Anchoring and Adjusting in Questionnaire Responses

Anchoring and adjusting in questionnaire responses

Hunter Gehlbach
Harvard Graduate School of Education

Scott Barge
Goshen College

Abstract

When ordering items on attitude/opinion questionnaires, do survey designers bias respondents' answers by the mere act of choosing to organize their survey in a particular way? We hypothesize that, under specific frequently-occurring conditions, respondents employ an anchoring and adjusting strategy in which their response to an initial survey item provides a cognitive anchor from which they (insufficiently) adjust in answering the subsequent item. Three experiments indicate that respondents anchor and insufficiently adjust in certain situations, anchoring and adjusting leads to higher inter-item correlations between adjacent items, and these inflated correlations can (spuriously) increase the reliability estimate of the scale that they comprise and affect the resultant correlations with other measures. These effects are not consistently accounted for by a "superior memory search" explanation. In organizing their surveys, researchers may wish to combat this bias by intermixing items designed for different, but related constructs.

KEYWORDS: Survey design, questionnaire, anchoring and adjusting heuristic, order effects, respondent motivation, satisficing

The prevalence of survey research within psychology is vast (Schwarz, 1999). When surveying respondents, psychologists – particularly in the personality and social subfields – frequently attempt to measure attitudes, opinions, beliefs and other "fuzzy" constructs. To assess these types of constructs which are often unclear in respondents' minds, hard to access in their memories, and/or unstable over time, scholars often try to mitigate error by asking a series of questions which they later aggregate into a summated rating scale (Spector, 1992). Despite the pervasiveness of scales, researchers have little empirical guidance when deciding how to organize these sets of similar questions on a survey. Should they group all the items from the same construct sequentially, or should they intermix items from different constructs? What consequences follow from one approach versus the other?

To address these questions, it is important to understand how people respond to surveys. The prevailing theory is that respondents engage in four processes to develop answers to survey items. Specifically, respondents must comprehend the item that is asked, search their memories to retrieve relevant information, consolidate that information into a judgment, and select an answer to report (Tourangeau, Rips, & Rasinski, 2000). As fatigue or disinterest sets in during the response process, respondents may be tempted to take shortcuts or rely on heuristics to ease the cognitive burden. In other words, respondents might engage in what survey Krosnick (1991) has labeled "satisficing" – failing to put forth optimal effort in completing surveys.

This article investigates whether respondents satisfice (in a previously undocumented way) by relying on *anchoring and adjusting* when adjacent items on a survey are similar. We begin by briefly reviewing literature on ordering surveys so as to situate our investigation. Next, we describe the anchoring and adjusting heuristic and a rival explanation that focuses on memory retrieval. Then we articulate how anchoring and adjusting (as well as the rival memory search explanation) might apply to the survey design context. Finally, we present our research questions and the results of three studies designed to test these questions. To the extent that the phenomenon of anchoring and adjusting generalizes to the domain of survey research, this article contributes to the anchoring and adjusting literature. However, the main contribution of this work (and the focus of the discussion) is to better understand the impact that this heuristic has on data fidelity and provide practical guidelines for survey designers.

Order Effects in Surveys

Many survey design textbooks raise the issue of question-order effects (Dillman, Smyth, & Christian, 2009; Fowler, 2009)[1]. Much of this research has addressed idiosyncratic cases of these effects. For example, Dillman, et al. (2009) describe substantial differences in respondent judgments about whether students should be expelled for plagiarism. It turns out that respondents' answers depend upon whether or not participants are first asked about whether a professor should be fired for the same offense. They also describe a historical example which randomized the order of asking respondents (a) whether communist reporters should be allowed to report on visits to the United States or (b) whether U.S. reporters should be allowed to report on visits to the U.S.S.R. Though the findings are dramatic, they appear to be specific to the content in question rather than broadly generalizable.

Of more relevance to the present study, Dillman et al. (2009) describe two overarching types of order effects. Assimilation effects are those in which answers to items become more similar. As an example, *priming* effects may emerge when early items bring to mind particular beliefs or memories that then become more accessible as respondents answer later items. Contrast effects are those in which responses become more distinct. For example, in *subtraction*, after respondents weigh certain considerations in answering an early item, those considerations are "subtracted" out and not considered in their evaluations of later items.

Though Dillman et al. (2009) indicate that some order effects extend to attitude items, the focus of this prior research has been on isolated items and there is little indication as to how these effects might manifest themselves across a whole scale. However, research by Harrison and McLaughlin (1996) on the physical groupings of items does pertain more directly to the issue of scales and how to organize scale items on a survey. They compared a "uniform grouping condition" in which items assessing a single construct were presented either in the same item-block or randomly interspersed with items from other constructs. They conclude that, "Physically grouping items on a questionnaire slightly enhances internal consistency and discriminant validity, by enhancing the within-set commonalities and between-set distinctions that guide respondents to retrieve relevant caches of information." (p. 329). However, the concept of anchoring and adjusting provides an alternative way to understand their findings.

Anchoring and Adjusting

[1] Question-order effects are distinct from order effects in response options (e.g., primacy/recency effects).

People frequently rely on heuristics to make decisions. One such mental shortcut, anchoring and adjusting (Tversky & Kahneman, 1974), entails anchoring with what is well-known, easily recalled from memory, or salient and then adjusting from that anchor. Use of this heuristic has been documented throughout several decades and across a wide span of cognitive tasks. In one example, Tversky and Kahneman (1974) showed evidence of participants anchoring and adjusting in trying to guess the number of African nations represented in the United Nations after evaluating randomly chosen anchors. Epley and Gilovich (2004) found similar anchoring and adjusting effects from participants' self-generated anchors when asking questions such as what year George Washington was elected president (where 1776 served as the anchor) and what the freezing point of vodka is (where 32°F anchored respondents' guesses). In the very different context of social perception, Ames (2004) found that perceivers who were trying to "read" others tended to anchor with what they (as perceivers) thought they would feel in that situation and adjust from that starting point. Across these and a multitude of other situations in which people engage in this type of anchoring process, a critical aspect of this heuristic is that people's subsequent *adjustments are consistently insufficient* (Epley & Gilovich, 2004, 2006).

For anchoring and adjusting to occur, two facilitating conditions need to be present. First, the anchor needs to be compatible by appearing similar in some way to the guess that is requested or the decision that is to be made (Chapman & Johnson, 2002). In other words, estimating the number of jelly-beans in a jar is unlikely to anchor one's guess as to how many African nations are in the United Nations. Second, the question needs to entail uncertainty (Tversky & Kahneman, 1974). People do not need to rely on this heuristic if they can easily recall the correct factual information.

As a competing theory, Mussweiler and Strack (1999) propose a model in which anchoring effects occur in part because people engage in hypothesis-consistent testing and semantic priming – a substantially different notion than the adjustment process described above. In this conception, people first test the hypothesis that the anchor presented by an experimenter is the right answer to some question. Although individuals reject that hypothesis, they are primed to begin recalling evidence consistent with that anchor. Through this selective memory search, they recall a disproportionate number of data points that are congruent with the anchor (as compared to subjects provided with a different anchor). Thus, when asked to make an absolute

estimate in answer to the question posed, they proffer a judgment that is relatively close to the anchor provided.

Although these competing models aptly explain many studies done in the classic anchoring paradigm, they have not yet been applied to the context of surveys.

Anchoring and Adjusting in Survey Responses

Epley and Gilovich's (2006) investigation of anchoring and adjusting begins to illuminate how this mental shortcut might lead to satisficing on surveys. They posit that in situations where anchoring and adjusting occurs, the adjustments are insufficient because people stop adjusting once they come to a value that falls within a plausible range or seems "close enough." Thus, anchoring and adjustment in surveys may occur as follows: A respondent marks an initial survey item (perhaps by choosing the fourth response option on a five-point scale); if the subsequent item is similar, the respondent might begin evaluating the response options by deciding whether the 4th response on the scale is reasonable; if it does not seem quite right, the respondent might proceed to the 3rd or 5th response option. Once a "good enough" response is reached, the respondent is likely to stop adjusting. If adjustments from these self-generated anchors are insufficient (Epley & Gilovich, 2006), the respondent is disproportionately likely to answer the second survey item at or near that 4th response option on the scale. Thus, over the course of a whole survey, responses to adjacent item-pairs are likely to be more similar (and consequently have smaller distances between them) than responses to the same item-pairs if they are in non-adjacent positions (see Figure 1). In addition, because respondents only need to evaluate response options until a "good enough" response is reached, respondents who are anchoring and adjusting might complete their surveys faster than those who evaluate all response options.

Presumably the aforementioned facilitating conditions of similarity and uncertainty also need to be present. Thus the adjacent item-pairs may need to be compatible by appearing similar in some way. For example, both items in the pair might address the same topic and/or use the same response anchors. In addition, both items should be of the same valence – the common practice of reverse-scoring items would not be conducive to anchoring and adjusting.[2] Second, the adjacent item-pairs need to entail uncertainty. In other words, survey items asking

[2] Although a common practice, using reverse scored items is not a recommended one (Benson and Hocevar, 1985; Swain, Weathers, and Niedrich, 2008).

respondents about their attitudes, opinions, or beliefs (rather than easily recalled factual information) would be most likely to produce anchoring and adjusting.

Anchoring and adjusting signals sub-optimal mental effort and lower fidelity in responses that researchers should strive to minimize. However, the observable consequences of anchoring and adjusting – similar answers to adjacent items assessing the same construct – could arise for another reason. We refer to this competing possibility as the "superior memory search" explanation. Mussweiler and Strack's (1999) conception of anchoring lays the foundation for this model. Their research indicates that once respondents are primed to think about a certain topic – perhaps through the presentation of one or two initial survey items on a particular construct – they may recall more and more information about that construct through a selective memory search. Of particular relevance to the present context is the idea that they recall evidence consistent with the initial anchor and ultimately offer judgments that are relatively close to that anchor.

Although their research tends to focus on a paradigm in which the experimenter provides the anchor (unlike the survey context), Knowles (1988) provides an empirical illustration using surveys that lends support to this superior memory search possibility. He suggests that as respondents answer more items about a given topic, they begin to recall more information that informs their reporting on that topic. As a result, their responses become more consistent as they continue reporting on the same topic. Taken together, the theory from Mussweiler and Strack (1999) and Knowles' (1988) findings suggest that grouping related items together in the same item-block might facilitate a superior memory search in which respondents stay focused on a particular topic, recall more, and as a result, produce more accurate, reliable responses which happen to be similar to one another.

To summarize the main distinctions between these alternative possibilities: if respondents anchor and adjust when pairs of adjacent items are related, the distances between the two items will be smaller than if the respondents had been required to answer intervening items. If respondents engage in this mental shortcut, their answers will be biased away from their actual beliefs towards their response to the previous items. As a primary consequence, individuals' scores would consist of more error and less true score (DeVellis, 2003). Alternatively, when similar items are grouped together in an uninterrupted block, perhaps the superior memory search effect facilitates respondents' memory search and makes responses for *all the items in the*

scale manifest smaller between-item distances than if they had been intermixed with other items. As Knowles (1988) and Harrison and McLaughlin (1996) suggest, in this case, respondents' answers may be more accurate and the overall scale would have a higher true score. Thus, for scales that are presented in a uniform block – assuming the ordering of the items was random – the anchoring and adjusting prediction would be to find smaller distances and higher correlations between *adjacent* item-pairs within the block. However, the superior memory search explanation would not predict any differences between adjacent and non-adjacent items because all items are part of the same scale and each item facilitates the priming one's memory for all the ensuing items.

Research Questions

The present research uses a split-ballot design in which participants are randomly assigned to take one of two forms of a survey. Form 1 places item-pairs of interest adjacent to one another, thus potentially facilitating anchoring and adjusting. Form 2 (and Form 3 in the case of Experiment 3) separated the item-pairs of interest with intervening items that were dissimilar in some way on the assumption that this would mitigate anchoring and adjusting. None of the scales we examined included reverse-scored items. We use this research design to investigate the extent to which the anchoring and adjusting heuristic occurs in survey responses, what the consequences are for the data, and whether these consequences might reasonably be accounted for by the superior memory search explanation. Within each experiment, we test four main research questions:

1) To what extent does anchoring and (insufficient) adjusting occur between adjacent item-pairs that focus on similar topics and use similar response scales?
2) Will anchoring and adjusting lead to higher correlations between items within the scale (and therefore ostensibly higher reliabilities as assessed by coefficient alpha)?
3) Can the superior memory search explanation adequately account for any differences in response patterns that we find between the two survey forms?
4) Will the strength of the associations between the focal scales and other related scales differ between Form 1 and Form 2?

The fourth research question requires additional explanation. All other things being equal, a more reliable measure will correlate with another measure more strongly than a less reliable measure of the same construct, provided that the actual relationship is not $r = 0$ (Glass & Hopkins, 1996). However, we anticipated that the reliabilities on Form 1 would be *artificially* higher. In other words, although the estimates of coefficient alpha might appear greater on the Form 1 scales in a mathematical sense, the actual proportion of true score within those scales would not be higher (see DeVellis, 2003 for an explanation of reliability in terms of proportions of signal and noise). Consequently, we anticipated that this difference in correlations between the focal scales and other measures would not emerge.

Our analytic approach was similar across all three experiments. We describe differences between respondents and differences between the relevant item-pairs across each survey form throughout the results. Because the distances between items across the two forms of the survey were approximately normally distributed, we report parametric tests of mean differences (*t*-tests and an ANOVA) for these analyses. For research questions 2, 3, and 4, we compare sets of correlation coefficients against one another. In these cases, we are reluctant to assume that these distributions are normal and thus, conducted our statistical testing using non-parametric tests as a more conservative approach. Specifically, we use Wilcoxon's signed rank test for research questions 2 and 4 where the correlations are paired and Wilcoxon's rank-sum test (also known as the Mann-Whitney-Wilcoxon or Mann-Whitney *U*) for research question 3 where they are unpaired.

Experiment 1

We initially investigated anchoring and adjusting within a larger study of teachers' multi-cultural competencies in the classroom (Irizarry & Gehlbach, 2007). Using a between-subjects design, we randomly assigned teachers to two different forms of the survey. We presented the item pairs of interest to respondents as adjacent to one another in Form 1 and non-adjacent in Form 2.

Method

Participants. The participants ($N = 172$) included 103 pre-service teachers and 69 of their cooperating teachers in the northeastern United States. Participants were predominantly female (78%) and White (92%). All spoke fluent English.

Measures. The first author and a colleague developed 50 items to create a new measure of teachers' multicultural competence through five different scales of 10 items each. Building off of a predominantly dispositional theory of multicultural competence (Washington & Evans, 1991) we assessed how *aware* (overall α = .87), *knowledgeable* (overall α = .88), *motivated* (overall α = .89), and *skillful* (overall α = .86)[3], teachers perceived themselves to be. Response anchors for these four constructs were formed by adding each construct label to the following 5-point response anchors: not at all, slightly, moderately, quite, and extremely (e.g., "not at all aware," "slightly aware," etc.). The fifth scale, assessing how frequently teachers took action (overall α = .88), used response anchors of "almost never," "once in a while," "sometimes," "often," and "almost all the time." To also incorporate the ideas from a competing, knowledge-based theory of multicultural competence (Banks & Banks, 2001), 2 of the 10 items within each scale addressed each of the following: epistemology, content, equity, prejudice, and cultural change.

Participants were randomly assigned to complete Form 1 (n = 94) or Form 2 (n = 78). These forms varied only in the order in which items were presented. Form 1 grouped items that were intended to address the same construct together in a cohesive item block, while Form 2 intermixed these items so that items of the same construct were not adjacent. More specifically, Form 1 presented all the *awareness* items first, in the following order: 2 epistemology items, 2 content items, 2 equity items, 2 prejudice items, and 2 cultural change items. Each subsequent construct proceeded in the same fashion. As shown in Appendix A, this organization resulted in Form 1 having 20 item-pairs that were similar in the specific content of the question and the wording of the response anchors. Thus, we examined each of these 20 item-pairs that were adjacent on Form 1 and non-adjacent on Form 2.

To explore our fourth research question of how our focal scales were associated with other measures, we included two additional scales that were presented identically to all participants at the beginning of the survey. First, we assessed participants' teaching efficacy by adapting one of the scales from the Patterns of Adaptive Learning survey (Midgley et al., 2000). This 8-item scale (α = .84) assessed how confident teachers were in their teaching ability through items such as, "How confident are you that you can teach even the most challenging

[3] In each study, the "overall" alphas refer to reliability estimates computed on the entire sample. These estimates are distinct from the computations of alpha that we compute on each Form of the survey as a part of research question 2 (which are presented in each Results section).

students?" Second, we assessed participants' propensity to take the perspective of others. This 7-item social perspective taking scale (α = .88) was based on Davis' (1983) scale and included items such as "How often do you try to figure out how the people around you view different situations?"

Procedure. The pre-service teachers completed the survey during class and then gave a copy of the survey to their cooperating teachers. Cooperating teachers completed the survey on their own time and returned it in a pre-paid envelope. All surveys were paper and pencil.

Results

Research Question 1. To investigate the extent to which anchoring and insufficient adjusting occurred, we compared the difference in the distances between focal item-pairs across Forms 1 (where the items were adjacent) and 2 (where the items were non-adjacent). First, we computed the absolute value of the difference between item-pairs of interest for each respondent. In other words, if a respondent marked the 4th response option for the initial item and the 3rd response option for the subsequent item in the pair, the absolute value of the difference would be 1 (as shown on the left hand side of Figure 1). Next, scores for each item-pair of interest were computed and aggregated so that each participant received an overall "anchoring and adjusting" score representing the mean absolute difference between the item-pairs of interest for that person. These mean scores were then compared between respondents who completed Form 1 versus Form 2. If respondents anchored and then adjusted insufficiently on Form 1 as expected, then their mean anchoring and adjusting scores would be smaller than for respondents of Form 2. Following this procedure, we found solid evidence of anchoring and adjusting on Form 1. The overall mean of participants' anchoring and adjusting scores was .63 (*sd* = .25) for Form 1 as compared to .75 (*sd* = .31) for Form 2, ($t_{(170)}$ = 2.88, p = .004, Cohen's d = .44). Dis-aggregating these results to specific item-pairs of interest, we found that Form 1 respondents had smaller between-item differences for 18 of the 20 relevant item pairs[4].

[4] Astute readers will notice that the design of the survey also allows us to test instances in which we might expect anchoring and adjusting to occur on Form 2 relative to Form 1. Specifically, there are 20 instances in Form 2 where adjacent items refer to similar content (according to the Banks and Banks, 1991 topics) which are non-adjacent on Form 1. For example, we could compare the second ("How would you rate your awareness of the ways that knowledge is constructed within your discipline(s)?") and third ("How knowledgeable are you of methods to help students understand multiple sides of debates in your discipline(s)?") "Epistemology" items across forms. This approach provides a test for whether anchoring and adjusting occurs in the case of similar content but different response anchors. For this analysis the overall mean of participants' anchoring and adjusting scores was .86 (*sd* = .37) for Form 1 as compared to .76 (*sd* = .36) for Form 2, ($t_{(170)}$ = 1.73, p = .09, Cohen's d = .27). In dis-

Research Question 2. We next investigated whether these insufficient adjustments led to stronger between-item correlations on Form 1 relative to Form 2 and whether these correlations, in turn, affected the internal consistency of the scales. As expected, the correlations between the adjacent item-pairs on Form 1 were higher than when those items were non-adjacent on Form 2 in 16 out of 20 instances. A Wilcoxon signed rank test confirmed that the median correlation for all the relevant item-pairs on Form 1 (*mdn* = .50) was significantly greater than for the corresponding item-pairs on Form 2 (*mdn* = .39; $z = 2.69$, $p = .007$; $r_{effect} = .85$).

Using Feldt's (1969) test, we compared the reliabilities for the five scales: *awareness, knowledge, motivation, skill,* and *frequency of action*, across the two forms. The Form 1, reliabilities were significantly higher than Form 2 for the *awareness* scale ($\alpha = .89$; versus $\alpha = .84$; $W = 1.51$, $p = .03$) and marginally higher for the *knowledge* scale ($\alpha = .89$; versus $\alpha = .85$; $W = 1.38$, $p = .07$). For the *motivation*, *skill*, and *frequency of action* scales the reliabilities on Form 1 were higher than on Form 2 but not significantly so.

Research Question 3. The third research question predicted that these differences in item-pair distances, item-pair correlations, and scale reliabilities could not be attributed to respondents engaging in more thorough memory searches (though they could be explained by anchoring and adjusting). To test this possibility we looked only at the respondents who completed Form 1. If these individuals engaged in a superior memory search, then *all* the inter-item correlations for the items within that scale should be similar. In other words, because all items pertain to the same topic on Form 1 if a superior retrieval process is to account for these results, the average distance between items within a block should be roughly the same for adjacent and non-adjacent item-pairs provided that there are no intervening items on a different topic to interrupt the retrieval process.

In testing the plausibility of this hypothesis, we examined whether the inter-item distances and correlations were higher for adjacent versus non-adjacent item-pairs within the Form 1 respondents only. Table 1 shows the descriptive statistics for each scale. To test whether these distances and correlations differed overall, we aggregated the adjacent and non-adjacent distances and correlations across all scales and then compared them. Results of the *t*-

---

aggregating these results, we found that Form 2 respondents had smaller between-item differences for 16 of the 20 relevant item pairs than Form 1 respondents. Because these results were not significant at the conventional level despite the similar item content, we do not test for anchoring and adjusting on Form 2 relative to Form 1 in the remaining experiments, although we do revisit this issue in our discussion of future directions.

test for all relevant item-pairs indicated smaller between-item distances for adjacent .61 (*sd* = .23) as compared to non-adjacent .72 (*sd* = .25) items ($t_{(93)} = 10.20$, $p < .000$, Cohen's $d = .45$). Using the Wilcoxon rank-sum test, we also found larger between-item correlations for the adjacent (*mdn* = .54) versus non-adjacent items (*mdn* = .42; $z = 4.78$, $p < .001$; $r_{effect} = .32$).

Research Question 4. To investigate whether the anchoring and adjusting on Form 1 affected the relationships with other variables, we examined the five scales' correlations with two other scales used in the questionnaire – teaching efficacy and social perspective taking. Usually scales with higher reliabilities correlate more strongly with measures of other constructs that are theoretically related (Glass & Hopkins, 1996). However, because we assumed that the reliabilities from the Form 1 scales were artificially inflated, we expected this not to be the case.

To examine these between-scale associations, we correlated individual's scores on each of the five focal scales with a measure of teacher efficacy and with a measure of social perspective taking – measures we anticipated would correlate positively with the five scales. As shown in Table 2, six of these correlations were higher on Form 1 and four were higher on Form 2. Congruent with our expectations, when we tested these differences using the Wilcoxon signed-rank test, we found no differences between these correlations ($z = .05$, *ns*).

Discussion

Overall, this experiment shows that when respondents were presented with a series of sequential items assessing the same construct and using the same response options, they anchored and adjusted. Specifically, respondents used their responses on initial items as anchors and then adjusted insufficiently from that anchor in responding to subsequent items. As a result, Form 1 respondents produced smaller distances between adjacent item-pairs than Form 2 respondents produced on those same items when they were not adjacent. The correlations between these same item pairs of interest were higher for Form 1 as compared to Form 2 respondents. Thus, we found support for research question 1 and part of research question 2; some evidence for differential reliabilities across forms was present but, as discussed below, more muted.

We demonstrated that these findings did not result from respondents engaging in a more effective memory search. Within the scales on Form 1, we found that responses to adjacent items were more similar and more highly correlated than the non-adjacent items. This

discrepancy cannot be accounted for by the memory search explanation because all items are part of the same construct. Thus, although we cannot rule out the possibility that grouping similar items together facilitates the memory search process, we can say that the memory search explanation does not account for the item-pair differences in this experiment. The anchoring and adjusting explanation appears more plausible. The fourth research question was also supported – there were no differences by form with respect to the correlations between scales.

One puzzle from these findings concerns the minimal differences in reliability across forms. These minimal differences may stem in part from a particular characteristic of coefficient alpha. The formula for alpha is a function of all the inter-item correlations within a scale (DeVellis, 2003). Thus, for a 3-item scale, there are two opportunities for respondents to anchor and adjust and three inter-item correlations that help determine alpha. On the other hand, for a 10-item scale the ratio is much different. There are 9 opportunities for anchoring and adjusting but 45 inter-item correlations. In other words, as the number of items on a scale increases, the proportion of the inter-item correlations that might be affected by anchoring and adjusting decreases rapidly. In sum, we suspect that the minimal impact on reliabilities might be due to our choice to investigate relatively long scales (i.e., 10 item scales). In Experiment 2, we further explore this issue and investigate the extent to which anchoring and adjusting generalizes to a different respondent population and different survey characteristics.

Experiment 2

To help assess the generalizability of our initial results, we investigated anchoring and adjusting on a survey of university alumni by looking at scales of different lengths, assessing different constructs, using different response scales, examining a web-survey, expanding the testing of the correlations between the focal scales and other constructs, and selecting participants from a very different cultural and linguistic context. We tested the same four main research questions.

Method

Participants. Students (N = 506) who had attended an Eastern European university completed an alumni satisfaction survey as part of the university's institutional research. Alumni were randomly assigned to Form 1 (n = 254) or Form 2 (n = 252) of the survey. Congruent with the gender balance at the university, more females (n=371) than males (n=135) completed the survey. The participants were not native English speakers, although all academic work at the

university was in English. To gain admission, students met a minimum standard on the Test of English as a Foreign Language. Alumni completing the survey represent a variety of current language contexts. Some currently live and/or work in English-speaking environments, while others have returned to contexts in which their native language is predominant. We suspected that the alumni who were no longer speaking English regularly (particularly older graduates who have been away from the university for longer) may have lost some of their language skills, thus introducing an additional level of uncertainty in their survey responses.

Measures. We examined scales that asked the alumni, "Please rate how important each of the following skills and competencies has been in your life since college. Please choose the appropriate response for each item:" All items used the same set of six partially-labeled response anchors ranging from "not important" to "very important," with the intervening points numbered 1, 2, 3, and 4. The *academic* scale (overall $\alpha = .72$) focused on the alumni's valuing of different academic abilities since graduating from college. This measure consisted of seven items such as "Write effectively in English." The *social* scale (overall $\alpha = .81$) contained three items inquiring about students' understandings of societal issues. "Understand current social problems" was a representative item. The seven-item *interpersonal* scale (overall $\alpha = .83$) assessed students' skills in working with and relating to others. It used items such as, "Resolve conflicts between people positively."

As before, Form 1 grouped certain items that were intended to address the same construct together in a cohesive item block, while Form 2 intermixed these items with items from other scales. On both forms, the 17 items were presented together on a single screen – only the ordering of the items differed across forms. On Form 1, we expected anchoring and adjusting to occur between "Write effectively in English" and "Communicate well orally in English" because these items were the first and second in that section of the survey. On Form 2 these same items were in the first and 5th positions, and thus we expected no anchoring and adjusting. Thus, across the three scales there were 14 item pairs where we anticipated that anchoring and adjusting would occur. See Appendix B for all items and the respective ordering of items on Form 1 and Form 2.

To explore the fourth research question we included nine additional scales that were presented identically to all participants. The first three scales paralleled the three focal scales (αs = .76, .82, and .89, respectively), but asked about the *institution's contribution* to alumni

development of the skill/competency (rather than asking about *how important* each skill/competency was to the alumni at present). Thus, for these three scales, the items were identical except for the response anchors. Two additional scales were similar in that they asked alumni to rate how important specific business core skills (7 items; $\alpha = .75$) and content areas (10 items; $\alpha = .79$) were in their lives at present. Another two scales asked alumni to rate how involved they were in various extra-curricular activities, and then to evaluate how much their involvement in each activity contributed to their personal development (10 items each; $\alpha = .79$ and $\alpha = .80$, respectively) The final two scales captured altruistic behaviors (11 items; $\alpha = .76$) and the extent to which they re-evaluated their values and beliefs while students (7 items; $\alpha = .90$).

Procedure. University staff administered the survey online by emailing: an introductory note about the survey from the university's president, an invitation to take the survey itself, and three follow-up reminders (sent to non-responders). Respondents completed the approximately 20 minute survey on the university's website.

Results

Research Question 1. As before, we first assessed the extent to which anchoring and insufficient adjustment occurred by computing each individual's overall anchoring and adjusting score and comparing across forms. Congruent with expectations, significantly more anchoring and adjusting occurred where similar items were adjacent to one another on Form 1 ($M = .77$, $sd = .46$) than on Form 2, where items were intermixed ($M = .92$, $sd = .48$; $t_{(483)} = 3.58$, $p < .001$, Cohen's $d = .22$). The average absolute value of distances between item pairs of interest was smaller on Form 1 (when the items were adjacent) than on Form 2 (when the items were non-adjacent) 13 out of 14 times. Thus, we also found additional evidence of anchoring and adjusting in this new survey context.

Research Question 2. These insufficient adjustments generally led to stronger between-item correlations in Form 1 as compared to Form 2. Results from the Wilcoxon signed rank test showed that correlations between the adjacent item-pairs on Form 1 ($mdn = .51$) were higher than when those items were non-adjacent on Form 2 ($mdn = .40$; $z = 2.79$, $p = .005$; $r_{effect} = .75$).

These correlations, in turn, affected the internal consistency of the scales. Using Feldt's (1969) test, we found that the *academic* scale was significantly more reliable on Form 1 ($\alpha = .76$) than on Form 2 ($\alpha = .68$) $W = 1.35$, $p = .01$. Similarly, for the *social* scale we found that Form 1

(α = .85) produced a more reliable scale than Form 2 (α = .77) W = 1.53, p < .001. The reliabilities were no different for the interpersonal scale (α = .83 in both cases).

Research Question 3. In examining whether better memory searching by Form 1 respondents could account for these results, we again looked only at the Form 1 respondents and compared their adjacent and non-adjacent item-pairs. Results of the *t*-test for all three scales indicated smaller between-item distances for adjacent .76 (*sd* = .46) as compared to non-adjacent .87 (*sd* = .50) items ($t_{(242)}$ = 6.24, p < .000, Cohen's d = .22). In testing the differences between correlations using the Wilcoxon rank-sum test, we found that the correlations were significantly higher for the adjacent items (*mdn* = .61) than the non-adjacent items (*mdn* = .50; z = 2.29, p = .02; r_{effect} = .34). The descriptive statistics of each scale are presented in Table 3.

Research Question 4. To investigate whether the anchoring and adjusting found on Form 1 affected the associations with other variables, we examined the three scales' correlations with other scales used in the questionnaire. Specifically, we looked at their convergent validity with nine related scales and their discriminant validity with two single-item indicators. For each of the nine other scales, we expected a positive correlation with our three scales of interest; for the single item indicators, we expected no relationship. We actually found the correlations with other scales to be higher for Form 2. Of these 27 correlations where positive associations were expected, the correlations with Form 2 were greater than the correlations with Form 1 in 20 instances. The Wilcoxon signed-rank test confirms that, taken together, correlations are significantly higher on Form 2 (*mdn* = .50) than on Form 1 (*mdn* = .35; z = -3.26, p = .001; r_{effect} = .63). As expected, the correlations between the three scales of interest and the single-item indicators were close to 0 (see Table 4).

Discussion

This experiment replicated the anchoring and adjusting phenomenon documented in Experiment 1 and provided additional support for each research question. The results indicate that anchoring and adjusting may generalize to different respondent populations and survey contexts – e.g., non-native English speakers and web-based surveys.

In assessing the fourth research question, we expected to find no difference in the correlations between the Form 1 and Form 2 scales in their relationships with other measures. This null finding would indicate that the reliabilities were indeed artificially inflated. We

actually found that the scales from Form 2 correlated more highly with related measures – an issue that warrants further exploration.

Experiment 3

In Experiment 3 we strove to accomplish three main goals. First, we wished to see whether our findings generalized to a younger population of students. Because satisficing is (negatively) associated with respondents' cognitive sophistication (Krosnick, 1991), we anticipated that we might find differences between respondents of different grade levels. Second, we wished to test one additional hypothesis. If anchoring and adjusting is a form of conserving effort on surveys, then logic dictates that the more that respondents anchor and adjust, the faster they should work through different sections of the survey. Third, we explored a potential solution to respondents' use of this heuristic, i.e., whether anchoring and adjusting might be mitigated by presenting survey items one at a time (via a web survey).

Method

Participants. Data for this experiment came from a broader study of teacher-student relationships. High school students (N = 214; 43% male) attending a parochial school in the northeastern United States participated. Students were randomly assigned to one of three survey forms. Form 1 respondents (n = 77) saw items presented in blocks that corresponded to the constructs being measured. Form 2 (n = 78) respondents received items presented in the same order as Form 1 respondents, but the items were presented one at a time. Thus, because respondents were directed to a new screen for each item, their anchor visually disappeared before they read the next item. In Form 3 (n = 59), items from the different constructs were mixed together so that items assessing the same construct were never adjacent. Students spoke a mix of English (52%), Spanish (32%), and other languages (including multiple languages) as their primary home language. Participants included 9th (30%), 10th (28%), 11th (26%), and 12th (17%) grade students.

Measures. We investigated anchoring and adjusting on four different scales. The *anxiety* (overall α = .67), *enjoyment* (overall α = .92), and *boredom* (overall α = .85) scales were adapted from Pekrun, Goetz, Titz, and Perry (2002). Each scale contained five items designed to assess the extent to which students felt each of those three emotions during a specific academic class (e.g., "How tense do you feel during this class?" as a representative item for *anxiety*). The *similarity* scale (overall α = .80) was developed for the research on teacher-student relationships.

"Overall, how similar do you think you and <teacher's name> are?" constitutes a representative item. Each scale employed fully-labeled, 5-point response anchors.

Form 1 and 2 presented items in the same order. For example, "How excited are you about going to this class?" was followed immediately by "How enjoyable is being in this class?" as the 6th and 7th items presented in this section of Forms 1 and 2. For Form 3, these items were in the 3rd and 7th positions, respectively. Thus, no anchoring and adjusting was expected to occur on Form 3. In total, Form 1 contained 16 items pairs where we thought anchoring and adjusting might occur. See Appendix C for the exact items and their respective order on each form of the survey.

To explore the fourth research question, we compared the four focal scales to scales that assessed students' perceptions of their relationship with their teacher, their propensity to take their teacher's perspective, how much effort they put into class, their sense of self-efficacy in class, and their sense of belonging at their school. The teacher-student relationship scale consisted of positively (9-items; $\alpha = .89$) and negatively (5-items; $\alpha = .71$) valenced sub-scales and asked items such as "How friendly is <teacher's name> towards you?" and "How often do you ignore something <teacher's name> says?", respectively (Gehlbach, Brinkworth, & Harris, in press). The social perspective taking scale (7-items; $\alpha = .85$) represented a slight adaptation from the scale used in Experiment 1 – specifically, we asked students to focus on their propensity to take the perspective of their teachers (as opposed to people in general). The 5-item effort ($\alpha = .70$) and self-efficacy (.80) scales were also used in previous studies (Gehlbach, et al., in press). They consisted of items such as "How much effort do you put into your homework for this class?" and "How confident are you that you can learn all the material presented in this class?" Finally, the 4-item sense of belonging ($\alpha = .76$) measure (Roeser, Midgley, & Urdan, 1996) presented statements such as, "I feel like I matter in this school." Respondents had to assess how true each statement was for them.

Procedure. The research team administered this online survey in the computer lab of the school as classes of students came down during their English class to participate. Most students took approximately 30 minutes to complete the survey.

Results

Research Question 1. We again found evidence of anchoring and adjusting. In assessing overall anchoring and adjusting across all 16 item-pairs of interest, Bonferroni post-

hoc tests from an analysis of variance revealed that respondents to Form 1 (M = .76, sd = .26) differed significantly from respondents to Form 3 (M = .96, sd = .35) but not Form 2 (M = .81, sd = .26); $F_{(2, 209)} = 8.22$, $p < .001$, $\eta_p^2 = .05$. Because Form 2 did not differ significantly from Form 1, we did not find support for the possibility that presenting items one at a time mitigates anchoring and adjusting. Thus, Form 2 is not discussed further in the results. Of the 16 item-pairs of interest, distances were shorter on Form 1 than on Form 3 in 12 of 16 instances.

Research Question 2. These insufficient adjustments led to stronger between-item correlations in Form 1 as compared to Form 3. The Wilcoxon signed rank test showed that correlations between the adjacent item-pairs on Form 1 (mdn = .61) were higher than on Form 3 (mdn = .50), when those items were non-adjacent (z = 3.07, p = .002; r_{effect} = .77).

These correlations, in turn, impacted the internal consistency of the scales. Using Feldt's (1969) test, we found that the *anxiety* scale was significantly more reliable on Form 1 (α = .75) than on Form 3 (α = .48; W = 2.08, p = .002). Form 1 of the *similarity* scale (α = .84) was more reliable than Form 3 (α = .71; W = 1.78, p = .01). The reliabilities for the enjoyment and boredom scales, though both slightly higher on Form 1, were not significantly different.

Research Question 3. We used the same procedures as the previous experiments to examine the memory search explanation. Comparisons of between-item distances for all the adjacent items (m = .76, sd = .26) versus non-adjacent (m = .83, sd = .29) items in the four scales, revealed significant differences between the two ($t_{(76)} = 2.91$, p = .005, Cohen's d = .25). This time the Wilcoxon rank-sum test showed no differences between the correlations for the adjacent versus non-adjacent items (z = .58, p = .56). See Table 5 for the mean distances and mean correlations.

Research Question 4. To investigate whether the anchoring and adjusting on Form 1 affected the relationships with other variables, we examined the 4 scales' correlations with other scales used in the questionnaire. Specifically, we looked at their relationships with a series of scales where we expected to see both positive and negative correlations. We found no evidence that, despite the generally higher reliabilities of the scales in Form 1, these scales produced stronger relationships with other measures. On the contrary, a Wilcoxon signed-rank test revealed that the correlations between scales were stronger (in 20 of 24 cases) for respondents of Form 3 (mdn = -.45) as compared to Form 1 (mdn = -.31; z = 2.99, p = .003; r_{effect} = .61). See Table 6 for the correlations between scales.

Additional analysis. Because this web-survey was conducted using a web-survey application which allows for the collection of certain types of meta-data – we were able to test an additional hypothesis. According to Epley and Gilovich's (2006) conception of anchoring and adjusting, the respondents for Form 1 should complete the focal scales only by evaluating response options until they reach a plausible response. Thus, they should complete that section of the survey faster than respondents of Form 3 (who would need to evaluate response options until they reach the most accurate response). Congruent with that expectation, Form 1 respondents (M = 130 seconds, sd = 56) completed this section of the survey more quickly than Form 3 respondents (M = 152 seconds, sd = 58; $t_{(132)} = 2.21$, $p = .03$, $d = .39$)

Discussion

This experiment replicated the anchoring and adjusting phenomenon in yet another population using another group of survey measures. Support was again found for the notion that respondents who are presented with survey items in a block of similar questions with similar response anchors will answer these questions more similarly than if the items were distributed throughout the survey with other items interspersed. The ramifications of anchoring and adjusting were similar to the first two studies – higher between-item correlations and artificially inflated reliability estimates (as computed by coefficient alpha). For this experiment, the results were slightly mixed in our investigation of the memory search explanation. However, given the context of the previous results, it seems unlikely that the memory search explanation can fully explain these findings. Particularly in light of this experiment's additional finding that respondents of Form 1 completed the items more quickly than respondents to Form 3, it is hard to imagine that these respondents are searching their memories more exhaustively (although there is some chance that Form 1 respondents are searching their memories more efficiently because the content of the adjacent items is similar). Finally, despite the higher reliabilities of the Form 1 scales, these scales actually correlated more weakly with other measures – thus replicating the finding from the second experiment[5].

General discussion and implications for survey researchers

[5] As a final set of exploratory analyses we sought to understand who anchors and adjusts more. We found no differences in mother's educational level (Study 1); English fluency, grade point average, or age (Study 2); grade level, primary language spoken at home, or parents' educational level (Study 3); or gender (all studies).

Taken as a whole, these results illustrate that anchoring and adjusting occurs on attitude/opinion questionnaires between adjacent items that use the same set of response anchors and contain related content. Specifically, when survey respondents face items which are grouped according to the constructs they are intended to measure they invoke a heuristic in which they use their response to an initial item as an anchor. In responding to the subsequent item that is presented, they (insufficiently) adjust from that anchor. These findings differ from "straight-line responding" in which respondents mark the same answer throughout a section or whole survey – we screened out all such respondents before beginning our analyses.

The concern for survey researchers is that, when anchoring and adjusting does occur, data may be compromised, particularly for shorter scales (e.g., 3-7 items). Specifically, because respondents give artificially similar responses for adjacent items, they introduce error into their responses. This error leads to spuriously high correlations between items within the scale and can artificially inflate estimates of the scale's internal consistency. In other words, researchers may be tricked into thinking their scales are significantly more reliable than they actually are.

For each experiment, we also examined an alternative explanation that presenting conceptually similar items adjacent to one another might facilitate respondents' cognitive search and retrieval capabilities (Harrison & McLaughlin, 1996; Tourangeau, et al., 2000). Potentially, this superior memory search process, not anchoring and adjusting, could explain the similarity of respondents' answers. Although it remains plausible that this approach to organizing surveys does facilitate respondents' memory searches, it does not account for the effects that we found. In each experiment, when we compared the adjacent and non-adjacent items within the focal scales for Form 1 respondents, we found that the overall adjacent between-item distances were smaller than their non-adjacent counterparts. As described above, these adjacent/non-adjacent differences should not result from differences in how respondents' search their memories – every question within the item-block pertained to the same construct – though these differences are expected as a consequence of anchoring and adjusting.

In relating the focal scales of interest to other measures we found no evidence that the (ostensibly) more reliable scales from Form 1 respondents produced stronger between-scale correlations. On the contrary, we found substantial evidence of the exact opposite – in Experiments 2 and 3, correlations between the focal scales and other measures were stronger on the forms where anchoring and adjusting was mitigated. We posit that this potentially

counterintuitive finding is due to the fact that the reliabilities were artificially inflated by anchoring and adjusting and that the true reliabilities of these measures are lower than (or towards the lower bound of) the coefficient alpha estimate. The data that Harrison and McLaughlin (1996) gathered are also consistent with this conclusion. That the ordering of survey items can have such a substantial impact on the results of correlational findings demonstrates how important it is for researchers to attend to item order in designing their surveys. For example, the finding that students' propensity to take the perspective of others is unrelated to their enjoyment of a class (r = .14; as was found on Form 1) is a very different finding than a significant association of r = .50 (as was found on Form 3).

In our final experiment, we also gained insight into the behaviors associated with anchoring and adjusting during a survey administration. Specifically, because anchoring and adjusting is a mental shortcut, we reasoned that those who employed the anchoring and adjusting heuristic would complete the relevant survey sections more rapidly than those who were not employing the heuristic. Experiment 3 supported the notion that the anchoring and adjusting heuristic serves as a time-saving technique for survey respondents.

Our results extend the previous work on anchoring and adjusting into a context of particular interest to educational and psychological researchers. In particular, the results provide evidence that anchoring and adjusting generalizes from asking factual questions in which the respondent is uncertain (e.g., Epley & Gilovich, 2006) to self-report contexts in which people record their own beliefs and attitudes. However, the major implications of this work address the pragmatic concerns of survey designers in providing guidance on how to order their surveys.

Balancing Competing Tensions

Our data indicate that questionnaire designers need to think strategically about the best way to organize survey instruments to alleviate anchoring and adjusting. Our results suggest that survey designers should avoid grouping items that assess the same concept together – data fidelity can be substantially degraded by doing so. Yet, we also reviewed literature indicating that to the extent that respondents do search their memories more effectively, they may produce more accurate opinions (Knowles, 1988; Tourangeau, et al., 2000). Thus, mixing items from different constructs might not be optimal. Furthermore, taking Bradburn, Sudman, and Wansink's (2004) idea that surveys are extensions of conversations, designers want to be careful not to come across as scattered and disorganized – conversations usually follow a clear topical

trajectory. Thus, survey designers need to balance the competing tensions between mitigating anchoring and adjusting while still facilitating respondents' memory search and retrieval processes.

We expected that presenting survey items one at a time (at least for web-surveys) would be an effective compromise – the logical flow of the survey would be preserved but respondents anchors would visually disappear before each subsequent item was presented. In fact, Tourangeau, Couper, and Conrad (2004) found that grouping similar items with the same response scale together on the same screen in a web survey produced higher reliability coefficients ($\alpha = .62$) than when each item was presented on a separate screen ($\alpha = .51$; see experiment six). In light of the present findings, it seems possible that their findings illustrate another instance of anchoring and adjusting when all items were presented on a single screen simultaneously and that presenting items one-at-a-time reduced respondents' anchoring and adjusting. However, without access to their raw data, we can only speculate as to whether their respondents also employed the anchoring and adjusting heuristic. Although this potential solution for ordering items on web-surveys appeared promising, it did not appear to substantially mitigate the anchoring and adjusting for the high school students in our third experiment. Additional studies (with larger data sets) would be illuminating here.

A second alternative is to group items from distinct but related scales into the same section of a survey and intersperse them within that section (taking care to avoid placing items from the same construct adjacent to one another). For example, a designer might provide section instructions stating that the next group of questions will ask them how interesting, important, and enjoyable they find their psychology class. Respondents would then answer questions on all three topics that were mixed together. Although three conceptually different scales would emerge for interest, import, and enjoyment, the respondents could continually reflect on their psychology class. Similarly, survey designers could ask respondents to think about their motivation in a particular domain and might ask about their goals, efficacy, and values in that domain.

Although our data do not speak directly to the viability of this option, it may be possible to mitigate anchoring and adjusting by varying the wording of response scales used for each item within a scale. For example, a *course satisfaction* scale might ask "Overall, how satisfied were you with the course?" by using responses ranging from "not at all satisfied" to "extremely

satisfied" and then ask "How much did you enjoy the lectures?" by using "did not enjoy at all" to enjoyed a great deal" as response options. Although both items would contribute to the course satisfaction scale, the different response options might be sufficient to mitigate using the first item as an anchor. Future studies that test this possibility of using heterogeneous response anchors would also be particularly valuable.

Although it certainly requires more work on the part of the survey researchers, a final option is to randomly assign participants to different forms of the survey as was done in the present experiments. Through this approach, researchers can document the extent to which respondents are anchoring and adjusting and can better assess the heuristic's impact on reliability estimates and subsequent correlations with other measures.

Limitations and Future Directions

Although this investigation presents findings with important implications for survey designers, several limitations of the study could be explored through future research. First, these experiments shed little light on the question of who is particularly susceptible to anchoring and adjusting. Although we found no clear differences among subpopulations in our studies, our samples may have suffered from a restriction of range on key characteristics. For example, respondents of lower educational levels might make use of this heuristic more frequently as a means to reducing cognitive effort, and we simply could not detect this tendency because each sample was relatively homogeneous with regard to educational level (see footnote 5). Assessing respondents' motivation while taking the survey (or at the end of the survey) might also provide insights into who anchors and adjusts, although results have been mixed in the classic anchoring and adjusting literature as to whether motivation accounts for differences in the use of the heuristic (Chapman & Johnson, 2002; Epley & Gilovich, 2006).

The experiments (particularly the first and third) would benefit from greater power. Each experiment was built into pre-existing surveys that were designed to answer a substantive research question. The benefit of this approach is that our results represent real illustrations of what can happen to survey data in the field – each experiment has high ecological validity. However, this choice came at the cost of dedicating large numbers of participants to testing our hypotheses exclusively.

Finally, if future studies were to examine anchoring and adjusting in more controlled environments, scholars might gain more direct evidence of anchoring and adjusting as the

mechanism in question. A study that tracked participants' response times *for each item* might be especially useful. Studies comparing scales under more conditions might also yield more information about the conditions necessary for anchoring and adjusting to occur. A 2 X 2 design in which items of the same construct are placed adjacently versus non-adjacently and items have the same or different response anchors would also illuminate the extent to which anchoring and adjusting is driven by similar content in the question stem versus having identical response options (see footnote 4).

Because surveys are so central to social science research, it seems critical that scholars have a strong empirical basis from which they can make survey design decisions. This article borrows the concept of anchoring and adjusting from social psychology to demonstrate how the ordering of survey scales can inadvertently introduce substantial error into survey responses – error that could mislead researchers' understandings of the theoretical and applied implications of their work. Hopefully, additional research can build on this investigation to develop best practices to help investigators help respondents to overcome this bias.

Acknowledgements:

This research was made possible through generous support from the Teachers for a New Era project at the University of Connecticut. We are grateful for the tremendous assistance of Jason Irizarry in Experiment 1 and Kirstie Paul, Madeline Scott, Rebecca Zazove, Anna Harris, and Maureen Brinkworth in Experiment 3. The earlier drafts of this manuscript benefitted greatly from the incisive comments of Ben Cook and Hahrie Han. Daniel Koretz provided invaluable feedback on a later draft. Portions of this research were presented at the American Educational Research Association's annual conference in Chicago (April, 2007), at the Center for Survey Research at the University of Massachusetts-Boston (August, 2008), and at the American Association for Public Opinion Research conference in Hollywood, FL (May, 2009).

Correspondence concerning this article may be sent to Hunter Gehlbach at Hunter_Gehlbach@gse.harvard.edu.

References

Ames, D. (2004). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*(5), 573-585.

Banks, J. A., & Banks, C. A. M. (2001). *Multicultural education: Issues and perspectives* (4th ed.). New York: John Wiley.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design -- for market research, political polls, and social and health questionnaires* (Rev. 1st ed.). San Francisco: Jossey-Bass.

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment.* (pp. 120-138). New York, NY: Cambridge University Press.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113-126.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Newbury Park, CA: Sage.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: J. Wiley.

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin, 30*(4), 447-460.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4), 311-318.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 363-373.

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage Publications.

Gehlbach, H., Brinkworth, M. E., & Harris, A. D. (in press). Changes in teacher-student relationships. *British Journal of Educational Psychology*.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Harrison, D. A., & McLaughlin, M. E. (1996). Structural properties and psychometric qualities of organizational self-reports: Field tests of connections predicted by cognitive theory. *Journal of Management, 22*(2), 313-338.

Irizarry, J., & Gehlbach, H. (2007, April). *Differences in multicultural competence: Moving toward a model of multicultural teacher development.* Paper presented at the American Educational Research Association, Chicago.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*(2), 312-320.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). Patterns of Adaptive Learning Study. Retrieved November 8, 2000, from http://www.umich.edu/~pals/PALS%202000_V13Word97.pdf

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35*(2), 136-164.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*(2), 91-106.

Roeser, R. W., Midgley, C., & Urdan, T. C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology*, *88*(3), 408-422.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly, 68*(3), 368-393.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Washington, J., & Evans, N. J. (1991). Becoming an ally. In N. J. Evans & V. A. Wall (Eds.), *Beyond tolerance: Gays, lesbians and bisexuals on campus*. Alexandria, VA: American College Personnel Association.

Table 1: Between-item distances and mean correlations within Form 1 for Experiment 1 (Research Question 3).

	Distance		Correlation	
	Adjacent *M* (*sd*)	Non-adjacent *M* (*sd*)	Adjacent *r*	Non-adjacent *r*
Awareness	.53 (.28)	.65 (.30)	.57	.40
Knowledgeable	.53 (.34)	.64 (.28)	.57	.42
Motivated	.55 (.26)	.70 (.28)	.61	.44
Skillful	.60 (.38)	.66 (.38)	.45	.38
Frequency of Action	.84 (.42)	.92 (.42)	.48	.41

Table 2: Correlations between focal scales and related measures for Forms 1 and 2 in Experiment 1 (Research Question 4).

	Teaching efficacy			Social perspective taking		
	r for Form 1 (n = 94)	r for Form 2 (n = 78)	Difference	r for Form 1 (n = 94)	r for Form 2 (n = 78)	Difference
Awareness	.59***	.49***	-.10	.14	.40***	.25
Knowledgeable	.58***	.41***	-.16	.14	.24*	.10
Motivated	.67***	.42***	-.25	.07	.22*	.15
Skillful	.42***	.40***	-.03	.29**	.49***	.20
Frequency of Action	.45***	.30**	-.15	.23*	.18	-.05

† $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$

Notes: 1) Stars underneath the columns labeled "Form 1" or "Form 2" indicate whether the correlation differed significantly from 0.

2) The difference scores presented here intended to be descriptive; omnibus significance tests were conducted on all difference scores and are reported in the text.

3) Because the pattern of correlations looks so different for teaching efficacy (all correlations being greater for Form 1) as compared to social perspective taking (all but one correlation greater for Form 2), we also assessed whether these differences in correlations were significant. Only the *motivated* X *teaching efficacy* correlation differed by Form (*Fisher's z* = 2.33, $p = .02$).

Table 3. Between-item distances and mean correlations within Form 1 for Experiment 2 (Research Question 3).

	Distance		Correlation	
	Adjacent *M* (*sd*)	Non-adjacent *M* (*sd*)	Adjacent *r*	Non-adjacent *r*
Academic	.99 (.70)	1.04 (.65)	.43	.31
Interpersonal	.58 (.51)	.72 (.63)	.54	.38

Note: Items from the *social* scale were not included because they consisted of only three items, thus, these comparisons of distance and correlations between items pairs were based on too few items to be meaningful.

Table 4. Correlations between importance scales and related measures for Forms 1 and 2 in Experiment 2 (Research Question 4).

	Academic scale			Social scale			Interpersonal scale		
	Form 1	Form 2	Diff.	Form 1	Form 2	Diff.	Form 1	Form 2	Diff.
Scales									
Academic scale, development	.36**	.53**	.17	.36**	.56**	.20	.47**	.55**	.08
Social scale, development	.35**	.36**	.01	.50**	.65**	.15	.43**	.62**	.19
Interpersonal scale, development	.28**	.35**	.07	.46**	.59**	.13	.52**	.69**	.17
Business core skills, importance	.65**	.62**	-.03	.31**	.50**	.19	.48**	.49**	.01
Business curricular areas, importance	.49**	.52**	.03	.33**	.51**	.18	.47**	.60**	.13
Co-curricular involvement	.10	.08	-.02	.24**	.21**	-.03	.24**	.17**	-.07
Co-curricular development	.13*	.08	-.05	.26**	.30**	.04	.27**	.25**	-.02
Altruism	.28**	.35**	.07	.42**	.53**	.11	.39**	.52**	.13
Re-think beliefs	.07	.19**	.12	.32**	.33**	.01	.22**	.21**	-.01
Single-item indicators									
Own financial support	.02	.01	-.01	-.14*	.1	-.24	-.14*	.07	.21
Family financial support	-.10	-.02	-.08	-.03	-.13*	-.1	.03	-.06	-.09

† $p < .1$; * $p < .05$; ** $p < .01$; *** $p < .001$

Notes: 1) Stars underneath the columns labeled "Form 1" or "Form 2" indicate whether the correlation differed significantly from 0.

2) The difference scores presented here are descriptive; omnibus significance tests were conducted on all difference scores and are reported in the text.

Table 5: Between-item distances and mean correlations within Form 1 for Experiment 3 (Research Question 3).

	Distance		Correlation	
	Adjacent *M* (*sd*)	Non-adjacent *M* (*sd*)	Adjacent *r*	Non-adjacent *r*
Anxiety	.86 (.50)	1.07 (.55)	.42	.36
Boredom	.86 (.68)	.89 (.58)	.53	.55
Similarity	.64 (.54)	.68 (.56)	.55	.49
Enjoyment	.65 (.49)	.65 (.48)	.74	.71

Table 6. Correlations between importance scales and related measures for Forms 1 and 3 in Experiment 3 (Research Question 4).

	Anxiety			Enjoyment			Boredom			Similarity		
Scales	Form 1	Form 3	Diff.	Form 1	Form 3	Diff.	Form 1	Form 3	Diff.	Form 1	Form 3	Diff.
TSR-negative	-.04	.34**	.38	-.47**	-.49**	-.02	.52**	.58**	.06	-.39**	-.40**	-.02
TSR-positive	.08	-.10	-.17	.77**	.79**	.02	-.71**	-.64**	.07	.59**	.70**	.11
SPT propensity	-.06	.10	.16	.14	.50**	.36	-.27*	-.36**	-.09	.24*	.41**	.17
Effort	.12	-.14	-.26	.58**	.60**	.01	-.51**	-.56**	-.05	.35**	.46**	.11
Self-efficacy	-.14	-.45**	-.30	.55**	.59**	.04	-.49**	-.53**	-.05	.51**	.49**	-.02
Belonging	.06	-.20	-.26	.14	.29*	.16	-.28*	-.12	.16	.02	.18	.16

† $p < .1$; * $p < .05$; ** $p < .01$

Notes:

1) TSR = Teacher-student relationship (positive or negative) scales; SPT = Social perspective taking
2) Shaded cells indicate anticipated negative correlations.
3) Stars underneath the columns labeled “Form 1” or “Form 3” indicate whether the correlation differed significantly from 0. The difference scores presented here are descriptive; omnibus significance tests were conducted on all difference scores and are reported in the text.

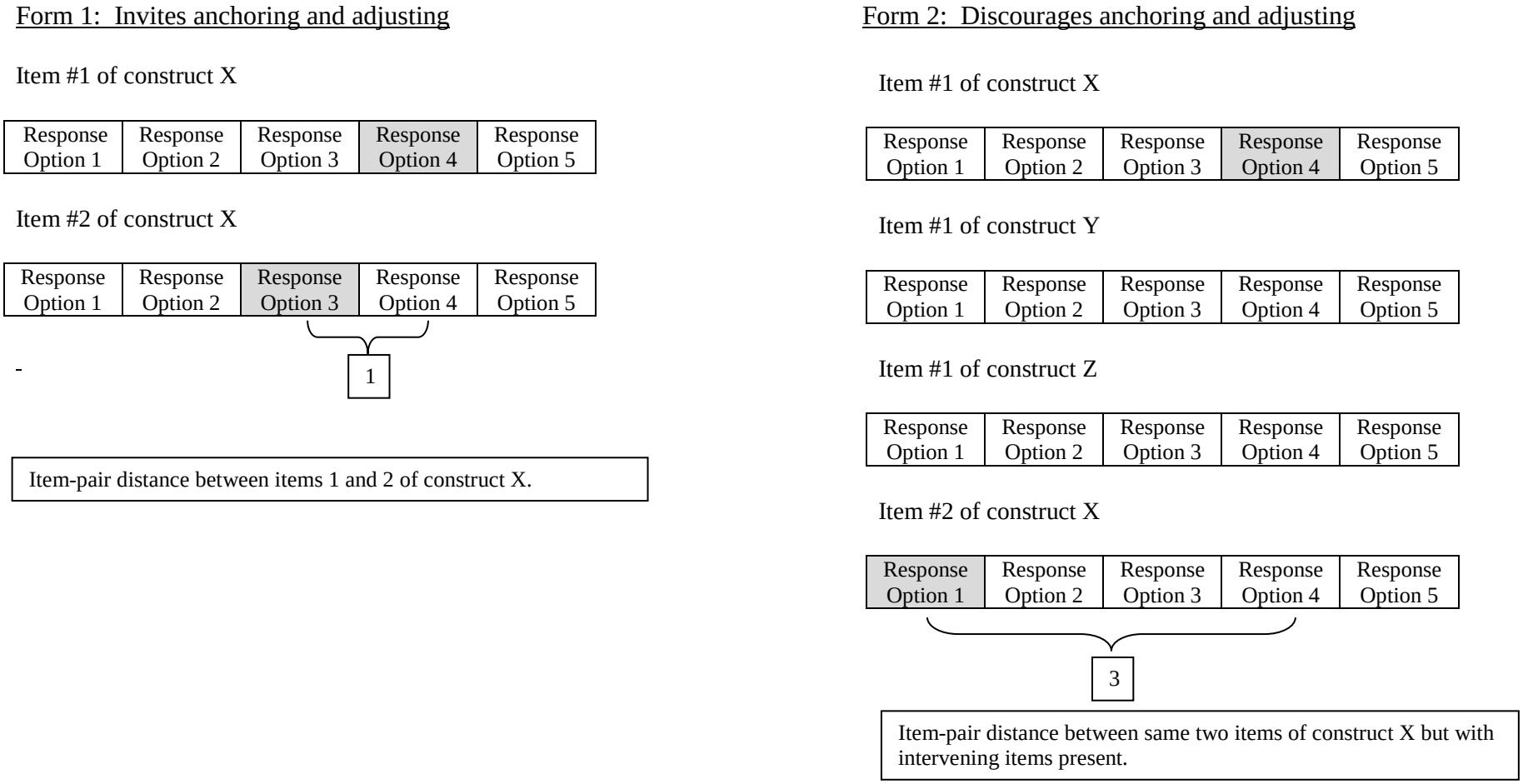

Figure 1. Comparing how different organization of identical survey items might lead respondents to anchor and adjust under some survey conditions. Grey cells indicate selected responses on the item-pairs of interest.

Appendix A: Experiment 1 section of the survey form, question order, and item-pairs of interest

Question	Banks & Banks' (2001) topics	Washington & Evans' (1991) scale	Survey Order Form 1	Survey Order Form 2
When teaching, how aware are you of the need to help students understand where knowledge comes from?	Epistemology	Awareness	16	16
How would you rate your awareness of the ways that knowledge is constructed within your discipline(s)?	Epistemology	Awareness	17	17
How aware are you of opportunities to get culturally diverse resources for your lesson plans?	Content	Awareness	18	26
While you are creating lesson plans, how aware are you of the need to include culturally diverse content?	Content	Awareness	19	27
While preparing lessons, how aware are you of the need to address issues of equity	Equity	Awareness	20	36
How aware are you of the need to modify your teaching practices to address the needs of culturally diverse learners?	Equity	Awareness	21	37
How aware are you of prejudice against your students?	Prejudice	Awareness	22	46
How aware are you of the ways in which prejudice influences students' learning in your classes?	Prejudice	Awareness	23	47
How aware are you of the influence of your school's culture on the experiences of culturally diverse students?	Change culture	Awareness	24	56
How aware are you of the ways your own culture influences the classroom environment?	Change culture	Awareness	25	57
How knowledgeable are you of methods to help students understand multiple sides of debates in your discipline(s)?	Epistemology	Knowledge	26	18
In terms of knowing different ways that knowledge is constructed in your field, how knowledgeable do you feel that you are?	Epistemology	Knowledge	27	19
How knowledgeable are you regarding resources that indicate similarities between distinct cultures?	Content	Knowledge	28	28
Regarding culturally diverse content in your discipline(s), how knowledgeable do you consider yourself?	Content	Knowledge	29	29

Notes: 1) The five scales of interest consisted of 50 items, 14 of which are presented here for illustrative purposes. Please contact the first author for the full scales.
2) "Survey order" – as presented in the two right-hand columns of the table – reflects the actual number of each item on the respective forms of the survey. (The *Teaching Efficacy* and *Social Perspective Taking* scales were presented in items 1-15).
3) The boxes identify the item-pairs of interest, i.e., those items which are adjacent on Form 1 (but non-adjacent on Form 2), assess the same construct, and use the same response anchors. Because each scale contains four item-pairs of interest and there are five scales, we examined a total of 20 item-pairs of interest.

Appendix B: Experiment 2 item arrangement and response anchors for forms 1 and 2

Question stem

Please rate how IMPORTANT each of the following skills & competencies has been in your life since college.

Response anchors

Not Important – 1 – 2 – 3 – 4 – Very Important

Items	Form 1 Order	Form 2 Order
Academic Scale		
Write effectively in English	1	1
Communicate well orally in English	2	5
Use quantitative tools	3	3
Synthesize and integrate ideas and information	4	11
Gain in-depth knowledge in a field	5	13
Acquire new knowledge/skills on my own	6	16
Understand the processes of science and experimentation	7	8
Social Scale		
Identify moral and ethical issues	8	9
Understand current social problems	9	17
Consider how my beliefs and/or faith inform my actions	10	15
Interpersonal Scale		
Resolve conflicts between people positively	11	2
Function effectively as a member of a team	12	6
Act in the interests of the communities I belong to	13	7
Relate well to people different from me	14	10
Identify and understand cultural differences	15	12
Lead and supervise tasks and groups of people	16	4
Consider the role of the leader in helping others	17	14

Appendix C: Experiment 3 item arrangement and response anchors for Forms 1 and 3

Question stem

Please indicate the most accurate response in each case.

Response anchors

Not at all – A little – Somewhat – Quite – Extremely

Items	Form 1 Order	Form 3 Order
Anxiety Scale		
How concerned do you feel about understanding the material in this class?	1	2
When you think about this class, how uneasy do you feel?	2	6
How nervous does this class make you?	3	10
How worried are you that you have prepared enough for this class?	4	18
How tense do you feel during this class?	5	14
Enjoyment Scale		
How excited are you about going to this class?	6	3
How enjoyable is being in this class?	7	7
At the end of class, how eager are you for your next class with Ms. G?	8	15
How enjoyable is it to listen to Ms. G in this class?	9	11
Overall, how enjoyable is the learning that you do in this class?	10	19
Boredom Scale		
How dull do you find the discussions in class?	11	20
When you are sitting in class, how hard is it to keep your mind from wandering?	12	4
In class, how frequently do you think about what else you might be doing?	13	12
During class, how hard is it to stay alert?	14	8
Overall, how boring do you find this class?	15	16
Similarity Scale		
How similar are your values to Ms. G's values?	16	1
How similar is your background to Ms. G's background?	17	9
How interested are you and Ms. G in the same activities?	18	5
How easy is it for you to think of things that you and Ms. G have in common?	19	17
Overall, how similar do you think you and Ms. G are?	20	13