3% residual improvement, which are significantly different at $p = .039$. These results are summarized in Table 2.

From this validation experiment we can conclude that the stereotyped model succeeds at zeroing in on actual cases of misalignment better than a model that provides the same amount of clarification at random.

TABLE 2
Comparison of Improvement in Stereotyped and Random Models

Variable	All Mappings		Complicated Mappings	
	Stereotyped	Random	Stereotyped	Random
Improvement				
Model-predicted improvement	28.2%	29.2%	53.5%	49.5%
Residual improvement	26.3%	21.1%	33.3%	53.8%
<i>p</i> of predicted versus residual	.404	.134	.039	.384

Time to answer each question. How much overall time will be added to an interview if these various methods of providing clarification are used? The general trend was that the more clarification was given, the more time it took each participant to get through each question. Question–response sequences went fastest when little clarification was given, in the no clarification condition (16.1 sec for the younger group and 19.2 sec for the older group) and the respondent-initiated clarification condition (younger = 15.0 sec, older = 20.2 sec). More time was needed per question in the generic respondent model condition (younger = 39.4 sec, older = 42.5 sec) and slightly more in the stereotyped respondent model condition (younger = 50.0 sec, older = 51.2 sec). As may be expected, the most time was needed when participants heard clarification for every question in the required clarification condition (younger = 60.9 sec, older = 66.1 sec). These differences were reliable by condition, $F(4, 84) = 173.76, p < .001$; and by age group, $F(1, 84) = 6.95, p = .01$; with no significant interaction between the two, $F(4, 84) = 0.31, ns$.

As seen in Figure 3, the younger group is consistently faster. Moreover, respondents took reliably less time to answer questions under the stereotyped respondent model method than they did under the required clarification method: $F(1, 35) = 36.84, p < .001$. This difference is worth noting because the earlier comparison of these two conditions showed no reliable difference in accuracy. Respondents who received stereotyped modeling were just as accurate as respondents who received clarification for every question but spent significantly less time on each question.

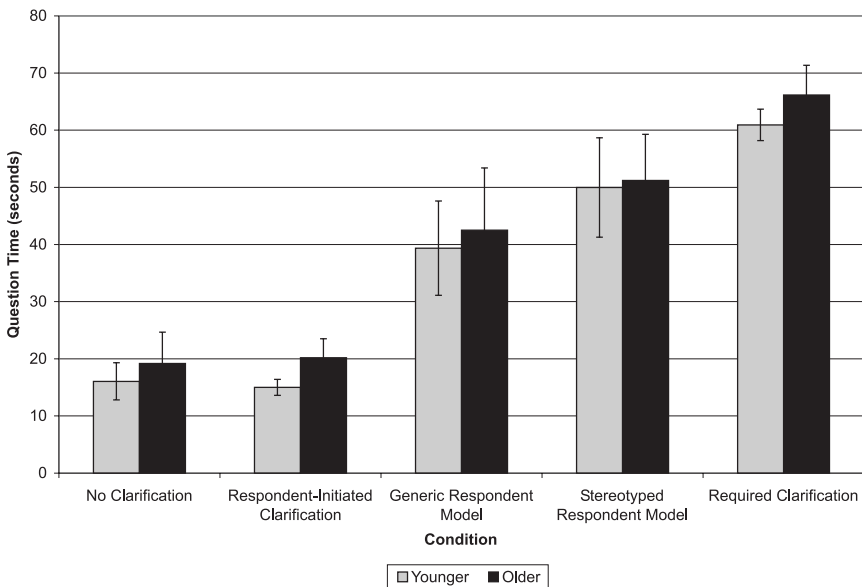


FIGURE 3 Average time to answer each question.

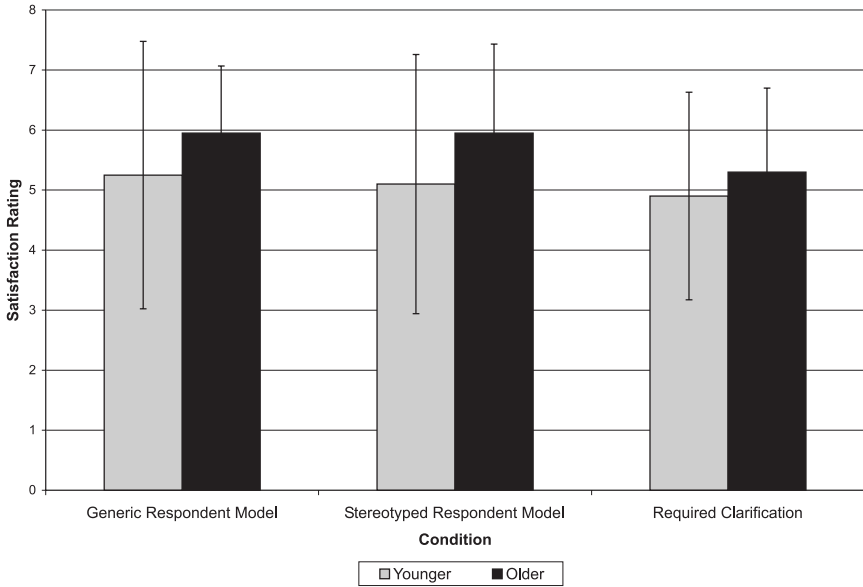


FIGURE 4 Respondent satisfaction with clarification.

Respondent satisfaction ratings. Regardless of whether clarification was offered automatically according to our models or was simply provided along with every question, respondents did not reliably prefer one method over another. Overall mean satisfaction was 5.41, with a marginally higher rating coming from older respondents ($M = 5.73$) than younger ones ($M = 5.08$), showing no reliable effect by age group, $F(1, 54) = 2.11$, *ns*; or condition, $F(2, 54) = 0.49$, *ns*. Satisfaction results are shown in Figure 4.

DISCUSSION

Our purpose here was to look into exploiting that “second plane” of communication in which signals like misalignment cues help to manage the dialogue process and to see if generic or stereotyped models of those signals could help diagnose cases of conceptual misalignment in an automated survey system. After deriving both generic and stereotyped models of various misalignment cues, we found the best cue is conspicuous in its absence: a critical *Goldilocks range* in response latency, outside of which people are more likely to give conceptually misaligned responses.² A generic Goldilocks range of 2 to 7.2 sec is effective at predicting overall levels of conceptual alignment. However, that range is more effective when tailored for older and younger speakers, where older speakers are less likely to pro-

vide a conceptually aligned response if their answers do not fall within a range that is *brief* and slightly *sooner* (2.6–4.35 sec) than the range used by younger speakers (4.6–10.2 sec).

The “*brief* and *sooner*” Goldilocks range for older speakers is a curious observation. Why would older speakers show a shorter and earlier range in which to exhibit signals of conceptual alignment, given their tendency toward more disfluent behaviors?

One possibility is that older speakers have more trouble assuming a frame of reference that is different from their own, so they do not try as much as younger speakers. Another possibility is that they use a different cognitive strategy in the process of alignment by, for instance, adhering to a stronger *presumption of interpretability* (Clark & Schober, 1991) that leads them to jump more quickly to the conclusion that someone else’s frame of reference is already aligned to their own. People aged 65 to 81 at the time of this study could be more willing to believe that words in a survey questionnaire presented to them by someone they view as an authority (whether the government or a psychologist in a laboratory) have been carefully chosen and do not need to be questioned, whereas the younger speakers were more apt to spend time questioning the frames of reference put to them by these so-called authorities. Or, it may be that the subject matter of the questions we asked (jobs and housing) was less relevant to older speakers than to younger ones, causing them either to spend more time analyzing a question or to jump to a prompt, summary conclusion.

Another possible strategy difference is that older speakers simply “know what they know” and have learned over time how to either access that information quickly or just move on—that is, they may have more experience and self-confidence than younger speakers do when faced with challenging problems like the complicated mapping scenarios presented in this study, and are able to assess more quickly than younger speakers whether they can offer an informed response or, if not, just guess quickly and move on.

Whatever the differences in answering strategy between older and younger people, an eager survey practitioner may wonder how well these particular Goldilocks ranges would fare if extended to other dialogue domains and whether they would serve as effective predictors of conceptual alignment for any set of questions. Here we must issue a word of caution and emphasize that the goal of this study is not to

²It should be noted that the latencies seen in these *Goldilocks ranges* may be attributable solely to variations in processing effort on the part of the speaker and are perhaps better described as symptoms—or, at best, *collateral signals* (see Clark & Fox Tree, 2002)—rather than signals with overt communicative purpose. The same goes for many of the paralinguistic events in our second plane of communicative activity. Our point of inquiry here is not whether such events—like a tear from an eye or a frightened yelp—are born from the intention to communicate, but rather whether they are exploitable as such in diagnosing and modeling conceptual misalignment.

offer an absolute predictor for conceptual misalignment that can be immediately applied to other interrogative dialogues; in fact, that very notion would be contrary to our theoretical stance. Rather, we seek to investigate how well this kind of modeling can work in principle, and we have used a set of fairly uniform and innocuous fact-based questions to do so. This uniformity was necessary to prevent differences in the response data from being washed out by confounding variables that would arise from using different types of questions (such as questions about opinions or questions on more sensitive issues), which are likely to yield different results. The task of determining effective Goldilocks ranges that could be applied to various types of dialogue requires a more extensive experiment that would pose different types of questions and solicit answers that would number far higher than the number of observations we collected here.

One general observation we can make, however, is that “help helps—and tailored help helps better.” Whatever the modeling method, both of our models led to more offers of clarification, and more clarification led to greater accuracy. Whether they received this clarification as a result of modeling through the Goldilocks ranges or simply received it for every question, respondents were reliably more accurate when given clarification automatically than when they were left to ask for it on their own or were not given clarification at all. Therefore, help helps, but tailored help helps better because both older and younger age groups were just as accurate under stereotyped modeling as when they received clarification every time, yet took reliably *less time* to answer questions in the stereotyped modeling condition. In addition, they were just as satisfied with the clarification they received under stereotyped modeling as they were with receiving it every time.

Therefore, it seems expectations about the speech of different types of respondents can be successfully modeled to help assess problematic answers to questions or conceptual misalignment in general; and computer interviewing systems could be designed to permit flexible interactions that help respondents provide data that are in line with the intentions of the question authors, in speech systems like this one, as well as in text-based on-line interviews (see Conrad, Schober, & Coiner, 2007).

On a final note, although older speakers showed more disfluent behaviors than younger speakers when verbally answering questions asked by a computer interviewer—using reliably more fillers, restarts, repeats, repairs, reports, repeat requests, and mutters—most of these other disfluent behaviors were not highly significant predictors in regression models that sought to identify cases of conceptual misalignment. Does this mean that these other misalignment cues would not be effective in models that seek to predict conceptual misalignment? On the contrary, these other cues could prove much more effective if modeled on an individualized, case-by-case basis. We have shown how the interpretation of some misalignment cues can fare well using a dynamic, interactive process that relies on the expecta-

tions people bring to a dialogue in the form of stereotyped modeling. However, conversational partners attend to much more than their prior expectations about how each thinks the other will speak; they work from a constant stream of feedback signals that they monitor in each other, and use that feedback to make moment-by-moment adjustments to the models that determine how they interpret each other's behaviors (Clark & Krych, 2004; Pickering & Garrod, 2004; Schober & Brennan, 2003). Therefore, a more ambitious test of the dynamic workings of misalignment cues calls for an approach that creates and adjusts its model of a person's style of speech—that is, works from an *individualized model*—even while that person is still speaking.

ACKNOWLEDGMENTS

This research was generously facilitated by a fellowship from the Charles Cannell Fund in Survey Methodology of the Survey Research Center at the University of Michigan, a Dissertation Fellowship from the New School for Social Research, and National Science Foundation ITR Grant No. IIS-0081550.

We thank John Ehlen and Michelle Levine for their invaluable assistance in this study.

REFERENCES

- Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, *27*, 195–225.
- Blyth, W. G. (1997). Developing a speech recognition application for survey research. In L. E. Lyberg, P. P. Biemer, M. Collins, E. de Leeuw, C. S. Dippo, N. Schwarz et al. (Eds.), *Survey measurement and process quality* (pp. 249–266). New York: Wiley.
- Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*, 123–147.
- Bradburn, N., Sudman, S., & Associates. (1979). *Improving interview method and questionnaire design*. San Francisco: Jossey-Bass.
- Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*, 383–398.
- Clark, H. H. (1994). Managing problems in speaking. *Speech Communication*, *15*, 243–250.
- Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.
- Clark, H. H., & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*, 73–111.
- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81.

- Clark, H. H., & Schober, M. F. (1991). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 15–48). New York: Russell Sage Foundation.
- Cole, R. A., Novick, D. G., Fenty, M., Vermeulen, P. J. E., Sutton, S., Burnett, D., & Schalkwyk, J. (1994). A prototype voice-response questionnaire for the U.S. census. In *Proceedings of the International Conference on Speech and Language Processing (ICSLP 94)* (pp. 683–686). Yokohama, Japan: ICSLP.
- Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, *64*, 1–28.
- Conrad, F. G., & Schober, M. F. (Eds.) (2007). *Envisioning the survey interview of the future*. New York: Wiley.
- Conrad, F. G., Schober, M. F., & Coiner, T. (2004). Bringing features of dialogue to Web surveys. *Applied Cognitive Psychology*, *21*, 165–187.
- Draper, N. R., & Smith, H. (1981). *Applied regression analysis*. New York: Wiley.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 225–242). New York: Seminar Press.
- Johnston, J. (1972). *Econometric methods* (2nd ed.). New York: McGraw Hill.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190.
- Schober, M. F. (1998). Conversational evidence for rethinking meaning. *Social Research*, *65*, 511–534.
- Schober, M. F., & Bloom, J. E. (2004). Discourse cues that respondents have misunderstood survey questions. *Discourse Processes*, *38*, 287–308.
- Schober, M. F., & Brennan, S. E. (2003). Processes of interactive spoken discourse: The role of the partner. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 123–164). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, *61*, 576–602.
- Schober, M. F., Conrad, F. G., & Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology*, *18*, 169–188.
- Schow, R. L., Christensen, J. M., Hutchinson, J. M., & Nerbonne, M. A. (1978). *Communication disorders of the aged*. Baltimore, MD: University Park Press.
- Shrivastav, R., Hollien, H., Brown, W. S., Jr., Rothman, H. B., & Harnsberger, J. D. (2003, November 10–14). *Shifting perceptions of age in voice*. Poster presented at the 146th meeting of the Acoustical Society of America, Austin, TX.
- Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*, 25–38.
- Yairi, E., & Clifton, N. F., Jr. (1972). Disfluent speech behavior of preschool children, high school seniors and geriatric persons. *Journal of Speech and Hearing Research*, *15*, 714–719.