The Role of Context-Driven Response Bias on the Standard Anchoring Effect

by

Cory Tam

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Psychology
University of Alberta

Abstract

Anchoring is judgmental bias in which quantitative estimates assimilate to seemingly irrelevant numerical reference values (Tversky & Kahneman, 1974). While Tversky and Kahneman (1974) originally proposed that anchoring results from the application of a deliberate anchoring-and-adjustment heuristic, other researchers have stressed the role of automatic processes with the introduction of priming-based accounts (Mussweiler & Strack, 1999). In this paper, we present a new perspective on anchoring called *consistency theory.* On this view, people first determine whether the true target value is above or below the anchor value, and then they provide an estimate that is consistent with the "Greater" or "Less" judgment. Differing from the selective accessibility account, consistency theory assumes that people can be affected by factors such as the response format of the initial comparative judgment. As predicted, we obtained context effects—participants' judgments of *target* items were influenced by their judgments of *filler* items. That is, participants responded "Greater" more often for the target items when they had made fewer "Greater" judgments for the fillers items, and vice versa. Overall, these findings suggest that people can interact with numerical information in a number of different ways, which challenges the view that anchoring is driven by automatic, activation-based processes.

Keywords: judgmental anchoring, heuristics and biases, real-world estimation, context effects, response bias, selective accessibility

Preface

This thesis is an original work by Cory Tam. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name "MAKING AND JUDGING NUMERICAL ESTIMATES", No. 34351, October 11, 2012.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Norman Brown, for his support, guidance, and expertise throughout this entire process. Thank you also to the members of my committee, Dr. Anthony Singhal, Dr. Peter Dixon, and Dr. Sarah Moore, for their thoughtful questions and insightful comments. Finally, I am grateful for having Oliver Schweickart and Connie Svob as my lab mates and friends; the PSYCO 496/498 students for helping me collect the data; and the "Tamily" for their love and support.

Table of Contents

List of Tables

List of Figures

The Role of Context-Driven Response Bias on the Standard Anchoring Effect

The standard anchoring effect is one of the most prominent biases in the judgment and decision making literature. It refers to the finding that quantitative estimates are commonly biased in the direction of numerical reference values, so-called *anchors* (Kahneman, 2011; Tversky & Kahneman, 1974). The standard anchoring paradigm consists of a two-step procedure: first, participants have to determine whether a specific quantity (e.g., the number of babies born in Canada last year) is greater or less than an anchor value (e.g., 500,000). Then, they are asked to give an estimate for that quantity (e.g., "How many babies were born in Canada last year?"). Experiments using this task show a robust anchoring effect, such that participants' estimates, on average, are closer to the anchor when compared to unanchored estimates (Jacowitz & Kahneman, 1995).

Over the past 40 years, the anchoring effect has been widely replicated in a number of domains. Anchors have been shown to affect people's probability estimates (Chapman & Johnson, 1999; Tversky & Kahneman, 1974), willingness to pay for consumer goods (Ariely, Loewenstein, & Prelec, 2003), estimates of real-estate prices (Northcraft & Neale, 1987), and criminal sentencing decisions (Englich, Mussweiler, & Strack, 2006). Anchoring effects occur even when people are explicitly told that the provided values were randomly generated (Chapman & Johnson, 2002; Russo & Shoemaker, 1989). In fact, in the landmark anchoring experiment conducted by Tversky and Kahneman (1974), the anchor values for the initial comparison question were presented as being arbitrary, coming from a (rigged) wheel of fortune that was spun in front of the participants.

Given that we are constantly exposed to new numerical information, it is important to understand how we interact with and manage this information. Current theories of anchoring suggest that people often have difficulties discriminating between relevant and irrelevant numerical information, and that even arbitrary numbers can contaminate people's judgments (e.g., Mussweiler & Strack, 1999; Tversky & Kahneman, 1974; Wilson, Houston, Etling, & Brekke, 1996; for a recent review, see Furnham & Boo, 2011). In this paper, we outline an alternative account, called *consistency theory*. Contrary to priming-based accounts in which anchor values bias people's beliefs about quantities, consistency theory predicts that the comparative assessment of anchor values, as well as subsequent estimates are affected by a response bias. It is predicted that the anchor values of *filler* items will affect responses to *target* items. We should observe a difference for target items that have been mixed in with filler items with *high* anchor values, compared to target items that have been mixed in with filler items with *low* anchor values. We test this prediction against the selective accessibility account, and end with a discussion of the broader implications of our findings with respect to how people deal with numerical information.

Theories of Anchoring

Tversky and Kahneman (1974) proposed that anchoring results from the application of a deliberate anchoring-and-adjustment heuristic. According to their classic account, people first determine whether they think the presumed true value of the target is greater or less than the provided anchor value. Then, they make adjustments, starting from the anchor to where they think the presumed true value lies. However, the adjustments are considered to be insufficient because final average estimates still remain shy of average unanchored estimates. The most common explanation for the insufficiency of adjustment is that people terminate the adjustment process once they reach the range of plausible values, which is typically too soon (Epley & Gilovich, 2006). Note that from the participants' viewpoint, they believe they have arrived at a suitable estimate, but the experimenter still sees the estimate as "insufficient" because it does not match estimates from a control group.

However, there are a set of findings that are commonly interpreted as being inconsistent with the idea that people use a deliberate anchor-and-adjust strategy. First, if people were deliberately adjusting from the anchor, providing an accuracy incentive (e.g., offering participants a chance to win money if they provided the most accurate estimate) should motivate them to adjust further away from the anchor value, and beyond just the beginning of the plausible range. However, incentives typically do not decrease the gap between anchored and unanchored estimates (Epley & Gilovich, 2005; Tversky & Kahneman, 1974; but see Simmons, LeBoeuf, & Nelson, 2010). In addition, explicit forewarnings about people's tendency to adjust insufficiently did not affect the amount of adjustment (Epley & Gilovich, 2005). Again, if people were deliberately adjusting away from the anchor, they should be able to modify their performance after being warned. Finally it appears that the adjustment process is unaffected by cognitive load in a dual-task scenario (Blankenship, Wegener, Petty, Detweiler-Bedell, & Macy, 2008; Epley & Gilovich, 2001). Participants in the high cognitive load condition were given a more cognitively demanding task (e.g., answering questions while listening to a string of letters and reporting the number of vowels) than participants in the low cognitive load condition; the former should have interfered with the effortful adjustment process and lead to greater anchoring effects, but it did not. Therefore, these results, among others, contributed to the rise of the current consensus view of anchoring as being driven by automatic processes.

In contrast to the anchoring-with-insufficient-adjustment position, automatic accounts propose that anchoring can still occur without an effortful adjustment process[1] (Mussweiler & Strack, 1999; Strack & Mussweiler, 1997). According to this view, people engage in a confirmatory hypothesis-testing process, in which they assess whether the provided anchor is equal to the presumed true value of the target question. As a result, semantic knowledge consistent with the anchor becomes increasingly accessible. This selective accessibility model suggests that people produce estimates that are biased toward the anchor because they rely on this easily accessible, anchor-consistent knowledge for generating their post-comparison estimates. On this view, anchoring is driven by an automatic semantic priming process, and as a result, the anchoring bias is extremely difficult, if not impossible, to avoid (Mussweiler, Strack, & Pfeiffer, 2000).

More recently, anchoring researchers have proposed accounts that incorporate both automatic and controlled processes: dual process models (e.g., Kahneman, 2011; Wegener, Petty, Blankenship, & Detweiler-Bedell, 2010a, 2010b; see Frederick, Kahneman, & Mochon, 2010 for a critical review), and integrative accounts of adjustment and selective accessibility (Chaxel, 2014; Simmons et al., 2010).

One dual-process approach is based on an "attitude change" perspective, which applies the Elaboration Likelihood Model to anchoring (Petty & Cacioppo, 1986). The model suggests that anchors can be processed in relatively non-thoughtful (peripheral or heuristic), or relatively thoughtful (central or systematic) ways. Different mechanisms are assumed to underlie anchoring effects elicited under these high-elaboration and low-elaboration conditions. These include selective accessibility, for high-elaboration conditions, and the activation of knowledge via the anchor (e.g., numeric or magnitude priming), using the anchor as a "hint" for the target judgment (Schwarz, 1994), and a shallow form of selective accessibility in which people only partially employ a confirmatory hypothesis test, for low-elaboration conditions (Wegener et al., 2010a). In addition, a person's ability to elaborate depends on his or her ability (i.e., knowledge level) and motivation (Blankenship et al., 2008; Wegener, et al., 2010a, 2010b; Wegener, Petty, Detweiler-Bedell, & Jarvis, 2001).

Most recently, another dual-process approach has been proposed. Here, anchoring is described as a result of two mechanisms: anchoring as a priming effect, an operation of the automatic, involuntary "System 1", and anchoring as adjustment, an operation of the deliberate, effortful "System 2" (Kahneman, 2010). System 1 retrieves anchor-consistent information from memory (i.e., the anchor is the true value), and then System 2 works on this (biased) information

(Kahneman, 2010). The emphasis of this dual process approach is on the two systems interacting with one another.

Then there are integrative theories of anchoring, which are based on a framework that merges the anchoring-with-insufficient adjustment and selective accessibility theories. Simmons et al., (2010) proposed that selective accessibility is not the only process underlying anchoring effects, and suggests that people do effortfully adjust from anchors. They found that the ability to increase the distance between anchors and final estimates (called anchor-estimate gaps), was contingent upon whether people were certain about the direction of adjustment, and their motivation (Simmons et al., 2010). However, these findings largely focused on the anchoring-and-adjustment process, and the specific role of selective accessibility was unclear.

To fill in this gap, Chaxel (2014) manipulated effort with a word-string memorization task, but also primed selective accessibility through a similarity search (i.e., asking participants to list similarities between two scenes). The main finding was that selective accessibility interacted with effort, which has traditionally been used to differentiate between theories of anchoring-and-adjustment and selective accessibility (Frederick et al., 2010). Though a different methodology was used, this provides support for the view that both anchoring-and-adjustment and selective accessibility processes may be operating in parallel.

Anchoring effects have the reputation of being "easy to generate but hard to explain" (Frederick et al., 2010, p.17). This review of the literature suggests that a variety of processes, which work more or less in unison, could underlie the standard anchoring effect. However, we propose an alternative account, which incorporates minimal assumptions, and can be used as a framework for a unified account of judgmental anchoring.

An Alternative Account: Consistency Theory

Consistency theory assumes that people use different strategies when evaluating anchors (Brown & Moore, 2003; Wegener et al., 2001). It also assumes that people have a great deal of control over how they interact with numerical information, and that people provide target estimates that are consistent with the just-preceding comparative judgment (Jacowitz & Kahneman, 1995).

On this view, people assess the potential relevance of the anchor value for a target judgment, and take it into consideration if they deem it relevant (i.e., plausible; also see Ariely et al., 2003). However, there is a *range of metric indifference* due to people's varied knowledge of real-world quantities. This range encompasses values that people view as being roughly equally plausible. For a low knowledge item (e.g., "What is the distance between Edmonton and Sydney, Australia?"), the range is rather large. In this instance, there are many possible anchor values (e.g., 10,000km, 15,000km, and 20,000km) that might be viewed as a plausible answer. On the other hand, for a high knowledge item (e.g., "How many calories are in a McDonald's 'Big Mac'?"), the range of metric indifference is most likely much smaller.

If the anchor value falls outside the plausible range, some participants will reject it and instead, form an independent estimate. Consistent with this view is the finding that extreme anchors sometimes elicit a smaller anchor size effect than moderate anchors (Wegener et al., 2001), and that implausible anchors are rejected much more frequently than plausible ones (Brown & Moore, 2003). If the anchor value falls within the range of metric indifference, people will have to guess because the comparative judgment is difficult, and then adjust to provide a consistent estimate. Adjustments are small (Simmons, et al., 2010), and are usually rounded to the nearest spontaneous number (Albers, 2001). Therefore, as long as the anchor falls inside the range for at least a subset of participants, an aggregate anchoring effect is virtually guaranteed to emerge.

Finally, even though the comparative judgment may be made under some uncertainty, it is almost inevitably followed by a numerical estimate that is *consistent* with it (Jacowitz & Kahneman, 1995). Given the structure of the standard anchoring paradigm, once people have decided whether the true target value is likely to be above or below the anchor, they adjust *away* from the anchor in the indicated direction. And the estimate that follows will fall within the subjective range of plausible values. Note that like Tversky and Kahneman (1974), we are proposing an anchoring-and-adjustment-like account. However, we argue that the adjustments are almost always *sufficient* given the task demands.

The Current Research

The aim of the present study was to test one important implication of consistency theory. Here, we focus on a key finding of the standard paradigm that is typically considered to support priming-based accounts of anchoring. The finding is that even when a *high* anchor is presented (at the 85th percentile of a distribution of estimates in a calibration group), it is still often judged as being too low compared to the target quantity. Because the high anchor is taken from the 85th percentile, it is expected that participants will respond “Greater” for the target judgment 15% of the time. However, Jacowitz and Kahneman (1995) found that 27% of judgments were “Greater” responses. And because anchored estimates are almost always consistent with the prior judgment (99.5%; Jacowitz & Kahneman, 1995), the percentage of estimates greater than the anchor value was also greater than expected, compared to unanchored estimates.

When Mussweiler and Strack (1999) changed the wording of the comparative question slightly, they reported the same result, but for both high and low anchors. Instead of asking participants whether the target was greater or less than the anchor (i.e., wording from the standard paradigm), they asked one group of participants whether the target object was larger than the anchor, and another group whether the target was smaller than the anchor. With this procedure, the target was still judged as *larger* than the *high* anchor, but also as *smaller* than the *low* anchor more than would be expected based on calibration data (Mussweiler & Strack, 1999). If people were not affected by the anchor, then the percentage of these so-called extreme judgments should be similar to the unanchored or calibration groups. Because of these differences, the finding is interpreted as supporting the view that anchors alter people’s beliefs, and that subsequent judgments about the anchor are based on these altered beliefs (Jacowitz & Kahneman, 1995).

In contrast, the consistency account holds that the effect is due to a response bias; specifically, people's tendency to equalize the use of different response categories (Erlebacher & Sekuler, 1971; Parducci, 1965; Schneiderman & Manis, 1978). In the standard paradigm, we predict that participants would try to respond "Greater" and "Less" about equally when possible, especially given the degree of uncertainty regarding some of the answers. In Jacowitz and Kahenman's (1995) study, the mean confidence rating for their set of questions was only 3.85 on a 10-point scale. Therefore, an *equal frequency response bias* could account for the present finding without assuming that the comparative judgment alters people's beliefs about the magnitude of the target object.

To test this response bias hypothesis, we manipulated context through the anchor values assigned to a set number of what were called, *filler* items. In the *high-context* condition, the anchors for these fillers were from the 90th percentile of the distribution of estimates in the calibration group. In the *low-context* condition, they were from the 10th percentile, and in the *median-context* condition, from the 50th percentile. When anchors for the fillers are taken from the high end of the distribution of normative estimates, we expect a few "Greater" judgments (~10%) and many "Less" judgments. Conversely, when the filler values are drawn from the low end of the distribution, we expect many "Greater" judgments (~90%) and few "Less" judgments. Because the fillers were the subjective medians in the median-context condition, the proportion of "Greater" judgments should be roughly equal to the proportion of "Less" judgments. Thus, on average, we expect the proportion of "Greater" judgments to be around 50%.

Critically, *target* items were also mixed in with the fillers items. The anchor values for the target items were always the medians of the calibration group. Selective accessibility posits that estimates following the comparative judgment should be biased towards anchors irrespective of how the anchor was judged. This means the expected proportion of “Greater” judgments would be predicted to be about equal for all target items, regardless of the context manipulation. Because the anchors are the calibration medians, there should be no overall anchoring effects.

In contrast, consistency theory predicts a context effect. The frequency of “Greater” judgments for fillers should impact the frequency of “Greater” judgments for targets. Therefore, assuming that the high-context condition will produce many “Less” responses for the filler items, it follows that we should observe a relatively large number of *greater than* responses for the target items. And assuming that the low-context condition will produce many “Greater” responses for the filler items, we should observe a relatively large number of *less than* responses for the target items. A difference in the average proportion of “Greater” judgments for target items in the high- and low-context conditions should also be expected overall.

This pattern, as a consequence of consistency responding, should be reflected in the target estimates as well. Specifically, estimates should be larger in the high-context, compared to the low-context condition. Finally, if the proportion of “Greater” judgments were around 50% for the fillers in the median-context condition, then the proportion of “Greater” judgments should be around 50% for the targets items as well. This median-context condition will serve as the control.

Method

Participants

One hundred and two University of Alberta undergraduate psychology students (aged 17-28, *Mdn* = 18) took part in this experiment. They were recruited from the Psychology Research Participation Pool, and received partial course credit for their participation. All participants were born in Canada and had English as their native language.

Materials and Design

The stimuli consisted of 38 estimation problems. Ten served as target items (Table 1), and 28 as filler items (Appendix, Table 1). The questions covered a wide range of topics, and the correct answers ranged from 8 (the number of main islands in Hawaii) to over 800,000 (the population of San Francisco). Target items were chosen to be of relatively low knowledge to ensure that participants would not know the exact answers.

Following the standard procedure, each item comprised of a comparative judgment (e.g., "Is the length of the Nile River greater than or less than 800 kilometres?"), and an absolute judgment (e.g., "How long is the Nile River?"). The anchor values for all questions were obtained from a calibration group. Participants in the calibration group (*n* = 74) were recruited first, and provided unanchored estimates and confidence ratings to 48 randomly selected questions from a larger set of 80.

The anchor values for the target items were always the median values. In other words, the anchors corresponded to the 50^{th} percentile of the distribution of estimates in the calibration group. The anchor values of the filler items were manipulated to correspond to the 90th percentile (high-context condition), 10th percentile (low-context condition), or 50th percentile (median-context, or control condition) of the distribution of estimates in the calibration group.

The context provided by the filler items was manipulated, via the anchor values, between participants. Furthermore, we counterbalanced (across participants) the order in which the comparative adjectives appeared in the response option (i.e., Greater/Less vs. Less/Greater). All anchor values were rounded to be consistent with the response format.

Procedure

Participants were tested individually on a computer in a lab-based setting. To increase the plausibility of the task, the instructions stated that the purpose of this research was to assess people's real-world knowledge, and that two different techniques would be used to do so. One technique was to have people make judgments about randomly generated answers, and another was for them to estimate target quantities directly. For the former, participants were told that the computer program would generate a different randomly generated answer for each judgment question.

In all conditions, participants made three consecutive judgments for each of the 38 items: a comparative judgment, an absolute judgment, and a confidence judgment. Participants were first presented with the comparative judgment in which the anchor value was substituted with a blank. Then they clicked a button to "generate a random number" that filled the blank, and caused the two response options (*Greater* vs. *Less*, or *Less* vs. *Greater*) to appear below the question. After participants had clicked on and confirmed their choice, they were prompted to come up with a specific answer for the question. When they had a number in mind, they pressed the spacebar and typed in their answer. Finally, participants indicated how confident they were that their estimate was accurate. The confidence rating scale ranged from 1 to 5; participants were instructed to use a 1 when they had absolute no confidence in the accuracy of their estimates, to use a 5 when they believed that their response was correct or very nearly correct, and to use the intermediate values to indicate intermediate levels of confidence.

Item order was pseudo-randomized such that targets and fillers were approximately equally distributed across trials. In the calibration group, participants only provided estimates and confidence ratings.

Results

In this section we focus on the two main dependent measures, proportion of “Greater” responses for the comparative judgments, and the anchoring bias for the post-comparison estimates. All statistical analyses reported below are based on mixed-effects models with participants and items as random effects (Baayen, Davidson, & Bates, 2008).

Table 1

Summary Statistics for Calibration and Anchored Estimates of Target Items

Question	Calibration Median	Confidence	High Context % Greater	High Context Median Estimate	Low Context % Greater	Low Context Median Estimate	Median Context (Control) % Greater	Median Context (Control) Median Estimate
How high is Mount McKinley (in metres)?	4,500	1.34	55.9%	4,500	41.2%	4,000	52.9%	5,000
How long is the Nile River (in km)?	800	1.52	58.8%	1,000	52.9%	900	82.4%	1,000
How many Canadian soldiers died in World War II?	11,000	1.58	70.6%	15,000	58.8%	15,000	70.6%	14,000
What is the manufacturer's suggested retail price of a 2012 Lexus LFA (in US dollars)?	65,000	2.33	64.7%	70,000	58.8%	72,500	55.9%	72,500
How many steps are there in the CN Tower?	1,000	1.54	76.5%	1,650	47.1%	900	73.5%	1,500
How many kilometres did Terry Fox run during his Marathon of Hope?	1,800	1.78	91.2%	3,000	61.8%	2,350	85.3%	2,400
How many rooms are in the MGM Grand Hotel in Las Vegas?	1,000	1.68	73.5%	1,500	55.9%	1,160	64.7%	1,500
How much did Bryan Cranston get paid for each episode of 'Breaking Bad' during its final season (in US dollars)?	200,000	1.84	73.5%	355,000	58.8%	250,000	64.7%	250,000
How many babies were born in Canada in 2012?	300,000	1.72	61.8%	350,000	55.9%	350,000	47.1%	260,000
What was the population of San Francisco in 2012?	2,500,000	2.03	64.7%	3,000,000	50%	2,425,000	50%	2,550,000

Note. Confidence ratings range from 1 (not confident at all) to 5 (very confident).

Context Effects

Filler items. To confirm the effectiveness of the context manipulation, we looked at the proportion of "Greater" judgments for the filler items (see right panel of Figure 1). The proportions were calculated separately for each participant, and then averaged by condition. The observed proportions closely matched the expected proportions. In the high-context condition, participants responded "Greater" for 22.3% of the filler items, compared to an expectation of 10.7%;[2] and in the low-context condition, participants responded "Greater" for 85.6% of the filler items, compared to an expectation of 89.3%. In the median-context (control) condition, they responded "Greater" for 55.5% of the filler items, compared to an expectation of around 50%.

Target Items. The left panel of Figure 1 shows the proportion of "Greater" judgments for the target items. In the high-context condition, the average proportion of "Greater" judgments was 69.1%, and in the low-context condition, the average proportion of "Greater" judgments was 55.3%. There was a critical difference between the two conditions, such that participants responded "Greater" more often in the high-context condition than in the low-context condition, *odds ratio* = 1.88; $z = 3.31$, $p < .001$. Thus, as predicted, context affected participants' target judgments.

This finding holds across the 10 target items. As shown in Table 1, there were more "Greater" judgments given in the high-context than low-context condition. Median estimates were also larger in the high-context than low-context condition for the majority of items.

In the median-context (control) condition, participants responded "Greater" for 63.5% of the target items. Though this is less than the proportion in the high-context condition, and more than the proportion in the low-context condition, it was greater than originally predicted based on the calibration data. However, this finding may provide some insight on another response bias that may be affecting judgments, and will be addressed in the General Discussion.

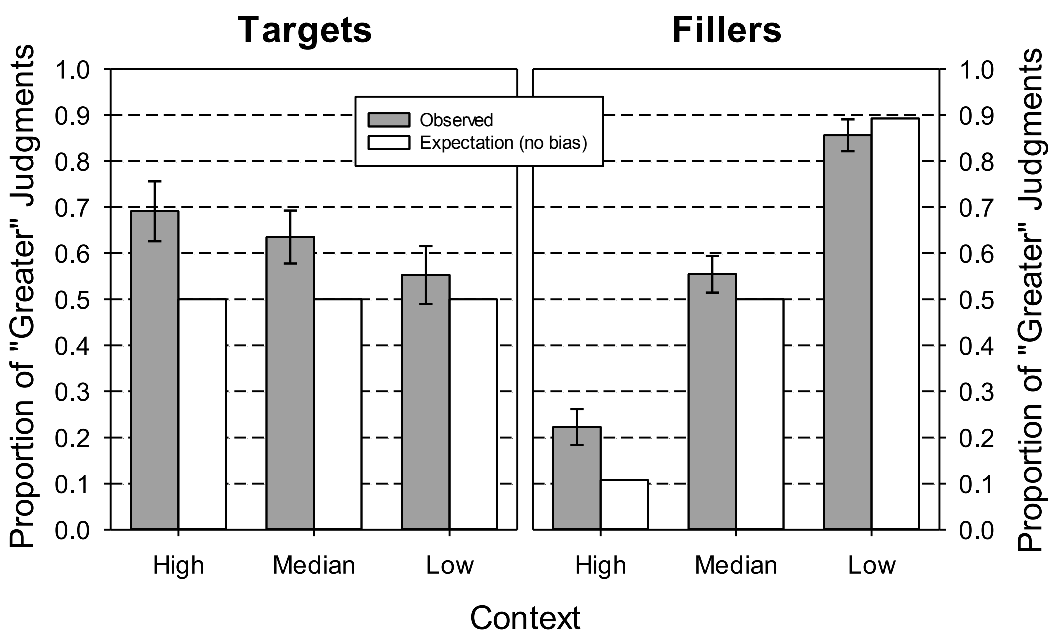

Figure 1. Responses to comparative judgments for target and filler items by condition. The left panel shows the observed and expected proportion of "Greater" judgments for the 10 target items for the high, median, and low context conditions. The right panel shows the observed and expected proportion of "Greater" judgments for the 28 filler items, also for each condition.

Consistency and Absolute Judgments

Consistency. Overall, participants provided estimates greater than the anchor when they had indicated that the target value was greater than the anchor in the comparative judgment (consistency = 99.3%), as predicted. Participants also provided estimates smaller than the anchor when they had indicated that the target value was less (consistency = 99.5%). This result directly replicates Jacowitz and Kahneman's (1995) consistency finding in which participants gave anchored estimates that were 99.5% consistent with the just-preceding judgment. Table 2 shows the data decomposed by question type and condition.

Table 2

Percentage of "Greater" Judgments and Percentage of Estimates Greater Than the Anchor for Target and Filler Items by Condition

	Targets		Fillers	
Condition	Percentage of "Greater" judgments	Percentage of estimates greater than anchor	Percentage of "Greater" judgments	Percentage of estimates greater than anchor
High Context	69.1%	68.5%	22.3%	21.5%
Low Context	55.3%	55.9%	85.6%	85.5%
Median Context	63.5%	63.5%	55.5%	55.7%

Absolute judgments.[3] To determine the degree in which target estimates were influenced by the anchor value, we computed the anchor-based *signed order of magnitude error* (*aSOME;* Brown, 2002; Brown & Siegler, 1993; Nickerson, 1981). It is defined as:

$$\text{aSOME} = \log_{10}(\text{estimate/anchor value})$$

This measure expresses how close an estimate is to a given reference value (the anchor). Positive aSOMEs indicate, on an order of magnitude scale, the degree to which an estimate is greater than the anchor, and negative aSOMEs indicate the degree to which it is smaller. An aSOME of 0 indicates that the estimate is equal to the anchor value.

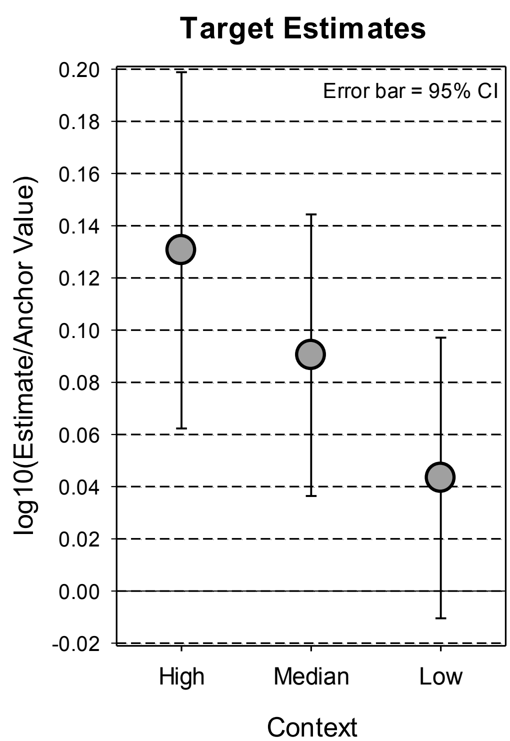

Figure 2. The mean anchor-based signed order of magnitude error (aSOME) for target items in the high, median, and low context conditions.

As predicted, participants gave higher estimates in the high-context condition (M_{aSOME} = 0.13), compared to the low-context condition (M_{aSOME} = 0.04), b = 0.09; t = 2.13, p = .04. In the results described above, we already saw how context affected the proportion of "Greater" judgments. And with consistent responding, we can see here, how the context effect is reflected in the estimates as well.

In all conditions, aSOMEs were positive, indicating that on average, estimates were greater than the anchor. This is also shown with the higher proportion of "Greater" judgments overall. This means that participants believed the target value was usually greater than the anchor, and will be discussed in the next section.

General Discussion

The results of the present study demonstrate an effect of context in the standard anchoring paradigm. We observed with the filler items, that when high anchors were presented, participants responded with fewer “Greater” judgments, and when low anchors were presented, they responded with more “Greater” judgments. The finding of interest though, was how participants judged the target items, in which the anchors were always the normative medians. As predicted, we found that participants attempted to equalize their responses by choosing the under-used alternative (Schneiderman & Manis, 1978). Participants responded “Greater” more often for the target items when they had made fewer “Greater” judgments for the fillers items, and responded “Greater” less often for the targets, when they had made more “Greater” judgments for the fillers. And we observed a difference between the two groups, such that the proportion of “Greater” judgments elicited by the filler items was always higher in the high vs. low context condition.

Furthermore, estimates were consistent with the prior judgments. The subsequent estimates were larger when participants had responded with a higher proportion of “Greater” judgments, and smaller when participants responded with a lower proportion of “Greater” judgments. The estimates reflected the difference between the two groups, with larger estimates given, again, in the high vs. low context condition.

Here, we also discuss an unexpected, but interesting finding from the present study. In the median-context, or control condition, it was expected that participants would respond “Greater” to the target judgment around 50% of the time because the anchor was from the 50^{th} percentile of the calibration group. What we found was that, on average, participants had judged the target as “Greater” 63.5% of the time; this was almost 14% above the expectation. A possible explanation for this comes from research in the domain of binary choice. Bar-Hillel, Peer, and Acquist (in press) reported that when people had to make a choice between two options, they were likely to choose the one that is, based on convention, presented first. When participants were asked to generate a sequence of coin tosses, around 80% of the participants had begun the sequence with “Heads” (Bakan, 1960; Goodfellow, 1940). This response bias is called *reachability* because people seem to favour the more reachable option (Bar-Hillel et al., in press).

In the standard paradigm, “Greater” is the more reachable option. In addition, when English speakers compare two objects that differ in magnitude, they are more likely to use words like *greater* and *higher* (called “larger” comparatives), than words like *less* and *lower* (called “smaller” comparatives; Matthews & Dylman, 2014). We find that this is replicated in our dependent variables. As we mentioned in the results, there were a higher proportion of “Greater” judgments overall, and on average, participants were providing estimates that were greater than the anchor. In addition, this “Greater” bias may have been mitigating the response equalization effect. Though the exact underlying mechanism remains unclear, we do know that this bias was operating in the proposed direction in one of the conditions, but in the opposite of the proposed direction in another condition. In other words, it may have been inflating the proportion of “Greater” judgments in the high-context condition when we predicted a high proportion of “Greater” judgments for the target items, but deflating the proportion of “Greater” judgments when we predicted a low proportion of “Greater” judgments for the target items. Because it is working in opposite directions however, we are confident that it does not change the overall pattern of results we obtained in our study.

Taken together, these effects are inconsistent with automatic, priming-based accounts of anchoring. According to the selective accessibility model, anchoring effects are “inevitable” due to the retrieval of accessible knowledge (Mussweiler et al., 2000; Strack & Mussweiler, 1997). Thus, people anchor to numerical information, regardless of the context. In contrast, consistency theory had proposed that a response bias could account for the findings. That is, even though people show anchoring effects, they are also responding to the task demands of the comparative judgment. As we demonstrated, people’s responses to the target items were influenced by their responses to the filler items. According to consistency theory, as long as the anchor values fall inside the range of metric indifference, people will make an assessment, then provide a consistent estimate.

Another way to distinguish between consistency theory and selective accessibility is to compare the distributions of anchored responses, relative to the distributions of the calibration group. Schweickart, Tam, and Brown (2014) used a comparative question that was framed as a hypothesis test (e.g., “Is 5000 metres a good or bad estimate of the height of Mount McKinley") to look at the distribution of estimates that followed the assessment of the anchor value as a good, or as a bad target estimate. This decomposition by assessment response revealed that there was a mixture of different underlying distributions resulting from the use of different strategies, which only consistency theory predicts. When participants indicated that the anchor value would be a *good* target estimate, participants either adopted the anchor or provided an estimate close to it. Here, the distributions of anchored estimates were tight and centered on the respective anchor values. However, when participants indicated that the anchor would be a *bad* target estimate, they subsequently provided estimates that were similarly distributed as unanchored estimates.

Consistency theory can also explain the striking effects of arbitrary and implausible (extreme) anchor values, and why even people with expertise may be influenced by anchors. In one experiment, participants were asked about the year that Attila the Hun was defeated in Europe (Russo & Shoemaker, 1989). Following the standard procedure, participants had to first consider whether the event occurred before or after a specific year. That year (i.e., the anchor value) was generated separately for each participant, and was *arbitrary* (the last three digits of their phone number plus 400). Russo and Shoemaker (1989) found a strong relationship between the answers to the target question, and the phone numbers that were turned into anchor values. The larger the resulting anchors were, the more recent, on average, the defeat of Attila the Hun was estimated to be. However, this selection procedure resulted in anchor values limited to the range of about 400 to 1400. Note that the correct answer is 451. Therefore, consistency theory would reason that because this is a rather difficult question, it is likely that for a majority of participants, the anchor values were viewed as reasonable answers and well within the plausible range, even if they were derived from people's phone numbers.

In another experiment, participants were presented with implausible anchor values, which deviated from the mean of the calibration group by more than 10 standard deviations. For example, participants were asked whether the mean winter temperature in the Antarctic is higher or lower than 45°C in the high anchor condition, or -210°C in the low anchor condition (Strack & Mussweiler, 1997). The authors found anchoring effects for even these implausible values, which on the surface, support their hypothesis that anchoring results from mechanisms of semantic priming. However, consistency theory would argue that extreme anchoring effects are likely an averaging phenomenon (Siegler, 1987). That is, anchoring effects are observed even if most participants reject the anchor and form an independent (and "unbiased") estimate, and a few (even a very few when responding to extreme anchors) participants are, for some reason,[4] influenced by the anchor value (Brown & Moore, 2003).

Finally, there have been anchoring studies that looked at the role of the decision-maker's level of expertise. These studies suggest that both experts and non-experts (e.g., real-estate agents vs. undergraduate business students in Northcraft and Neale, 1987; legal professionals vs. law students in Englich & Mussweiler, 2001; Englich et al., 2006) are equally influenced by anchor values.

In Englich et al., (2006)'s study, they found that despite the profession of the participants (i.e., legal professionals), final sentences for a defendant in a shoplifting case assimilated to the sentencing demands from a dice roll. However, it should be noted that like Tversky and Kahneman's (1974) rigged wheel of fortune, the dice in the study was fixed to land on a 3 in the low anchor condition, and a 9 in the high anchor condition. Consistency theory would suggest that even if low-knowledge participants have a wider range of plausible values than high-knowledge participants for a given target question (see Smith, Windschitl, & Bruchmann, 2013), high-knowledge participants would still be anchored if anchor values fell in the plausible range. Considering that the range of sentences was between 0 and 12 months, with a mean of 5.05 and a standard deviation of 3.18, this may have been the case.

Overall, the present study suggests that anchor values can elicit a number of different reactions from participants, and that automatic processes play only a minimal role. We conclude that people may not be adjusting towards an unbiased (and in many cases unknowable) value, but rather, adjusting away from the anchor. On this view, adjustments are *sufficient*, because people provide an estimate that is consistent with their prior *Greater* or *Less* assessment and that falls within the range of subjectively plausible values. In other words, when confronted with a plausible anchor value, people anchor and *sufficiently* adjust. On a more general level, the results of the present study therefore suggest that people have the ability to assess the potential relevance of new numerical information, and to incorporate this information along with existing knowledge when it is deemed relevant.

References

Albers, W. (2001). Prominence theory as a tool to model boundedly rational decisions. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality: The adaptive toolbox* (pp. 297–317). Cambridge, MA: MIT Press.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics, 118*, 73-105.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Bakan, P. (1960) Response-tendencies in attempts to generate random binary series. *The American Journal of Psychology, 73*, 127-131.

Bar-Hillel, M., Peer, E., & Acquisti, A. (in press) “Heads or tails?” A reachability bias in binary choice. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

Blankenship, K. L, Wegener, D. T., Petty, R. E., Detweiler-Bedell, B., & Macy, C. L. (2008). Elaboration and consequences of anchored estimates: An attitudinal perspective on numerical anchoring. *Journal of Experimental Social Psychology, 44*, 1465-1476.

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 41, pp. 321-360). New York, NY: Academic Press.

Brown, N. R., & Moore, S. E. (2003). *Mixed reactions: A reconceptualization of the post-comparison anchoring effect*. Unpublished manuscript, Department of Psychology, University of Alberta, Edmonton, Canada.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100*, 511-534.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes, 79*, 115-153.

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). Cambridge, England: Cambridge University Press.

Chaxel, A.S. (2014). The impact of procedural priming on selective accessibility on self-generated and experimenter-provided anchors. *Journal of Experimental Social Psychology, 50*, 45-51.

Englich, B., & Mussweiler, T. (2001). Sentencing under uncertainty: Anchoring effects in the courtroom. *Journal of Applied Social Psychology, 31*, 1535-1551.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin, 32*, 188-200.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*, 391-396.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making, 18*, 199-212.

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

Erlebacher, A., & Sekuler, R. (1971). Response frequency equalization: A bias model for psychophysics. *Perception and Psychophysics, 9*, 315-320.

Frederick, S. W., Kahneman, D., & Mochon, D. (2010). Elaborating a simpler theory of anchoring. *Journal of Consumer Psychology, 20*, 17-19.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*, 35-42.

Goodfellow, L. D. (1940). The human element in probability. *The Journal of General* Psychology, 23, 201-205.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). New York: Academic Press.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21*, 1161-1166.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Matthews, W.J., & Dylman, A.S. (2014). The language of magnitude comparison. *Journal of Experimental Psychology: General.* Advance online publication. doi: 10.1037/a0034143

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*, 136-164.

Mussweiler, T., & Strack, F. (2001a). Considering the impossible: Explaining the effects of implausible anchors. *Social Cognition, 19*, 145-160.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26*, 1142-1150.

Nickerson, R. S. (1981). Motivated retrieval from archival memory. In H. E. Howe, Jr. & J. H. Flowers (Eds.), *Nebraska symposium on motivation: Cognitive processes* (Vol. 28, pp. 73-119). Lincoln, NE: University of Nebraska Press.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*, 84-97.

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

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 19, pp. 123-205). New York, NY: Academic Press.

Russo, J. E., & Shoemaker, P. J. H. (1989). *Decision traps*. New York: Simon and Schuster.

Schneiderman W., & Manis, M. (1978). Response vs. prediction: An enduring question revisted in the domain of context effects. *Motivation and Emotion, 2*, 259-273.

Schwarz, N. (1994). Judgment in social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology*, 26, 123–162.

Schweickart, O., Tam, C., & Brown, N. R. (2014). *When "bad" is good: How evaluative judgments eliminate the standard anchoring effect*. Manuscript submitted for publication.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250-264.

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology, 99*, 917-932.

Smith, A. R., Windschitl, P. D., & Bruchmann, K. (2013). Knowledge matters: Anchoring effects are moderated by knowledge level. *European Journal of Social Psychology, 43*, 97-108.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology, 73*, 437-446.

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

Wegener, D. T., Petty, R. E., Blankenship, K. L., & Detweiler-Bedell, B. (2010a). Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Journal of Consumer Psychology, 20*, 5-16.

Wegener, D. T., Petty, R. E., Blankenship, K. L., & Detweiler-Bedell, B. (2010b). Elaboration and numerical anchoring: Breadth, depth, and the role of (non-) thoughtful processes in anchoring theories. *Journal of Consumer Psychology, 20*, 28-32.

Wegener, D. T., Petty, R. E., Detweiler-Bedell, B. T., & Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology, 37*, 62-69.

Wilson, T. D., Houston, C. E., Etling, K. M., Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125*, 387-402.

Footnotes

[1]According to Mussweiler & Strack (2001a), an adjustment process is only needed when the anchor falls outside the range of plausible answers. For example, if asked whether Mahatma Gandhi's age was greater or less than 271 when he died, people would arrive at an answer by first adjusting to a plausible range (e.g., to 100 years) before proceeding with the confirmatory hypothesis-testing process.

[2]The expectation is for participants to respond "Greater" for 3 out of the 28 filler items: (3/28)*100 = 10.7%.

[3]To determine the degree in which estimates for the filler items were influenced by the anchor value, we computed the anchor size effect using the *anchoring index* (AI) measure (Jacowitz & Kahneman, 1995). The AI is computed as follows:

AI = (Median [High anchor] – Median [Low anchor]) / (High anchor - Low anchor)

Over the 28 questions, the mean AI was .22 (see Appendix, Table 1). Values of AI typically range from 0 (no anchoring) to 1 (median estimates are equal to the anchor). Thus, for these filler items, participants moved almost a quarter of the way from an unanchored estimate toward the anchor.

[4]A potential reason for reliance on anchor values is that participants treat the anchor as a "hint" provided by the experimenter (Grice, 1975; Schwarz, 1994).

Appendix

Table 1

Summary Statistics for Filler Items

Questions	High anchor	Low anchor	Median anchor	Anchoring Index	Confidence
How old was Neil Armstrong when he landed on the moon?	45	28	34	0.18	2.32
How old is the Great Sphinx of Giza (in years)?	600,000	500	3,000	0.01	2.00
How many main islands are there in Hawaii?	16	3	5	0.15	2.91
How much water is in a standard water cooler jug (in litres)?	40	3	12	0.27	2.88
What is the world record for the most people crammed into an old style Volkswagen Beetle?	30	9	15	0.19	2.37
How many different prime ministers has Canada had?	65	13	30	0.23	2.22
How many football teams are there in the NFL?	50	20	32	0.18	2.46
How many homicides occurred in Edmonton in 2012?	240	15	44	0.12	2.64
How many gold medals did China win in the 2012 Summer Olympic Games?	61	8	21	0.28	2.43
What was the highest (hottest) recorded temperature for a day in Seattle, Washington (in degrees Celsius)?	50	35	43	0.33	2.50
How many countries are there in Africa?	54	8	30	0.22	2.43
How old was Ernest Hemingway when he died?	86	50	70	0.17	1.67
What is the average life expectancy of an elephant in the wild?	95	18	43	0.29	2.37
What is the average weight of a male German Shepherd (in pounds)?	150	42	83	0.28	2.75

How many games are played by each hockey team in the NHL during the regular season?	85	20	50	0.35	2.93
How many seats are there in the Canadian senate?	300	15	100	0.37	2.38
What is the fastest speed a cheetah can run (in km/hour)?	180	40	80	0.09	2.88
How long was the movie 'Forrest Gump' (in minutes)?	200	100	137	0.34	3.00
How many bones make up an adult human skeleton?	390	110	212	0.16	2.72
How many episodes of the television show 'Friends' were made?	390	59	180	0.33	2.49
How many calories are in a McDonald's 'Big Mac'?	1,400	440	600	0.31	3.11
What is the maximum seating capacity of a Boeing 747 (a jumbo jet)?	900	150	350	0.14	2.46
What is the distance between Edmonton and Toronto (in km)?	20,000	900	3,072	0.13	2.34
How many people died at the World Trade Center site in New York City on 9/11?	7,600	270	2,000	0.30	2.29
How many pages are there in the complete Harry Potter book series (UK version)?	27,000	1,800	3,500	0.10	2.79
What is the total student enrollment at the University of Calgary?	52,000	5,200	25,000	0.32	2.38
What is the base tuition fee at Harvard Law School for the 2013-2014 academic year (in US dollars)?	96,000	10,000	25,000	0.23	2.48
What is the annual salary of the President of the United States (in US dollars)?	5,600,000	120,000	450,000	0.11	2.24

Note. Anchoring Index (AI) = (Median[High] - Median[Low])/(High anchor - Low anchor); cf. Jacowitz and Kahneman (1995). Values of AI typicallyrange from 0 (no anchoring) to 1 (median estimates are equal to the anchor). Confidence ratings range from 1 (not confident at all) to 5 (very confident).