Category Rating Is Based on Prototypes and Not Instances: Evidence from Feedback-Dependent Context Effects

Alexander A. Petrov
Department of Psychology
Ohio State University

Context effects in category rating on a 7-point scale are shown to reverse direction depending on feedback. Context (skewed stimulus presentation probabilities) was manipulated between and feedback within subjects in two experiments with diverse stimulus sets. Prototype- and exemplar-based scaling models are contrasted on the basis of their diverging predictions in this paradigm. The critical factor is that prototype-based categories cannot increase their coverage on the continuum without decreasing their coverage on the opposite side. The range of qualitative behavioral patterns consistent with each model class is shown using computer simulations with two representative members: ANCHOR and an instance-based modification thereof. ANCHOR can exhibit context effects in either assimilative or compensatory direction depending on feedback. The instance-based model always exhibits assimilative context effects. The human data show a significant context-by-feedback interaction. The main context effect is assimilative in one data set and compensatory in the other. This pattern is consistent with ANCHOR but rules out the instance-based variant, which fails to account for the compensatory effect and the interaction. This suggests that human category rating is based on unitary representations.
Keywords: psychophysical scaling, prototypes, instance-based representations, categorization

Category rating is a widely used method of data collection in experimental psychology. Psychophysical scales (e.g., Stevens, 1957), similarity judgments, typicality judgments attitude questionnaires, and health self-reports—all these tasks involve classifying stimuli using an ordered set of relatively few categories such as 1 . . . 7 or *very dissimilar* . . . *very similar*. Such ratings are among our primary dependent measures. It is important to formulate a detailed, quantitative theory of how people produce these ratings (Petrov & Anderson, 2005). Moreover, the category rating task constrains theories of perception, categorization, and memory. As such, it is a fertile field for theoretical integration. The present study uses a psychophysical paradigm to differentiate between two prominent theories of categorization. It also makes a contribution to the psychophysical literature by demonstrating that external feedback can reverse the direction of context effects in category rating.

A classic controversy in the categorization literature (see, e.g., Ashby, 1992, for review) contrasts prototype and exemplar-based representations of categories. According to the prototype view (e.g., Posner & Keele, 1968; Smith & Minda, 1998; Rosch, 1975), each category is represented by a unitary description of the central tendency of its members. Novel instances are classified on the basis of their similarity to the prototypes of various categories. Alternatively, categories can be represented by storing the individual instances themselves (Kruschke, 1992; Medin & Shaffer, 1978; Nosofsky, 1986, 1992; Nosofsky & Zaki, 2002). According to this exemplar-based view, novel instances are classified on the basis of the weighted aggregate similarity to the known instances of various categories. Thus, prototype systems aggregate the information about known category members as they are stored in memory, whereas instance-based systems delay the aggregation until retrieval time. It has proven surprisingly difficult to discriminate between these competing views and a lively debate continues to this day (e.g., Busemeyer, Dewey, & Medin, 1984; Minda & Smith, 2001, 2002; Nosofsky, 2000; Nosofsky & Stanton, 2005; Olsson, Wennerholm, & Lyxzèn, 2004; Smith & Minda, 2000, 2002; Stanton, Nosofsky, & Zaki, 2002; Zaki, Nosofsky, Stanton, & Cohen, 2003). The difficulty stems in part from the great flexibility of instance-based models, which can mimic prototype models in certain parameter regimes (Nosofsky & Johansen, 2000; Nosofsky & Zaki, 2002; but see Myung, Pitt, & Navarro, 2007). Various hybrid schemes have also been proposed (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Huttenlocher, Hedges, & Vevea, 2000; Love, Medin, & Gureckis, 2004; Nosofsky, Palmeri, & McKinley, 1994).

A physical analogy clarifies the differences between the two classes of models. It is as if each category represen-

Alexander A. Petrov, Department of Psychology, Ohio State University. This article is accepted for publication in the *Journal of Experimental Psychology: Human Perception and Performance*.

The author thanks John R. Anderson, Rob Nosofsky, Roger Ratcliff, and James Todd. Correspondence concerning this article should be addressed to Alexander Petrov, Department of Psychology, 200B Lazenby Hall, Ohio State University, Columbus, Ohio 43210. E-mail: apetrov@alexpetrov.com

tation generates a gravitational field[1] across the stimulus space. When a new stimulus appears somewhere in this space, the categories compete to incorporate the new unit mass into their representation. By definition, all the mass of a prototype-based category is concentrated in a single point: the prototype. In this article, this will be referred to as the *unitary constraint* on category representation. The resulting gravitational field is radially symmetric and centered on that point. By contrast, instance-based representations are not subject to the unitary constraint. The mass of an instance-based category is dispersed across multiple points of unit mass each. The resulting gravitational field can have very irregular topography.

Prototype-based models have well-known limits of applicability (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1992, 1993; Medin & Shaffer, 1978; Nosofsky, 1992; Nosofsky et al., 1994; Nosofsky & Zaki, 2002). There are categories that instance-based models (and humans) can learn but prototype-based ones cannot. The complement is not true. Instance-based models can learn *any* stationary categorization task with enough practice with feedback (Ashby & Alfonso-Reese, 1995). Thus, they have great representational flexibility. It has been argued on the basis of this flexibility that instance-based accounts offer a general, all-encompassing model of human categorization (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 2002). Even in domains in which prototype-based theories can account for the data, the argument goes, instance-based theories should still be preferred on grounds of parsimony unless there is "clear and definitive evidence for the operation of prototype abstraction in people's category representations" (Nosofsky & Zaki, 2002, p. 939).

The present experiment provides one piece of such evidence. Our results establish a limit for the applicability of instance-based theories of categorization. Despite their spectacular success in other tasks, it appears that instance-based theories cannot provide an empirically adequate account of category rating without feedback.

The unitary constraint plays a key role in our argument. The two types of representations can be differentiated on the basis of one of its side effects. When a unitary category acquires a new member, the prototype moves toward the corresponding point in stimulus space. But this also moves it *away from* the opposite region of the space. Thus, there are locations where the gravitational field of the category *weakens* after the incorporation of a new member. Instance-based categories, on the other hand, can expand with impunity because they are free from the unitary constraint. Each new acquisition makes a positive contribution to the gravitational field across the entire stimulus space. It cannot happen that the field weakens anywhere after the incorporation of a new member.

Categorization models are often tested in binary classification tasks with multidimensional stimuli (e.g., Medin & Shaffer, 1978). Our task, by contrast, has seven response categories and unidimensional stimuli (distances between pairs of dots). If the instance-based theory is a truly universal account of categorization, it should apply to this task too. One advantage of our task is that it makes it easy to detect whether a category "loses ground" on one side when it "gains ground" on the opposite side. This is because the stimuli are linearly ordered and because most categories are flanked by other categories. Category rating also introduces the notion of systematic alignment—homomorphism—between stimuli and responses (Stevens, 1957). This homomorphism supports the formation of categories without any external feedback (Petrov & Anderson, 2005). This in turn allows for powerful experimental manipulations that are not possible in a two-choice categorization paradigm. The present experiments use such feedback manipulation.

Human categorization is sensitive to the frequencies of occurrence of various stimuli in the environment. This sensitivity gives rise to phenomena known as *context effects* in the psychophysical literature. The classification of a given stimulus depends not only on the stimulus itself but also on the distribution of other stimuli in a block of trials (e.g., Chase, Bugnacki, Braida, & Durlach, 1983; Marks, 1993; Parducci, 1974). One convenient way to manipulate the location of a category prototype is to increase the presentation frequency of stimuli near one end of the continuum or the other. The simplest experimental design is to compare the performance under a skewed stimulus distribution with the baseline performance under a uniform distribution. Two kinds of context effects are possible: assimilative and compensatory. For concreteness, suppose a group of observers is presented with predominantly long stimuli. By definition, if the stimuli tend to be systematically overestimated relative to baseline, there is assimilation—the responses are attracted toward the densely populated end of the scale. The rich get even richer. If the stimuli tend to be systematically underestimated instead, there is compensation. The response distribution is less skewed than the stimulus distribution.

Instance-based models predict assimilative context effects. Because all instances are stored separately and similarities always add, densely populated regions must have an attractive effect on any new instance. The predictions of prototype-based models are more complex because of the presence of two opposing mechanisms. One mechanism is assimilative—frequently used prototypes become more active, which gives them an advantage in the competition for new members. However, there is also a compensatory mechanism—when a prototype moves towards the densely populated region of the space, it "abandons" the less dense regions to competing prototypes. The overall context effect depends on the relative strengths of these opposing forces. Both assimilative and compensatory context effects are possible. A quantitative model is needed to weigh the relative strengths of the various factors.

In this article, we explore the qualitative patterns of behavior predicted by a representative example of each model class. The class of prototype-based models is represented by ANCHOR (Petrov & Anderson, 2000, 2005). It is a memory-

[1] Technically, this field is proportional to the product of the (estimated) base rate and the (estimated) probability density function of the category (cf. Ashby & Alfonso-Reese, 1995).

based scaling model that stores a single *anchor* per category. Each anchor is a weighted average of all stimuli labeled with the corresponding response. The class of instance-based models is represented by a variant of ANCHOR that stores a new exemplar on every trial. It is referred to as INST here and can be regarded as a modification of the well-known Generalized Context Model (Medin & Shaffer, 1978; Nosofsky, 1986, 1988). ANCHOR and INST are identical in all respects except their category representations. Thus, any differences in their observable behavior must follow from this representational difference.

Human categorization is a dynamic process that gives rise to sequential, practice, and other dynamic effects. ANCHOR has two incremental learning mechanisms that account for a comprehensive list of these effects in category rating and absolute identification (Petrov & Anderson, 2005). A competitive learning mechanism adjusts the location of each anchor along the magnitude continuum. A base-level learning mechanism updates the availability of the anchors. Under skewed stimulus distributions, these two mechanisms tend to push the average response level in opposite directions. Base-level learning is assimilative whereas competitive learning is compensatory.

The compensatory tendency is a direct consequence of the unitary constraint on ANCHOR representations, as detailed in the next section. Moreover, this compensatory tendency disappears in the presence of external feedback. Thus, ANCHOR predicts that the context effects can have opposite directions with and without feedback. This prediction was confirmed in two experiments that manipulated the stimulus frequencies within subjects and feedback between subjects (Petrov & Anderson, 2005). The experiments reported here have a complementary design: The frequencies are manipulated between and feedback within subjects, in order to trace the feedback-induced dynamics of the context effects.

INST also has two incremental learning mechanisms. Its base-level learning is a special case of that in ANCHOR—instances decay with time. The dynamic adjustment of anchor locations, however, is replaced in INST by the memorization of separate exemplars. Once committed to memory, the locations of these exemplars never change. Conceptual analysis and computer simulations demonstrate that this mechanism produces assimilative context effects. Thus, INST has no mechanism that can counterbalance the assimilative influence of frequent stimuli. INST can never produce compensatory context effects. It can only assimilate, and this assimilation is exacerbated in the absence of feedback.

The next section presents the ANCHOR theoretical framework and explains the key predictions in qualitative terms. ANCHOR can produce some behavioral patterns that INST cannot, even though the representational scheme of ANCHOR is more constrained than that of INST.[2] The results of Experiment 1 are compatible with ANCHOR but not INST. Although context effects are assimilative overall, there is a significant interaction with feedback. Lack of feedback makes the context effects less assimilative and can even reverse their direction under certain conditions. A second experiment replicates this context-by-feedback interaction with different stimuli and under tighter controls. The main effect of context is compensatory in the second data set, which is even more problematic for INST. On the basis of these empirical findings and a theoretical analysis of the task demands, we conclude that category rating is based on prototypes and not instances. Finally, the broader implications are discussed and the models are compared to other influential models from the literature.

Theoretical Framework

The two models presented in this article are based on the ANCHOR theoretical framework (Petrov & Anderson, 2000, 2005). It integrates three broad theories: memory-based categorization (e.g., Nosofsky, 1986; Rosch, 1975), Thurstonian psychophysics (Thurstone, 1927; Torgerson, 1958), and the theory of memory incorporated in the ACT–R architecture (Anderson & Lebiere, 1998; Anderson & Milson, 1989). The ANCHOR theory is described in detail elsewhere (Petrov & Anderson, 2005). The present analysis is predicated on the four principles summarized below.

Main Principles

Internal Magnitude Continuum. It is assumed that some sensory process maps the intensity of the physical stimulus onto an internal *magnitude*. It is this internalized quantity that can be committed to memory and compared against other magnitudes.

Content-Addressable Memory. It is possible to establish associations between a magnitude and the label of a response category. The *anchors* in ANCHOR and the *instances* in INST are such associations. They substantiate the mapping between magnitudes (and hence the stimuli represented by them) and responses. Given a new target magnitude for classification, the memory fills in the corresponding response label. This completion process is stochastic and depends on two factors: (a) the similarity of each memory element to the target and (b) the base-level activation of each memory element.

Explicit Correction Strategies. People are aware that their "first guess" is not always reliable and adopt explicit correction strategies. The product of memory retrieval is not always reliable because the memory system is noisy and biased in favor of frequent and/or recent items. The role of memory is to provide a reference point in the vicinity of the target, thereby converting the global scaling problem into a local comparison problem. The final response can increment or decrement the retrieved category label. An introspective report of a trial might go like this: "This looks like a 5. No, it's too short for a 5; I'll give it a 4." It is well known that people rely on such anchor-plus-adjustment heuristics in uncertain situations (Tversky & Kahneman, 1974). However, the profound impact that even occasional corrections can have on

[2] This illustrates that *representational* flexibility does not necessarily entail *behavioral* flexibility (R. Nosofsky, personal communication, April 28, 2010).

the dynamical stability of a memory-based system has only recently been appreciated (Petrov & Anderson, 2005). The correction mechanism is the major ANCHOR innovation relative to standard memory-based theories. It is what allows ANCHOR to unfold a rating scale without any external feedback. The correction strategy is not elaborated here because it does not inform our present focus on prototypes versus instances. We should keep in mind, however, that without corrections both models fall pray to runaway winner-takes-all dynamics in the absence of feedback.

Obligatory Learning. The state of the system is incrementally updated at the end of each trial. One learning mechanism updates the base-level activations of the memory elements and thus indirectly tracks the base rates of the corresponding responses. A second learning mechanism tracks the probability density of each category across the magnitude continuum. The latter mechanism is sensitive to the unitary constraint on category representation. The two models presented below differ mainly in the way they learn the probability densities of categories.

Two Models: ANCHOR and INST

The ANCHOR model instantiates these theoretical principles in a concrete, implemented system.[3] See Appendix A for a list of equations and Petrov and Anderson (2005) for a comprehensive treatment.

Briefly, ANCHOR uses prototype representations: There is one anchor per response category. The instructions in our experiments call for a 7-point response scale. Thus, under the unitary constraint, there are 7 anchors in the model. When a new stimulus is presented, its magnitude serves as a memory cue and the anchors compete to match it. The winning anchor represents ANCHOR's "first guess" and provides a reference point for the correction mechanism. The latter may increment or decrement the anchor response depending on the discrepancy between the target and anchor magnitudes. This introduces a relative-judgment component to ANCHOR (cf. Stewart, Brown, & Chater, 2005) and accounts for the growing evidence of negative generalization between highly dissimilar stimuli (Stewart, Brown, & Chater, 2002; Jones, Love, & Maddox, 2006).

The adjustments are systematic but conservative. Their systematicity promotes the homomorphism between stimuli and responses even without feedback. Because positive discrepancies trigger positive corrections, large stimuli tend to map to high responses in the long run. Because correction thresholds are conservatively high, it matters which anchor is used as reference—the response is often assimilated toward it. This is how factors such as frequency and recency exert their influence—anchor selection is sensitive to them and subsequent correction does not compensate for them fully. This insufficiency of adjustment is a common theme in the diverse literature on anchoring effects (e.g., Hastie & Dawes, 2001; Wilson, Houston, Etling, & Brekke, 1996).

After committing to a response, ANCHOR updates the corresponding anchor to reflect the incorporation of a new stimulus into this category. If there is feedback, this is it; otherwise the system's own response designates the anchor for update. The base-level activation of this anchor increases, whereas the activations of all other anchors decay. A separate learning rule adjusts the location of the anchor on the magnitude continuum. The basic idea is very simple—the anchor moves a little towards the new exemplar being incorporated into the category. Only the anchor for the current response is updated; all other anchors stay put. This *competitive learning rule* sets the location of each anchor to the (exponentially weighted) running average of the magnitudes of all stimuli classified under the associated response category.

INST is our representative member of the class of instance-based models. It is identical with ANCHOR in all respects except that it does not obey the unitary constraint on category representations. It stores a separate exemplar on each trial. Consequently, there is no need for competitive learning. It is replaced by simple memorization of individual exemplars. Each exemplar has a base-level activation that decays with time. The exemplars compete to match the target on each trial. This competition is governed by the same equations as in ANCHOR and is mathematically equivalent to that in the Generalized Context Model (GCM, Nosofsky, 1986, see Appendix A). The competition involves hundreds of memory elements in INST as opposed to seven elements in ANCHOR.

Once an instance is retrieved from memory, it is subject to the same correction strategy as in ANCHOR. This is a major difference from GCM. It is what allows INST to perform not only absolute identification with feedback (Nosofsky, 1997) but also category rating without feedback (Petrov & Anderson, 2005). The decaying activation of INST's exemplars is a second difference, although there are GCM variants that involve exemplar strengths and response biases (Nosofsky, 1988, 1991; Nosofsky & Palmeri, 1997).

Context Effects: Push and Pull

Both ANCHOR and INST are adaptive dynamic systems. Obligatory learning is one of their foundational principles. As new stimuli are presented and classified under various response categories, the internal representations of these categories change in systematic ways. This in turn affects the classification of future stimuli. Because stimulus frequency is a potent determinant of this dynamics, both models predict context effects on a principled basis.

INST always predicts assimilative context effects under skewed stimulus distributions. To illustrate, suppose long stimuli are more frequent than short ones. As each exemplar is stored individually, the memory pool contains many instances labeled 5, 6, or 7 and few instances labeled 1, 2, or 3. Now, suppose a target in the middle of the range is presented. Its correct classification is 4, but it is also quite similar to instances labeled 3 and 5. By sheer force of numbers as all instances compete to match the new target, the probability to retrieve an instance labeled 5 is greater than the probability to retrieve an instance labeled 3. Consequently,

[3] Open-source Matlab implementation of both ANCHOR and INST is available at `http://alexpetrov.com/proj/anchor/`

the new stimulus is misclassified as 5 more often than it is misclassified as 3. The responses tend to shift toward the densely populated regions on the scale—an assimilative context effect. None of the other INST mechanisms can reverse the direction of this effect. The activations of all old instances decay at the same rate. The correction mechanism operates on the single instance retrieved from memory on a given trial. Although some retrieval mismatches are corrected, many go undetected. Thus, the assimilative tendency induced by the uneven gravitational fields persists, though attenuated. The important contribution of the correction mechanism is to prevent this tendency from running out of control in the absence of feedback (Petrov & Anderson, 2005).

ANCHOR, on the other hand, is consistent with both assimilative and compensatory context effects. This is because the two learning mechanisms in ANCHOR push in opposite directions. The overall context effect depends on the parameter-dependent relative strengths and interactions of these opposing forces.

The base-level learning mechanism in ANCHOR has an assimilative influence. To continue the above example, if more stimuli have been labeled 5 than 3, the activation of anchor 5 will be stronger than that of anchor 3. This is how the base rates of the categories are represented in ANCHOR (and ACT–R more generally). Active anchors are more likely to be retrieved from memory. Categories with many members thus exert stronger gravitational fields than categories with few members, everything else being equal.

But not everything else is equal in ANCHOR because of the compensatory influence of the competitive learning mechanism. Figure 1 illustrates the qualitative situation. A configuration with uniformly located, equally active anchors serves as baseline (top). The cone around each anchor depicts its gravitational field. The rectangular areas delineate the resulting partitions of the magnitude continuum. (The stochasticity of the anchor selection mechanism is ignored for simplicity. See Appendix A for details.) Now, suppose a long stimulus is classified under the category represented by the middle anchor in Figure 1. After this stimulus is averaged in, the anchor location shifts to the right. The gravitational field of this anchor also shifts *without growing in size*. As a result, a region formerly labeled 2 is now labeled 1, and a region formerly labeled 3 is now labeled 2. The net result is a systematic decrement of the overt responses.

It is convenient to formulate a descriptive rule of thumb to refer to this effect. According to this *inversion rule*, whenever the location of any anchor increases, responses decrease on average, and vice versa. The fundamental reason for this inversion is the unitary constraint on category representations. Because there is only one anchor per category, when it moves towards some region of the stimulus space, it is forced to leave the opposite region behind. In the latter region (e.g., to the left of the black anchor in Figure 1b), the gravitational field *weakens* after the incorporation of a new member.

The competitive learning tracks the probability density of the magnitude distribution. In skewed stimulus contexts, all anchor locations shift toward the densely populated end of the continuum. For example, suppose there is a preponderance of long stimuli. As they are being averaged in, the anchor locations shift toward longer values. By the inversion rule, the overt responses tend to shift in the opposite direction—a compensatory context effect.

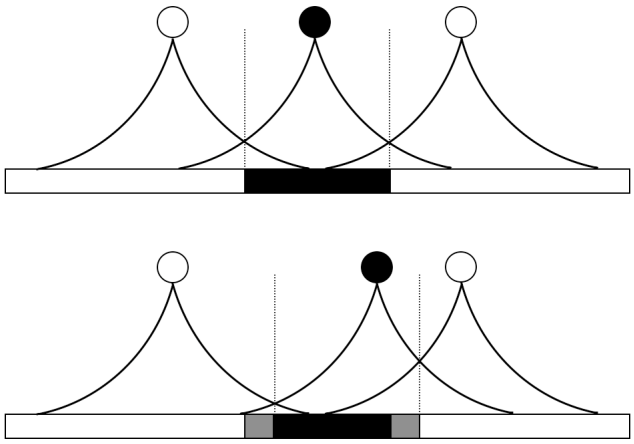

Figure 1. Illustration of the compensatory influence of the competitive learning mechanism in the ANCHOR model. The cone around each anchor (circle) depicts its "gravitational field" on the target magnitude continuum (horizontal axis). Assume the anchors are labeled 1, 2, 3 from left to right. Top: Baseline configuration with three equally spaced anchors. Bottom: Anchor 2 has shifted to the right after averaging in a long stimulus. As a result, a (grey) region formerly labeled 2 is now labeled 1, and a region formerly labeled 3 is now labeled 2. Thus, whenever an anchor shifts to the right, the responses shift to the left and vice versa.

This compensation counteracts the assimilatory tendency of the base-level learning. The opposition between the two learning mechanisms dampens any big fluctuations in either direction and aids the correction mechanism in preserving the stability of the system.

The Role of Feedback

ANCHOR is consistent with compensatory context effects where as INST is not. This difference can be used to dissociate the two models empirically. However, both models are consistent with assimilative context effects. Assimilative patterns are therefore ambiguous and must be dissociated on the basis of some other variable. The presence or absence of feedback is one such variable.

INST is relatively insensitive to feedback. None of its mechanisms is changed by feedback in any fundamental way.[4] ANCHOR, on the other hand, is affected by feedback. The differential predictions of the two models hinge again on the competitive learning mechanism (and hence on the unitary constraint enforced by it).

External feedback effectively switches competitive learning off. This is because when veridical feedback is available,

[4] This insensitivity depends on the correction mechanism. Instance-based models without corrections are generally extremely sensitive to feedback.

the model always updates the anchor representing the correct classification of the stimulus. This fixes the anchor locations to the internal images of the corresponding stimuli regardless of their presentation frequencies. The only remaining variability comes from perceptual fluctuations, which have no systematic effect on the overt responses.

ANCHOR's activation learning, on the other hand, exerts its assimilative tendency regardless of feedback. Skewed stimulus distributions always lead to skewed activation profiles, which in turn affect the retrieval probabilities and hence the responses. The assimilative tendency in INST persists regardless of feedback for similar reasons.

In summary, ANCHOR makes the following predictions. With feedback, context effects must be assimilative because the compensatory tendency of the competitive learning mechanism is switched off and all that remains is the assimilative tendency of the activation learning mechanism. Without feedback, the direction of the context effects is parameter-dependent. When feedback alternates across blocks, it interacts with context so that the context effect must be less assimilative with feedback than without.

Experiment 1

The present experiment is designed to differentiate the two classes of categorization models on the basis of these diverging predictions. A set of seven line lengths is used throughout. Three contexts—uniform, low, and high—are defined by different frequency distributions. Context is manipulated between subjects and feedback within subjects.

Method

Stimuli and Apparatus. The stimuli were pairs of white dots presented against a uniformly black background on a 17-inch AppleVision monitor. The viewing distance was approximately 600 mm. The independent variable was the distance between the centers of the two dots. The stimulus set consisted of seven dot pairs with the following distances: 420, 460, 500, . . . , 660 pixels (420 pixels $\approx$ 134 mm $\approx$ 13 degrees of visual angle [dva]; 660 pixels $\approx$ 211 mm $\approx$ 20 dva). The full width of the monitor was 1000 pixels (320 mm, 32 dva). The imaginary segment formed by the dots was always horizontal and was randomized with respect to its absolute horizontal and vertical position on the screen. The stimulus set for each participant was generated and randomized separately. Each dot was roughly circular in shape with a diameter of 16 pixels (5 mm, 0.5 dva).

Observers. Fifty-five undergraduate students at Carnegie Mellon University participated in the experiment to satisfy a course requirement.

Design. Each stimulus sequence consisted of 17 blocks of 28 trials each. The presentation frequencies in each block varied depending on context as follows: Uniform (U) blocks contained 4 presentations of each stimulus. Low (L, positively skewed) blocks contained 7, 6, 5, . . . , 1 presentations of Stimuli 1, 2, 3, . . . , 7, respectively. High (H) blocks were skewed in the opposite (negative) direction. The order of presentation was randomized within each block.

There were 5 experimental groups: Group U1 presented 17 uniform blocks. Groups L1 and L2 presented 1 uniform block followed by 16 low blocks. Groups H1 and H2 presented 1 uniform block followed by 16 high blocks. The first block was always uniform and always with feedback. The feedback-first Groups U1, L1, and H1 gave veridical feedback on blocks 2–5, 10–13 and no feedback on blocks 6–9, 14–17. The no-feedback-first Groups L2 and H2 gave no feedback on blocks 2–5, 10–13 and veridical feedback on blocks 6–9, 14–17.

Procedure. The participants were instructed that there were seven stimuli and seven responses and that their task was to identify each stimulus with a number from *1* to *7*. The instructions stated that the 7 stimuli would be shown "multiple times in random order." Nothing was mentioned about presentation frequencies.

Each trial began with a 500-ms alert sound followed by the presentation of a dot pair on the monitor. The dots remained visible until the participant entered their response on the keyboard. Then the dots were replaced by a big white digit indicating the correct identification in feedback blocks or by an uninformative "X" in no-feedback blocks. The character stayed for 1100 ms. Then the screen was cleared and the next trial began. Its 500-ms alert sound served as the inter-trial interval. Each session lasted about 40 minutes and consisted of 476 trials divided into 8 periods with short breaks after trials 70, 140, 196, 252, 308, 364, and 420.

Dependent Variable. Petrov and Anderson (2005) introduced a general method for tracking the *average response levels* (ARLs). Context effects manifest themselves as ARL deflections under different conditions. ARL is our primary dependent variable throughout this article. It is defined as the area under the Stevens function divided by the stimulus range (Petrov & Anderson, 2005). The Stevens function $\bar{R} = F(S)$ gives the expected category rating of each stimulus (Stevens, 1957). It is a good summary of the response policy across the stimulus range. The ARL is designed to condense this summary to a single number that can be estimated from behavioral data collected with arbitrary presentation frequencies. ARL is calculated in two steps. First, the coefficients of the Stevens power function $\bar{R} = R_0 + aS^n$ are estimated from the stimulus-response pairs. As the exponent n for physical length is virtually 1.0 (Stevens & Galanter, 1957; Petrov & Anderson, 2005), simple linear regression suffices for the present analyses. The second step of the ARL calculation is also very simple for linear functions. The ARL equals the predicted response to the stimulus in the middle of the range:

$$ARL = R_0 + a(S_{min} + S_{max})/2 \quad (1)$$

Our middle stimulus is 540 pixels long. Thus, $ARL = R_0 + 540a$. The coefficients R_0 and a are estimated by linear regression. The stimulus-response sequence is segmented into 9 nonoverlapping periods of 56 trials each.[5] A

[5] The first period, which is always uniform, is only 28 trials long.

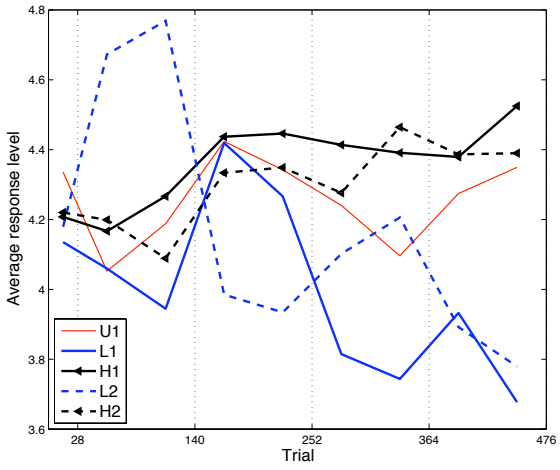

Figure 2. Average response levels for the line-length ratings in Experiment 1. Each line plots the mean of the 11 individual profiles in the corresponding group. Group U1: uniform context. Groups L1 and L2: low context (frequent short stimuli; upward shifts indicate compensation). Groups H1 and H2: high context (frequent long stimuli; upward shifts indicate assimilation). The vertical grid lines mark transitions from feedback to no-feedback blocks or vice versa. Groups U1, L1, H1: feedback on Trials 1–140, 253–364. Groups L2 and H2: feedback on Trials 1–28, 141–252, 365–476.

separate regression line is fitted and ARL calculated for each period. This converts the raw data to a profile of 9 average response levels per participant.

Results and Discussion

Figure 2 plots the mean ARL profiles for the five experimental groups. The data show clear-cut context effects modulated by feedback in agreement with the ANCHOR model. The overall direction is assimilatory—the average response levels in High context (thick lines with triangular markers) tend to exceed those in Low context (thick lines with no markers). The ARL in Group U1 (Uniform context) is between the ARLs in Groups H1 and L1.

Importantly, the overall assimilatory tendency is attenuated or even reversed during the no-feedback blocks. This feedback modulation is most evident for Groups L1 and L2. As the presentation frequencies are the same in both groups, the zig-zag pattern in their ARL profiles is driven entirely by the feedback manipulation. In particular, the highest ARL in the whole data set occurs during the initial no-feedback period in Group L2 (dashed line, Trials 29–140). Given the preponderance of short stimuli in this group, a high ARL indicates a compensatory context effect. When feedback is introduced in Group L2 on Trial 141, the average response level drops by 0.8 category units and the context effect becomes assimilatory. In Group L1, which is released from feedback at the same time, the ARL increases by 0.4 units and the context effect becomes compensatory. This context-by-feedback interaction is predicted by ANCHOR and inconsistent with INST.

The statistical significance of these findings is confirmed by a mixed-design ANOVA. For simplicity, Group U1 and the initial "warm-up" point on each ARL profile are not included. Context (H and L) and Order (feedback-first and no-feedback-first) enter as between-subject factors; Feedback (0–1) and Period (1–4) enter as within-subject factors. The temporal order of observations is ignored. The significant main effect of Context ($F(1,40) = 13.1$, $p < .001$, $\eta_p^2 = .25$) validates the overall assimilatory context effect in Figure 2. The significant Context by Feedback interaction ($F(1,40) = 17.7$, $p < .001$, $\eta_p^2 = .31$) validates the feedback modulation of the context effect. Some higher-order interactions are significant too (e.g., Context by Feedback by Order, $F(1,40) = 13.3$, $p < .001$, $\eta_p^2 = .25$). The main effect of Order is not significant ($F(1,40) < 1$).

There is a significant main effect of the Feedback factor ($F(1,40) = 20.9$, $p < .001$, $\eta_p^2 = .34$). Averaged across all contexts, the ARLs without feedback tend to exceed those with feedback. This tendency is evident, for example, in the control group U1 in Figure 2 (thin line). We attribute it to an idiosyncratic feature of our experimental setting. In general, the participants tend to overestimate the length of our stimuli—the baseline ARL in our data set (≈ 4.2, cf. trials 1–28) overshoots the halfway point (4.0) on the scale. Similar overshoot was observed in earlier experiments with these stimuli (Petrov & Anderson, 2005). Explicit feedback tends to bring the ratings closer to ideal performance. That is, there is downward pressure on the ARLs during the feedback blocks. Without feedback, the pressure is released and the ARLs tend to increase. This tendency amplifies the context-by-feedback interaction in low contexts and obscures it in high contexts. This explains why the zig-zag pattern is much more pronounced for Groups L1 and L2 than for Groups H1 and H2. The interpretability of the data is not jeopardized, however, because the theoretically relevant interaction is strong enough to overcome this tendency. In particular, the ARL in Group H2 tends to be *lower* without feedback than with feedback. Once again, absence of feedback promotes the compensatory context effect consistent with ANCHOR.

Assuming the idiosyncratic tendency discussed above does not interact with Context, we can cancel it out by subtracting two ARL profiles obtained in complementary contexts. Thus we define a new dependent variable *Assimilation*:

$$Assimilation = ARL(H) - ARL(L) \qquad (2)$$

Positive and negative values indicate assimilative and compensatory context effects, respectively. The ARL profiles in Figure 2 (ignoring the uniform group) combine into the Assimilation profiles in Figure 3. The overall assimilative context effect is evident from the positive sign of nearly all Assimilation values. The modulatory effect of feedback is manifested in the interlocking zig-zag pattern. Feedback blocks (marked by symbols) consistently show more assimilation than the no-feedback blocks (no symbols).

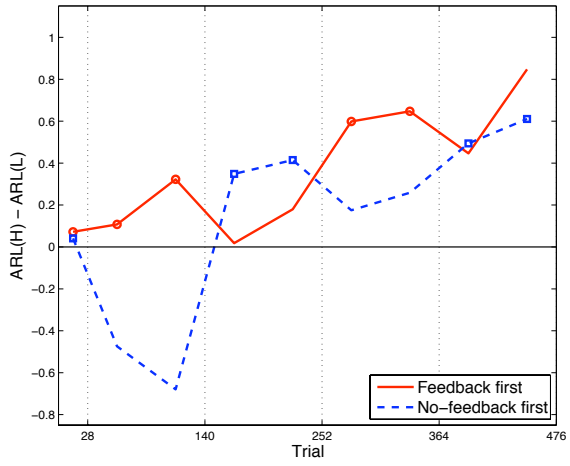

Figure 3. Context effects in the line-length rating task from Figure 2. The average response level in low context [ARL(L)] is subtracted from that in high context [ARL(H)] to measure assimilation. The vertical grid lines mark transitions from feedback to no-feedback blocks or vice versa. The assimilative effect is stronger in the feedback blocks (circles and squares) than in the no-feedback blocks (plain lines). This interaction is predicted by the prototype-based model but inconsistent with the instance-based model.

Model Fits

To assess how well INST and ANCHOR can account for the ARL profiles, each model was fitted to the data in Figure 2 by minimizing the root mean square error (RMSE) between predicted and observed ARLs (see Appendix B for details). ANCHOR fits better ($RMSE = 0.154$) than INST ($RMSE = 0.173$) with the same number of free parameters (Table B1 in Appendix B). Figure 4 plots the best-fitting ARL profiles and the corresponding Assimilation profiles.

Panels a and c in Figure 4 indicate that both models account for the overall assimilative context effect in Figure 2. Also, both models account for the tendency of the ARL to overshoot the midpoint of the response scale. The models produce this tendency by adjusting their correction thresholds so that upward corrections are more frequent than downward corrections (see Appendix A for details). This introduces a systematic upward drift of the average response levels (Petrov & Anderson, 2005). The drift is stronger in the absence of feedback and thus both models account for the main effect of the Feedback factor as well. The ARL drifts upward when there is no feedback, regardless of context. The zig-zag pattern in Uniform context (the line labeled U1 in Figures 2 and 4) is entirely driven by this effect.

Critically, ANCHOR accounts for the Context by Feedback interaction whereas INST does not. The difference is apparent in High contexts. When long stimuli are more frequent than short ones, INST predicts *higher* ARLs in the no-feedback blocks relative to the feedback blocks (thick lines with triangular markers in Figure 4, panel c). The pattern in the empirical data is exactly opposite—the ARLs are *lower* without feedback, particularly during the formative early period (Figure 2, Trials 29–140). The ANCHOR predictions in High contexts are in agreement with the data. In Low contexts, both models exhibit compensatory context effects during the no-feedback blocks. These effects, however, can be attributed to the main effect of Feedback rather than the *interaction* between Context and Feedback. It so happens that in Low contexts both the main effect and the interaction deflect the ARLs upwards.

The interaction effects are easier to interpret if we subtract out the Context factor (Equation 2). Panels b and d in Figure 4 plot the Assimilation profiles predicted by the two models. The INST profile never goes negative and shows little effect of the feedback manipulation. By contrast, ANCHOR captures the crucial compensatory tendency in the early no-feedback period (dashed line in panel b). It also reproduces the interaction pattern in Figure 3, at least qualitatively. The quantitative fit is not perfect because the parameters were optimized with respect to ARL rather than Assimilation.

Qualitative Patterns Consistent with Each Model

As we have just seen, ANCHOR fits the average response level profiles somewhat better than INST, but not dramatically better ($RMSE = 0.154$ vs. 0.173). Neither fit is really spectacular. Is this sufficient basis to prefer one model over the other? In addition to goodness of fit, it is also important to consider the range of qualitative patterns consistent with each model (Roberts & Pashler, 2000). To that end, the models were run with a range of parameter values in a simulation experiment that mirrors Experiment 1. See Appendix C for details on the simulation method.

Figure 5 plots the ARL profiles predicted by ANCHOR and INST for a range of values of the so-called history weight parameter H. The pattern of context effects depended mostly on this parameter. This is why we explored it systematically. Reasonable variations of the other parameters did not introduce any qualitatively new patterns. Recall that memory retrieval in the models is sensitive to two factors: (a) the similarity of each memory element to the target and (b) the base-level activation of each memory element. The history weight controls the strength of the second factor relative to the first (see Equation 12 in Appendix A). Memory retrieval is driven by similarity when H is low, and by the frequency and recency of past responses when H is high.

INST Predicts Assimilation

The instance-based model predicts assimilative context effects for all values of H. Panel e in Figure 5 plots the ARL profiles for $H = 0$; Panel f plots them for $H = .070$, which is a very high value for this parameter. The average response level shifts upward in high contexts and downward in low contexts. This assimilative tendency is a parameter-free prediction of the instance-based model. It is discernible for any parameter setting that generates less than perfect accuracy.

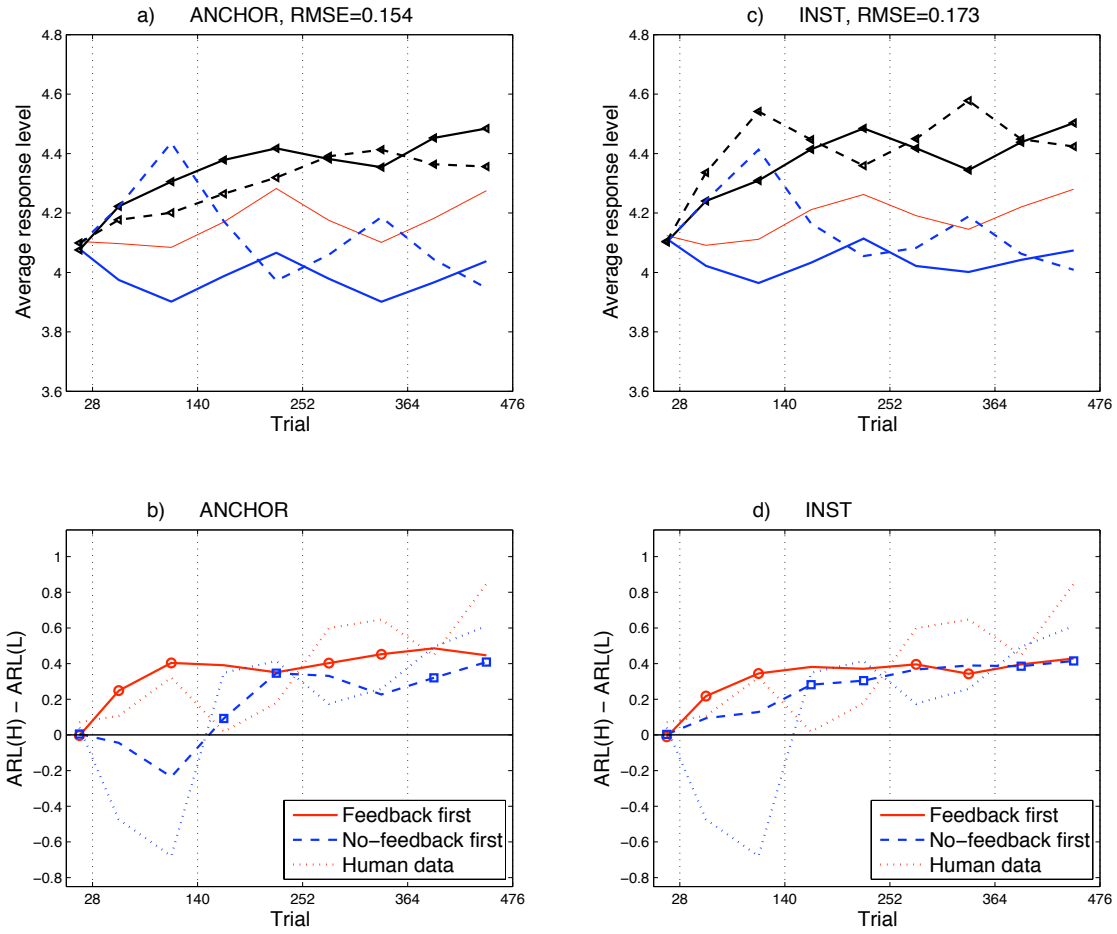

Figure 4. Fits of the prototype-based (ANCHOR) and instance-based (INST) models. Panels a and c: Average response levels (ARLs). Each profile is based on 250 simulated runs. Line styles as in Figure 2 (thick lines with triangular markers = High context, thick lines without markers = Low context, thin lines = Uniform control). RMSE = root mean square error. INST predicts feedback effects in the wrong direction in High context. Panels b and d: Assimilation profiles corresponding to panels a and c. The vertical grid lines mark transitions from feedback to no-feedback blocks or vice versa. Line styles as in Figure 3. See text for details.

The assimilative tendency is relatively insensitive to feedback. The vertical grid lines in Figure 5 mark transitions from feedback to no-feedback blocks or vice versa. The schedule is exactly the same as in Experiment 1. The ARL profiles are slightly different between feedback-first (solid lines) and no-feedback-first (dashed lines) runs, but the sign of the context effect is assimilative in all cases. This replicates the relative insensitivity to feedback in INST's fits to the empirical data in Figure 4, panel d.

Figure 6 shows the Assimilation profiles corresponding to the ARL profiles in Figure 5. The assimilative context effect is evident from the positive values (Equation 2). INST's assimilative tendency tends to increase slightly in the no-feedback blocks (e.g., trials 140–252 on panel f, solid line). This is opposite to the direction of the interaction effect in the empirical data (Figure 3, trials 140–252, solid line).

ANCHOR Is More Flexible

The prototype-based model, on the other hand, can produce three qualitatively different patterns of context effects illustrated in Panels a–d of Figure 5. The assimilative influence of the base-level learning counteracts the compensatory influence of the competitive learning. When H is high, the assimilative influence dominates and the average response levels resemble those of the INST model as illustrated in Panels d and f in Figure 5. ANCHOR can thus mimic INST. When H is low, the compensatory influence dominates, but only during the no-feedback blocks. This produces a strong context-by-feedback interaction. Panel a illustrates it for the extreme case of $H = 0$, which shows the effects of competi-

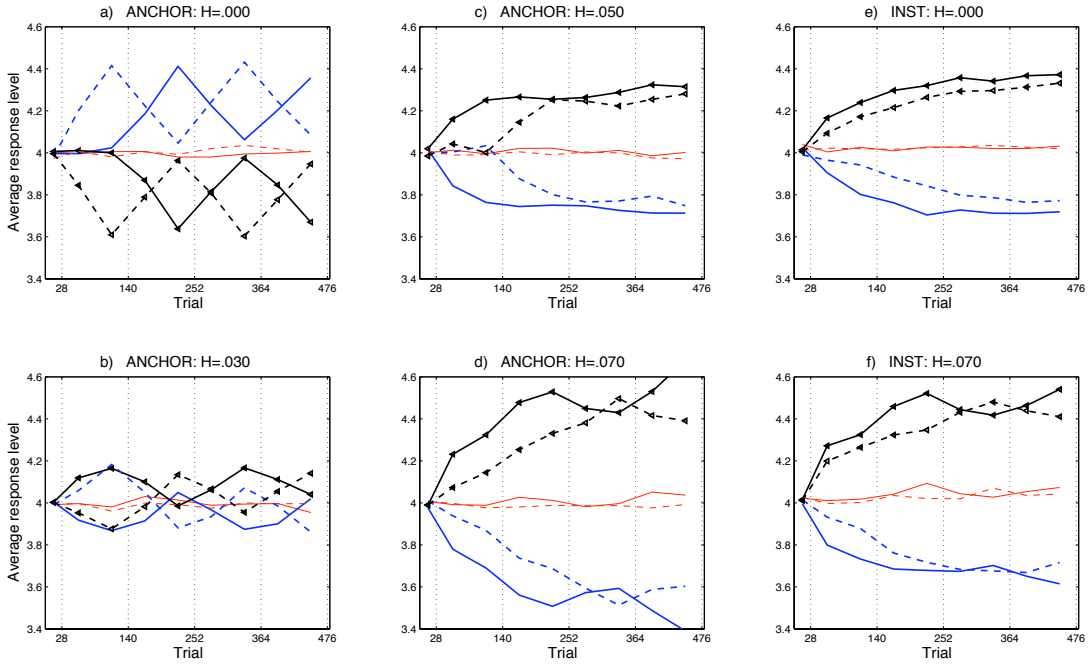

Figure 5. Average response levels (ARLs) predicted by the prototype-based (ANCHOR) and instance-based (INST) models. Each panel reports a batch of runs with history weight H indicated above the panel. The 6 ARL profiles correspond to the 3x2 factorial combinations of context and feedback. Thin lines: uniform context. Thick lines with triangular markers: high context (frequent long stimuli; upward shifts in ARL indicate assimilation). Thick lines with no markers: low context (frequent short stimuli; upward shifts indicate compensation). Solid lines: feedback-first sequences. Dashed lines: no-feedback-first sequences. The vertical grid lines mark transitions from feedback to no-feedback blocks or vice versa.

tive learning in pure form. Consider the two feedback-first groups (thick solid lines). Trials 1–140 and 253–364 are with feedback; Trials 141–252 and 365–476 are without. The flat segment up to Trial 140 shows that when both learning mechanisms are silenced, there are no context effects.[6] When feedback is discontinued, the ARL shifts downward in high context and upward in low context. This compensatory effect is driven by the inversion rule in competitive learning. When feedback is reintroduced at Trial 253, it gradually resets the anchors to their home positions and the ARLs converge back to the baseline. The no-feedback-first groups show a complementary pattern (thick dashed lines).

The corresponding Assimilation profiles (Figure 6) further illustrate these points. In particular, panel b demonstrates that ANCHOR can produce the qualitative pattern in the human data (Figure 2). When the history weight is such that base-level learning is allowed to operate but is weaker than competitive learning, ANCHOR predicts compensation during the initial no-feedback segment (dashed line, Trials 29–140), assimilation during the initial feedback segment (solid line), and continual context-by-feedback interaction during the subsequent segments.

Discussion

These qualitative considerations are our main basis for preferring the anchor-based model. Its superior quantitative fit reinforces the same conclusion. However, there is an alternative interpretation in terms of response bias (Parducci, 1974). A compensatory context effect can occur when the observers try to use all scale values equally often. The observed interaction effect can occur when the response bias is weaker with feedback than without.

The tendency of all ARLs in Experiment 1 to creep upward in the no-feedback blocks also complicates the interpretation of the data. We attributed it to an idiosyncratic feature of our stimuli. They were equally spaced but did not form a sequence with zero intercept—Stimulus 1 was not half as long as Stimulus 2, etc. Even though the instructions explicitly asked for an interval rather than ratio scale, some participants may have tried to preserve stimulus ratios (Stevens, 1957). If this strategy tended to overestimate the short stimuli more than the long ones, it could give rise to an upward bias in the average response levels. Our analyses assume that this bias does not interact with context and thus

[6] The flat segment also validates that the estimated ARLs remain unbiased even when the raw data are collected in skewed contexts.

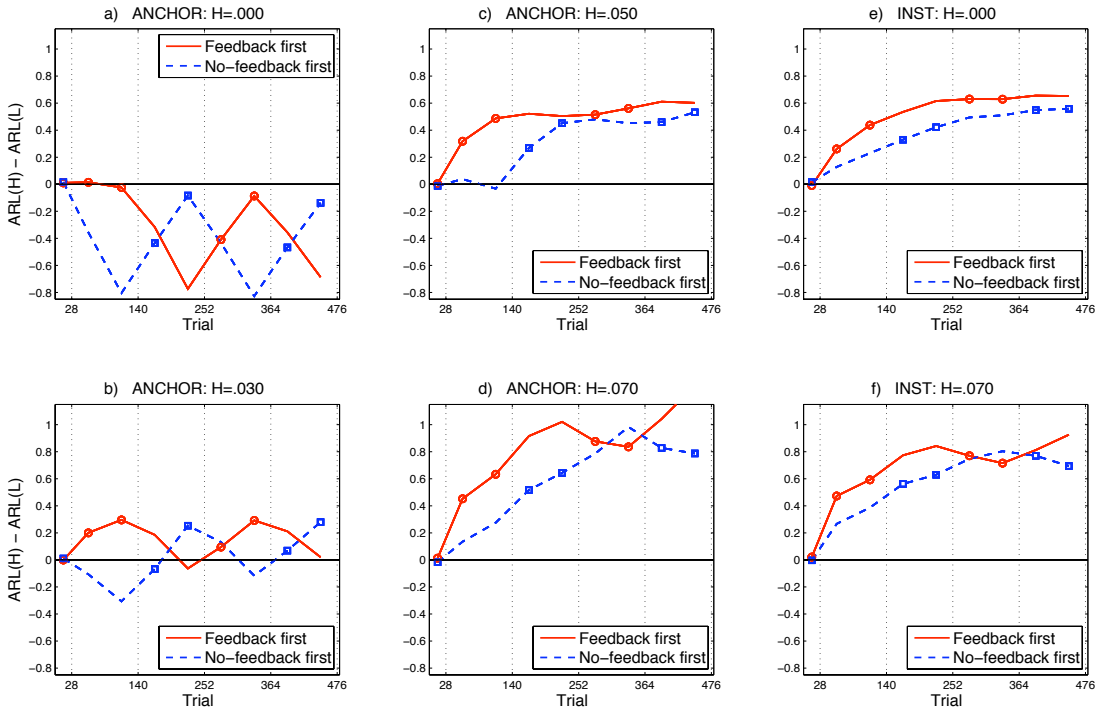

Figure 6. Context effects predicted by the prototype-based (ANCHOR) and instance-based (INST) models. Each panel reports a batch of runs with history weight *H* indicated above the panel. Based on the data in the corresponding panels of Figure 5. The average response level in low context [ARL(L)] is subtracted from that in high context [ARL(H)] to measure assimilation. The circles and squares mark blocks with feedback. INST always predicts assimilation. ANCHOR is consistent with context effects in either direction, modulated by feedback.

can be subtracted out. The Assimilation measure in Equation 2 depends on this assumption. However, there may be a weak interaction between the upward bias and context. It is obvious that Stimuli 2 and 1 are not in a 2:1 ratio but it is not obvious that Stimuli 9 and 8 are not in a 9:8 ratio. Thus, the upward bias may be stronger at the low end than the high end of the scale.

We ran a second experiment with different stimuli and tighter controls to rule out these alternative interpretations and test the generalizability of our results.

Experiment 2

Experiment 2 improves on Experiment 1 in three ways. First and foremost, the overall response-category frequencies are always uniform under the new design. A perfect responder will press each of the 7 response keys an equal number of times in each block. This makes it highly unlikely that response bias plays a significant role in this study. The context manipulation is preserved, but two different types of stimuli are mixed in each block, with presentation frequencies skewed in complementary directions. Concretely, there are motion stimuli and texture stimuli. The participants are instructed to rate the speed of motion on a scale of 1="slowest" to 7="fastest" and to rate the coarseness of texture on a scale of 1="lowest" to 7="highest."

We chose stimulus types that are as different from each other as possible in order to minimize the cross-talk between the two tasks in memory. When a motion stimulus is presented, for example, only motion anchors or instances compete to match it. Those involving textures are too dissimilar to ever be retrieved. If this is correct, Experiment 2 consists of two independent replications of Experiment 1.

The third improvement is the introduction of a monetary bonus contingent on accuracy. The bonus motivates the participants to use the interval scale prescribed by the instructions and to cast aside any *a priori* preferences for ratio scales, uniform frequencies, etc.

Method

Observers. Forty-one participants at Ohio State University were paid $6 plus a bonus that varied between $2.50 and $4.50 depending on their accuracy.

Stimuli and Apparatus. Each motion stimulus consisted of 150 black dots that moved coherently inside a grey circular aperture. The direction of motion was randomized across trials but all dots on a given trial moved in the same direction. The speed of each individual dot was constant throughout

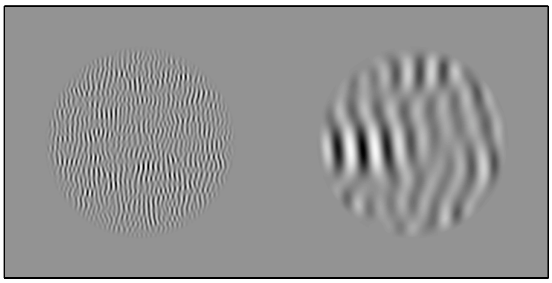

Figure 7. Sample filtered-noise texture stimuli for Experiment 2. Left: category 1, lowest "coarseness" (shortest wavelength). Right: category 7, highest "coarseness" (longest wavelength). Each patch was clipped within a circular aperture with diameter 7 degrees of visual angle. The orientation was randomized.

the lifetime of the dot and was drawn from a Gaussian distribution (e.g., Watamaniuk, Sekuler, & Williams, 1989). The mean of this Gaussian was 6, 7, 8, . . . , 12 degrees per second for categories 1, 2, 3, . . . , 7, respectively.[7] The standard deviation of the Gaussian was proportional to the mean (1 deg/sec for the slowest and 2 deg/sec for the fastest category). The participants were instructed to rate the average speed of the cloud of dots. The diameter of the aperture was 7 degrees of visual angle. As dots exited the aperture, they were replaced[8] with freshly sampled dots.

The texture stimuli were filtered-noise patches (Figure 7). Seven filters H_k were defined for the 7 categories k. Each filter had a Gaussian cross-section in the frequency domain:

$$H_k(f_x, f_y) = N_k \exp\left\{-\frac{1}{2}\left[\frac{(f_x - c_k)^2}{\sigma_{x,k}^2} + \frac{f_y^2}{\sigma_{y,k}^2}\right]\right\} + N_k \exp\left\{-\frac{1}{2}\left[\frac{(f_x + c_k)^2}{\sigma_{x,k}^2} + \frac{f_y^2}{\sigma_{y,k}^2}\right]\right\} \quad (3)$$

Equation 3 describes the amplitude profile of a Gabor wavelet (windowed sinusoidal grating) with spatial frequency c_k (Graham, 1989). The normalization constants N_k were chosen so that all filters had equal spectral energy. To generate one texture from a given category k, the algorithm generated a fresh matrix of iid Gaussian noise and applied the corresponding filter H_k. The center frequencies c_k were inversely related to the category labels k in a geometric progression with parameter $q = 0.4$ octaves (Equation 4). The "coarseness" of the texture was operationalized as the wavelength $\lambda_k = 1/c_k$. It varied from $\lambda_1 \approx 0.19$ to $\lambda_7 = 1.0$ degrees per cycle as illustrated in Figure 7.

$$c_k = 2^{q(7-k)} \approx 1.32^{(7-k)} \quad (4)$$

$$\sigma_{x,k} = \frac{2^b - 1}{(2^b + 1)\sqrt{2 \ln 2}}\, c_k \approx .146 c_k \quad (5)$$

$$\sigma_{y,k} = \frac{\pi\theta}{360\sqrt{2 \ln 2}}\, c_k \approx .333 c_k \quad (6)$$

The frequency bandwidth parameter b in Equation 5 controlled the uncertainty in spatial frequency. It was $b = 0.5$ octaves for all categories (full width at half height, in log-frequencies). The orientation bandwidth parameter θ in Equation 6 controlled the uncertainty in orientation. It was $\theta = 45$ degrees for all[9] categories. All textures were generated at vertical orientation and then rotated at a random angle.

All stimuli—moving dots and static textures—were generated in Matlab in real time and presented on a 21" NEC AccuSync 120 CRT at 96 frames/sec using PsychToolbox (Brainard, 1997) A software lookup table defined 255 evenly spaced luminance levels between $L_{min} \approx 2$ cd/m^2 and $L_{max} \approx 118$ cd/m^2. The displays were viewed binocularly from a chin rest placed 93 cm from the monitor.

Procedure. The participants were instructed that each block consisted of an equal number of motion and texture trials presented in random order and that there were feedback and no-feedback blocks. Nothing was mentioned about presentation frequencies within either stimulus type. The task was to rate the average speed of motion and the "coarseness" of the texture with a number from *1* to *7*. A brief demonstration presented examples of the slowest and fastest motion and of the finest and coarsest texture. The participants earned one bonus point for each correct answer. The current cumulative bonus was displayed above the fixation dot at all times except during the no-feedback blocks.

Each trial began with a brief beep. The stimulus was presented in the middle of the screen against gray background 500 ms later and continued until the observer pressed a key from 1 to 7. Invalid keys were ignored. Then the screen was cleared and a big white feedback digit (or an "X" in no-feedback blocks) appeared for 1100 ms. Each session lasted about 50 minutes and consisted of 700 trials.

Design. Block 1 had 28 trials; Blocks 2–13 had 56 trials. Block 1 presented 2 motions and 2 textures of each category, with feedback, in random order. The presentation frequencies in subsequent blocks varied in complementary ways depending on context: Fast-Low (FL) blocks contained 1, 2, 3, . . . , 7 presentations of motion stimuli 1, 2, 3, . . . , 7, respectively, and 7, 6, 5, . . . , 1 presentations of texture stimuli 1, 2, 3, . . . , 7. In Slow-High (SH) blocks the skewness of the two stimulus types was reversed. Note that a perfect responder would press each of the 7 response keys 8 times in each block regardless of context.

The participants were randomly assigned to 4 groups. Groups 1 (Fast-Low1) and 2 (Fast-Low2) presented 1 uniform block followed by 12 FL blocks. Groups 3 (Slow-

[7] We could not use a sequence with zero intercept because a parallax-like effect made the dot clouds appear to rotate rather than move sideways when the average speed was too low.

[8] Care was taken to correct the attrition bias that occurred as fast dots exited the aperture more often than slow dots.

[9] The standard deviation $\sigma_{y,k}$ in Equation 6 is proportional to the central frequency c_k for technical reasons involving a conversion from polar to Cartesian coordinates (Graham, 1989).

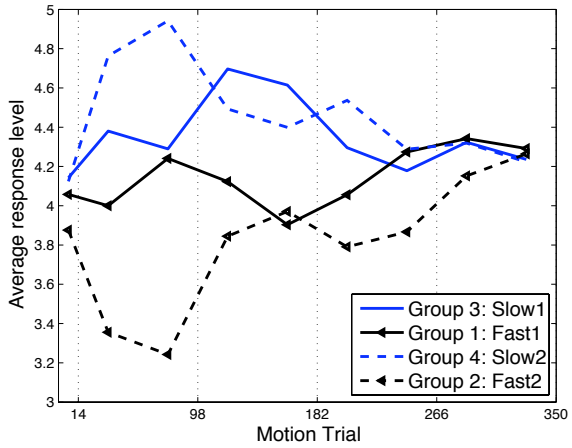

Figure 8. Average response levels for the motion-speed ratings in Experiment 2. Each line plots the mean of the individual profiles in the corresponding group. Groups 3 and 4: slow context (frequent slow stimuli; upward shifts indicate compensation). Groups 1 and 2: fast context (frequent fast stimuli; upward shifts indicate assimilation). The vertical grid lines mark transitions from feedback to no-feedback blocks or vice versa. Groups 1 and 3: feedback on Trials 1–98, 183–266. Groups 2 and 4: feedback on Trials 1–14, 99–182, 267–350.

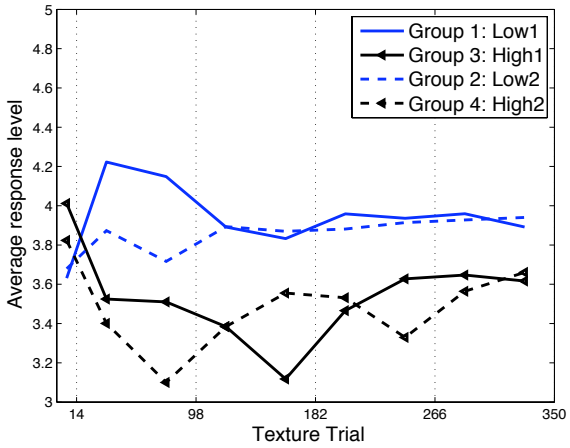

Figure 9. Average response levels for the texture-coarseness ratings in Experiment 2. Each line plots the mean of the individual profiles in the corresponding group. Groups 1 and 2: low context (frequent fine stimuli; upward shifts indicate compensation). Groups 3 and 4: high context (frequent coarse stimuli; upward shifts indicate assimilation). The vertical grid lines mark transitions from feedback to no-feedback blocks or vice versa. Groups 1 and 3: feedback on Trials 1–98, 183–266. Groups 2 and 4: feedback on Trials 1–14, 99–182, 267–350.

High1) and 4 (Slow-High2) presented 1 uniform block followed by 12 SH blocks. The feedback-first Groups 1 and 3 gave veridical feedback on blocks 1–4, 8–10 and no feedback on blocks 5–7, 11–13. The no-feedback-first Groups 2 and 4 gave no feedback on blocks 2–4, 8–10 and veridical feedback on blocks 1, 5–7, 11–13.

Dependent Variable. The dependent variable is the same as in Experiment 1—the average response level (ARL) calculated according to Equation 1. The two stimulus types are processed separately: 350 motion and 350 texture trials per participant. Each stimulus-response sequence is segmented into 9 nonoverlapping periods. Period 1 covers the initial uniform block and has 14 trials (per stimulus type). Period 2 covers block 2 and the first half of block 3 and has 42 trials. Period 3 covers the second half of block 3 and the entirety of block 4 and also has 42 trials. Periods 4 through 9 cover blocks 5 through 13 in an analogous fashion, each period spanning a block and a half and having 42 trials. The coefficients R_0 and a of the Stevens function are estimated by linear[10] regression from the 42 stimulus-response pairs in each period. The average response level for this period is $ARL = R_0 + 4a$, where 4 is the code of the middle stimulus. This procedure converts the raw data to two profiles—9 motion ARLs and 9 texture ARLs.

Results and Discussion

Figures 8 and 9 plot the mean ARL profiles for the motion-speed and texture-coarseness rating task, respectively. The line styles are the same as in Figure 2 to facilitate comparison with the line-length rating task of Experiment 1. The data show strong context effects, this time in a compensatory direction. The average response levels in Fast motion context (lines with triangular markers in Figure 8) are consistently lower than those in Slow context (lines with no markers). The compensatory effect is equally clear in the texture data—the ARLs in High (or coarse) context are lower than those in Low context (Figure 9). Both context effects are highly significant (motion $F(1, 37) = 21.4$, $p < .001$, $\eta_p^2 = .37$; texture $F(1, 37) = 34.4$, $p < .001$, $\eta_p^2 = .48$; mixed-design ANOVA as in Experiment 1).

This compensatory context effect falsifies the INST model (cf. Figure 5e,f) and challenges instance-based theories in general. The response-bias explanation does not seem to answer this challenge adequately. While it is impossible to rule out this explanation completely, it depends on the implausible assumption that the observers can keep separate counts of the response frequencies for the two stimulus types. Even if we assume for the sake of the argument that the participants could implement such bias and were willing to forfeit valuable bonus points in the process, it still remains unclear why the compensatory tendency is so much stronger in Experiment 2 than in Experiment 1.

ANCHOR, on the other hand, can generate compensatory context effects as discussed above. In fact, the empirical profiles in Figures 8 and 9 are very similar to the ANCHOR

[10] The correlation between the group-averaged ratings and the correct labels is 0.998 for motion and 0.997 for texture. Thus, the Stevens functions seem *locally* linear for our stimulus ranges.

profile in Figure 5a. The base-level activations in ANCHOR (and ACT–R more generally) have a strong but transient recency component (Equation 10 in Appendix A; see Petrov, 2006, for illustrative plots of the activation dynamics). When trials of different types are mixed in a block, each type dilutes the residual activation of the other. Thus, ANCHOR predicts weaker assimilation in heterogenous blocks than in homogenous blocks. This is exactly what we found—a compensatory effect in Experiment 2 and an assimilative effect in Experiment 1. Modeling this phenomenon in detail is a promising topic for future research.

Experiment 2 replicates the Context by Feedback interaction, particularly in the motion data. The zig-zag interaction pattern is most pronounced in Figure 8. Consider trials 15–98 for concreteness. With feedback (Groups 1 and 3, solid lines), there is hardly any context effect during this period. Without feedback (Groups 2 and 4, dashed lines), there is a massive compensatory effect. When feedback is discontinued in Groups 1 and 3, their ARLs diverge (trials 99–182); whereas when feedback is introduced in Groups 2 and 4, their ARLs converge.

The Context by Feedback interaction is statistically significant in the motion data ($F(1, 37) = 28.6$, $p < .001$, $\eta_p^2 = .44$) but not in the texture data ($F(1, 37) < 1$). This is driven by a general tendency of the texture ARLs to drift downwards during the no-feedback periods. The significant main effect of the Feedback factor (texture $F(1, 37) = 15.0$, $p < .001$, $\eta_p^2 = .29$) obscures the Context by Feedback interaction. The predicted zig-zag pattern is still evident in the High groups in Figure 9 (triangular markers). In the Low groups (no markers), however, the pattern is eliminated, even reversed perhaps. Recall that Experiment 1 produced analogous results, but there the ARLs tended to drift upwards without feedback. Thus, the interaction was strong in the Low groups in Figure 2 and weak in the High groups. Here it is the other way around. That a main effect can mask an interaction is well documented in the statistical literature (e.g., Keppel & Wickens, 2004). The length ARLs drifted upwards, the texture ARLs downwards. The motion ARLs are just right (no significant effect of Feedback, $F(1, 37) < 1$) and reveal the interaction in purest form.

We subtracted the ARL profiles obtained in complementary contexts (Equation 2) to calculate the Assimilation profiles in Figures 10 and 11. The overall compensatory context effect is evident from the consistently negative values. Note that Assimilation ≈ 0 during the initial uniform block (trials 1–14). The motion profile (Figure 10) is strikingly similar to the ANCHOR profile for low history weights (Figure 6a). The zig-zag pattern of Figure 3 is clearly replicated. The pattern is also discernible in the texture data (Figure 11).

Figure 11 contains two anomalous points in the feedback-first condition (solid line, trials 15–98). These are incompatible with ANCHOR (or INST). Tracing the problem back to Figure 9, it seems that the corresponding ARLs for Group 1 (Low1) are anomalously high (or that the subsequent ARLs have drifted downwards as discussed above). We have no good explanation for this. Given the noise in the data, however, it is not surprising to find two anomalous values among

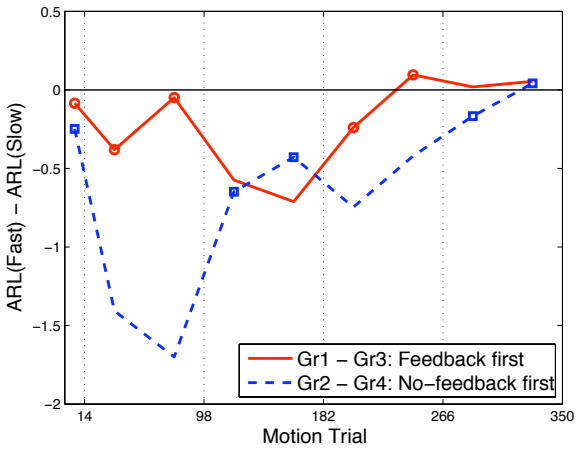

Figure 10. Context effects in the motion-speed rating task from Figure 8. The average response level in slow context [ARL(Slow)] is subtracted from that in fast context [ARL(Fast)] to measure assimilation. The vertical grid lines mark transitions from feedback to no-feedback blocks or vice versa. The consistently negative values indicate a compensatory context effect. It is stronger in the no-feedback periods (plain lines) than in the feedback periods (circles and squares). Both effects are predicted by the prototype-based model but inconsistent with the instance-based model.

the 117 points in Figures 2, 8, and 9. The compensatory effect during the late feedback blocks in Figure 9 are probably carried over from the preceding no-feedback blocks.

In conclusion, Experiment 2 replicated and improved upon Experiment 1. The motion data are particularly convincing. The compensatory main effect of Context and the Context by Feedback interaction rule out INST as a viable model of human category rating. ANCHOR, on the other hand, offers a natural and elegant account of this complex and interlocking behavioral pattern.

General Discussion

We presented evidence of assimilative (Experiment 1) and compensatory (Experiment 2) context effects in category rating with diverse stimulus sets. External feedback can reverse the direction of the context effect. These findings constrain the theory of direct psychophysical scaling and contribute to our understanding of how ratings are produced by human observers. They also constrain the theory of categorization.

Prototype- and exemplar-based theories make distinct predictions about the direction of context effects and their modulation by feedback. Computer simulations with representative members of each model class indicated that prototype-based models can exhibit both assimilatory and compensatory context effects, whereas instance-based models must always assimilate. Thus, the *unitary constraint* on the representational flexibility of the system can increase its behavioral flexibility via the *inversion rule* (Figure 1). Prototype-based categories cannot increase their coverage on the magnitude continuum without decreasing coverage on the oppo-

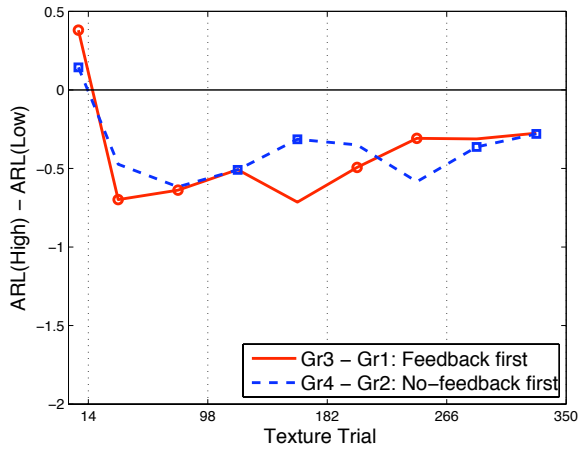

Figure 11. Context effects in the texture-coarseness rating task from Figure 9. The average response level in low context [ARL(Low)] is subtracted from that in high context [ARL(High)] to measure assimilation. The consistently negative values indicate a compensatory context effect, which is inconsistent with the instance-based model.

site side. This generates a compensatory tendency that counteracts the natural assimilative tendency in skewed contexts. These opposing forces can produce in ANCHOR an overall context effect in either direction (Figure 5).

Does this mean that ANCHOR is just too flexible and can fit anything but explain nothing (Roberts & Pashler, 2000)? The answer is an emphatic no because ANCHOR makes principled predictions about how context interacts with other variables (Petrov & Anderson, 2005). First, the assimilation originates in the base-level activation mechanism and hence any manipulation that weakens the activations should reduce the assimilative tendency in the data. This is what we found in Experiment 2. Mixing two stimulus types in the same block dilutes the recency component of the base-level activations of the anchors for each type. Thus, the behavioral pattern resembles the ANCHOR pattern generated with low history weight (Figure 5a). Second, the compensation originates in the competitive learning mechanism and hence any manipulation that constrains the anchor locations should reduce the compensatory tendency in the data. External feedback is one such manipulation. Petrov and Anderson (2005) demonstrated compensatory effects without feedback and assimilative effects with feedback. These studies manipulated context within and feedback between subjects. Here we replicate this result with the complementary design. ANCHOR, but not INST, makes a parameter-free prediction that the compensatory tendency in skewed contexts should be suppressed by feedback, leading to a characteristic zig-zag pattern. This is exactly what was observed.

In conclusion, the evidence suggests that category rating is based on unitary representations. Prominent theorists (e.g., Nosofsky & Johansen, 2000; Nosofsky & Zaki, 2002) have argued that people use instance-based representations in all categorization tasks. The present results identify a limit to such all-encompassing statements. Instance-based theories, despite their spectacular success in many other tasks, do not seem applicable to category rating without feedback.

Potential Challenges to Our Conclusions

INST retrieves a single exemplar per trial. While such single-exemplar proposals are not unprecedented (e.g., Ennis, Palen, & Mullen, 1988), most instance-based theories posit that the probability to classify a stimulus under a category is proportional to its *aggregate* similarity to all prior instances of this category. Thus, critics might argue that the failure of INST to fit our data does not constrain mainstream instance-based theory. A straightforward response to such criticism would be to fit the Generalized Context Model (GCM, Nosofsky, 1986) to our data. The problem is that GCM cannot do the no-feedback task. Without the stabilizing influence of a correction mechanism, a winner-takes-all dynamics sets in during the no-feedback blocks (Petrov & Anderson, 2005). The correction mechanism requires the retrieval of an individuated memory element on each trial. The corrections are based on the discrepancy between the remembered location and the target location (see Equation 14 in Appendix A).

All categorization models make *representation* assumptions and *retrieval* assumptions (see Ashby, 1992, for an excellent review). Our answer to the above criticism is that INST does embody the *representation* assumption central to all instance-based theories. It represents each category as a collection of instances in memory. Any model with non-unitary representations will fail to account for the compensatory tendencies in our data for the same reason that INST fails. The fundamental problem is that similarities always add (see Equation 17 in Appendix A) and thus a category can never lose strength in some region as it accrues a member in another region.

GCM has been extended to incorporate the idea that dissimilarity may play a role in categorization decisions. The similarity-dissimilarity model (SD-GCM, Stewart & Brown, 2005) assumes that the evidence for a category is the summed similarity to instances of that category plus the summed dissimilarity to instances of the opposite category. SD-GCM is motivated by the *category contrast effect*: The classification of a borderline stimulus is more accurate when preceded by a distant member of the opposite category than when it was preceded by a distant member of the same category (Stewart et al., 2002; Stewart & Brown, 2004). The problem with this approach is that it only works for binary classifications. Attempts to extend it to the rating task lead to highly implausible predictions. For example, SD-GCM predicts a strong tendency to respond *1* in High contexts because any short stimulus will be very dissimilar to the numerous long exemplars in memory. A further problem with SD-GCM is its apparent inability to account for the interaction between Context and Feedback that is the critical feature of our data.

In a different experimental paradigm, Smith and Minda (1998, 2000, 2002; Minda & Smith, 2001, 2002) docu-

mented many circumstances in which prototype models outperform instance-based models. Their results were challenged in various ways (Nosofsky, 2000; Nosofsky & Johansen, 2000; Nosofsky & Zaki, 2002; Stanton et al., 2002; Zaki et al., 2003). One controversy involves the response-scaling parameter γ in instance-based models (Smith & Minda, 1998, 2002; Myung et al., 2007). Nosofsky and Zaki (2002) argued that models without such parameter are artificially constrained. As our INST model lacks this particular parameter, our conclusions may seem vulnerable to the same criticism. They are not because they rest on qualitative patterns in the data rather than goodness of fit. INST's fundamental limitation in our paradigm is its inability to produce compensatory context effects. This is a structural limitation that cannot be circumvented by the introduction of a parameter that makes responding more or less deterministic.

Prototype models have also been challenged for being prone to overfitting (Olsson et al., 2004). Our simulations indicate that ANCHOR can indeed produce a broader range of qualitative patterns than INST (Figure 5). In that regard, the important outcome of the present experiments is that INST *cannot* fit the data, not that ANCHOR can. Note also that some potential outcomes could have falsified ANCHOR too. For example, it cannot fit compensatory context effects during Trials 29–140 in the feedback-first condition.

The Importance of Inductive Bias

It is not surprising to find evidence for unitary representations in category rating because they match the statistical structure of the target categories. Assigning a label to a novel exemplar is a form of induction. As such, it necessarily depends on *a priori* assumptions about the structure of categories (Hume, 1748/1962). Every representational scheme implicitly embodies such inductive bias. The foundational assumption of all memory-based classifiers is that similar items belong to the same category. In statistical terminology, the similarity-based inductive bias amounts to the assumption that categories have smooth probability density functions (Ashby & Alfonso-Reese, 1995; Nosofsky, 1990). Instance-based representations are equivalent to kernel density estimators and make no assumptions besides smoothness (Ashby & Alfonso-Reese, 1995). Prototype representations embody the additional assumptions of unimodality and symmetry. These are the conditions in which a distribution is well represented by its mean. In environments in which these assumptions are satisfied, the bias speeds up learning, improves classification accuracy, reduces the need for external feedback, and increases robustness. This is the case in category rating, where categories are contiguous regions on a unidimensional continuum and there are no exceptions. A prototype representation anticipates the regularities in these simple domains (Huttenlocher et al., 2000).

Flannagan, Fried, and Holyoak (1986) present convincing evidence that human observers are biased in favor of unimodal distributions. It was faster to learn a unimodal than a bimodal category. Also, subjects in the early stages of learning a bimodal category responded as if it were unimodal.

The compensatory tendencies in our data suggest an inductive bias for symmetry. Prototype-based representations enforce such symmetry; instance-based representations merely allow it. A bias for symmetry is beneficial for our task, even in nonuniform contexts, assuming symmetric perceptual noise. This is because each category in our experiment consists of a single stimulus. With feedback, both prototype- and instance-based systems converge to symmetric representations and thus behave identically. Without feedback, however, the systems' own mistakes violate the symmetry of categories. In skewed contexts, more misclassifications are made toward the frequent end of the continuum. This skews the representations in INST but not in ANCHOR where the unitary constraint enforces symmetry and thereby counteracts the destabilizing contextual influence.

Analyzing the two classes of systems in terms of their inductive biases helps explain why the decisive test occurs during the no-feedback blocks. The absence of feedback forces the system to rely on prior knowledge. Strongly biased systems have an advantage over weakly biased systems, provided of course that the bias matches the structure of the environment. Prototype-based systems have the strongest bias, followed by decision-bound systems, followed by instance-based systems (Ashby & Alfonso-Reese, 1995). All documented failures of prototype-based models (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1992, 1993; Medin & Shaffer, 1978; Nosofsky, 1992; Nosofsky et al., 1994; Nosofsky & Zaki, 2002) involve tasks that violate one or more prototype assumptions. The strong inductive bias of prototype-based systems is counterproductive in those cases.

Long training sessions with feedback reduce the importance of prior knowledge. In Bayesian terms, the likelihood dominates the prior. Systems with non-informative priors can behave optimally in such circumstances and instance-based models provide excellent accounts of the asymptotic strategy. This is consistent with converging evidence for "a progression from a strong reliance on prototypes to a strong reliance on exemplar memorization" (Smith & Minda, 1998, p. 1412). In that regard, it is notable that ANCHOR outperforms INST after more than 400 presentations of our seven stimuli. We attribute this to the confusability inherent in unidimensional perceptual continua.

Such behavioral data are very informative but should be interpreted with care because the link between behavior and the underlying representation is not always straightforward. In the current study, for example, the model with greater representational flexibility (INST) has lesser behavioral flexibility. The study of Nosofsky and Stanton (2005) is another example. It involved two-dimensional stimuli (Munsell color chips), binary classification, and probabilistic feedback for certain "critical" stimuli. In an ingenious manipulation, both categories had asymmetrical, kidney-shaped densities but the asymmetry of Category A mirrored that of Category B so that the optimal decision bound was still linear. Thus, a prototype-based classifier would maximize performance even though it misrepresented the kidney-shaped densities. The inductive bias of a prototype representation is that each *individual* category is symmetrical, not that the constel-

lation of categories is symmetrical. Only an instance-based scheme can represent kidney-shaped categories. Ironically, these accurate representations predict slower and less accurate responses to the critical stimuli in Nosofsky and Stanton's (2005) configuration. The data favored the instance-based model. One interpretation of this finding is that the objective of the human system is not only to optimize performance on the current task but also to build an accurate internal model of the environment in anticipation of future tasks. Such combination of task-driven and model learning has been shown to generalize better than purely task-driven learning in neural networks (O'Reilly, 2001).

Related Research

Cohen, Nosofsky, and Zaki (2001) manipulated category variability in ways similar to our context manipulation. An equidistant transfer stimulus was more likely to be classified into a low-variability than a high-variability category. This is analogous to the compensatory context effects in our paradigm and is consistent with ANCHOR. However, further increases of the variance of the high-variability category increased the probability to classify the transfer stimulus into it, which is consistent with neither ANCHOR nor INST. Subsequent experiments revealed complications (Stewart & Chater, 2002). This is a topic for further investigation.

There is middle ground between prototype- and instance-based classifiers. It is to store several memory elements per category but significantly fewer than the total number of exemplars encountered so far (e.g., Busemeyer et al., 1984; Homa, Dunbar, & Nohre, 1991; Smith & Minda, 2000). One advanced model along those lines is SUSTAIN (Love et al., 2004). It creates new elements (or *clusters*) only when a surprise occurs. With feedback, that is when the teacher corrects the model's response; without feedback, a surprise occurs when the similarity between a new item and any existing cluster is less than a threshold parameter. By varying this threshold, SUSTAIN can enforce the unitary constraint to varying degrees. The model begins with simple representations and introduces complexity only when necessary. It is designed for multidimensional spaces and is equipped with the requisite attentional machinery. It is not well equipped to handle context effects in category rating but can be extended with base-level activations and a correction mechanism. Would such an extended version be compatible with our data? The answer is no. To capture the compensatory context effects, SUSTAIN must keep a single cluster per response category. This may seem a simple matter of setting the recruitment threshold high. However, a problem occurs during the feedback blocks. SUSTAIN will make mistakes and be "surprised" by the feedback. Many mistakes will be blamed on the decision procedure but some trace back to irreducible perceptual overlap (B. Love, personal communication, 30 April 2008). These irreducible surprises will recruit multiple clusters for every response category. Thus, SUSTAIN seems bound to behave as an instance-based model in all tasks with perceptually confusable stimuli.

The Relative Judgment Model (RJM, Stewart et al., 2005) and the Memory and Contrast model (MAC, Stewart et al., 2002; Stewart & Brown, 2004) emphasize the importance of a comparison process that calculates differences between magnitudes. We agree that relative judgments are important and incorporate them in ANCHOR's correction mechanism. Memory retrieval in ANCHOR is based on absolute magnitudes whereas corrections are based on differences. The interplay between these two factors allows ANCHOR to work without external feedback, which neither RJM nor MAC can do.

This article focused on the distinction between prototype and instance-based models. Decision bound models are another prominent class in the categorization literature (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1993; Ashby & Townsend, 1986; Maddox & Ashby, 1993; Treisman & Williams, 1984). Under some reasonable assumptions, prototype models are equivalent to minimum-distance classifiers with linear bounds (Ashby & Gott, 1988; Ashby & Alfonso-Reese, 1995). The strict mathematical proof does not apply to ANCHOR because of its correction mechanism, which is an innovation relative to all decision bound theories. Still, all these theories seem consistent with the outcome of the present experiment—that category rating is based on unitary representations. With unidimensional stimuli, decision bounds are just points on the continuum and the system needs $N - 1$ criteria for N response categories (Torgerson, 1958; Treisman & Williams, 1984). Thus, the complexity of the internal representation is tied to the number of categories rather than the number of trials. In that sense, decision bounds are unitary representations consistent with our data. The Category Density Model (Fried & Holyoak, 1984) formalizes this idea. It assumes that category representations are (multivariate) Gaussians and incrementally updates the means and variances of these Gaussians. Decision bounds are then derived from likelihood ratios (Ashby & Townsend, 1986; Fried & Holyoak, 1984).

A growing number of theories posit two or more systems for categorization (e.g., Ashby & Ell, 2001; Ashby et al., 1998; Erickson & Kruschke, 1998; Nosofsky et al., 1994). A common assumption in these theories is that the different systems compete to categorize a given stimulus. ANCHOR also has multiple mechanisms but they cooperate rather than compete. ANCHOR's memory system is implicit, whereas its correction mechanism is explicit (cf. Ashby et al., 1998). The former is automatic, tracks the statistics of the environment, and is responsible for the "first guess" on each trial. Human observers, however, often second-guess themselves. This is captured by ANCHOR's explicit correction strategy. It embodies knowledge about the number and order of categories and generates the stimulus-response homomorphism that is the defining feature of scaling. The cooperative interaction between the implicit and explicit components in ANCHOR is crucial for its ability to unfold the scale and maintain stability in non-uniform and non-stationary environments without feedback (Petrov & Anderson, 2005).

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences*, *5*(5), 204–210.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 50–71.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 638–648.

Chase, S., Bugnacki, L. D., Braida, L. D., & Durlach, N. I. (1983). Intensity perception. XII. Effect of presentation probability on absolute identification. *Journal of the Acoustical Society of America*, *73*(1), 279–284.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, *29*(8), 1165–1175.

Ennis, D. M., Palen, J. J., & Mullen, K. (1988). A multidimensional stochastic theory of similarity. *Journal of Mathematical Psychology*, *32*, 449-465.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(2), 241–256.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 234–257.

Graham, N. V. (1989). *Visual pattern analyzers*. New York: Oxford University Press.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage Publications.

Homa, D., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 444-458.

Hume, D. (1962). *Enquiry concerning human understanding* (2nd ed.; L. A. Selby-Bigge, Ed.). Oxford: Oxford University Press. (Original work published 1748)

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*(2), 220–241.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 316–332.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70.

Marks, L. E. (1993). Contextual processing of multidimensional and unidimensional auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 227-249.

The MathWorks. (2004). Optimization Toolbox, version 3. For use with MATLAB [Computer software manual]. Natick, MA: The MathWorks, Inc.

Medin, D. L., & Shaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 775–799.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 275–292.

Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review*, *14*, 1043–1050.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, *34*, 393–418.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3–27.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in*

honor of William K. Estes, Vol. 1 (pp. 149–167). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nosofsky, R. M. (1997). An exemplar-based random-walk model of speeded categorization and absolute judgment. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 347–365). Mahwah, NJ: Lawrence Erlbaum Associates.

Nosofsky, R. M. (2000). Exemplar representation without generalization? Comment on Smith and Minda's (2000) "Thirty categorization results in search of a model". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1735–1743.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*(3), 375–402–233.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception & Performance*, *31*(3), 608–629.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 924–940.

Olsson, H., Wennerholm, P., & Lyxzèn, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 936–941.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*, 1199–1241.

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception: Vol. II. Psychophysical judgment and measurement* (pp. 127–141). New York: Academic Press.

Petrov, A. A. (2006). Computationally efficient approximation of the base-level learning equation in ACT-R. In D. Fun, F. Del Missier, & A. Stocco (Eds.), *Proceedings of the Seventh International Cconference on Cognitive Modeling* (pp. 391–392). Trieste, Italy: Edizioni Goliardiche.

Petrov, A. A. (2008). Additive or multiplicative perceptual noise? two equivalent forms of the ANCHOR model. *Journal of Social & Psychological Sciences*, *1*(2), 123–143.

Petrov, A. A., & Anderson, J. R. (2000). ANCHOR: A memory-based model of category rating. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Cconference of the Cognitive Science Society* (pp. 369–374). Hillsdale, NJ: Lawrence Erlbaum Associates.

Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based anchor model of category learning and absolute identification. *Psychological Review*, *112*(2), 383–416.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411–1436.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3–27.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 800–811.

Stanton, R. D., Nosofsky, R. M., & Zaki, S. R. (2002). Comparisons between exemplar similarity and mixed prototype models using a linearly separable category structure. *Memory & Cognition*, *30*(6), 934–944.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*(6), 377–411.

Stewart, N., & Brown, G. D. A. (2004). Sequence effects in the categorization of tones varying in frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 416–430.

Stewart, N., & Brown, G. D. A. (2005). Similarity and dissimilarity as evidence in perceptual categorization. *Journal of Mathematical Psychology*, *49*(5), 403–409.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 3–11.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*(4), 881–911.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 893–907.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, *91*(1), 68–111.

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

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: The integration of direction information. *Vision Research*, *29*(1), 47–59.

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*(4), 387–402.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1160–1173.

Appendix A
Model Equations and Parameters

ANCHOR and INST are governed by a set of mathematical equations that define conditional probability distributions of the following five variables: stimulus S, target magnitude M, anchors A_i, correction I, and overt response R. The perceptual processing in both models is described by Equation 7 that converts the stimulus intensity S into an internal magnitude M. The exponent $n = 1.0$ is determined from Stevens power law for line length[11] (Stevens, 1957; Stevens & Galanter, 1957). The scaling factor a is set arbitrarily to $1/1000$ so that magnitudes fall in the 0–1 range. The perceptual noise ε_p is drawn from a Gaussian distribution with zero mean and unit variance. Because of the multiplication in Equation 7, the standard deviation of the magnitude distribution is proportional to its mean. The coefficient $k_p = 0.04$ is estimated from the Weber fraction for line length (Petrov & Anderson, 2005). Equation 7 is consistent with both Weber's and Stevens's laws. An alternative, additive-noise equation can also be used without altering the predictions of the theory (Petrov, 2008).

$$M = aS^n(1 + k_p\varepsilon_p) \tag{7}$$

The magnitude A_i of each anchor i on a given trial is a noisy perturbation of its current location L_i. The memory-noise Equation 8 is analogous to the perceptual Equation 7. A new perturbation ε_m is drawn for each element on each trial from a Gaussian distribution with zero mean and unit variance. The coefficient k_m is a free parameter that scales the memory noise.

$$A_i = L_i\,(1 + k_m\varepsilon_m) \qquad \text{for each element } i \tag{8}$$

Each anchor has a base-level activation B_i that quantifies its availability as a function of the history of prior uses of the corresponding response. The base-level Equation 9 is taken verbatim from the ACT–R architecture (Anderson & Lebiere, 1998, p. 124). It is a logarithm of a sum of powers with decay rate $d = 0.5$. Each new use of the anchor adds another term to this sum, which then decays independently. The total count so far is denoted by n, and t_l are the individual time lags from the present.

$$B = \ln\left[\sum_{l=1}^{n} t_l^{-d}\right] \tag{9}$$

$$B \approx \ln\left[t_{last}^{-0.5} + \frac{2(n-1)}{\sqrt{t_{life}} + \sqrt{t_{last}}}\right] \tag{10}$$

As Equation 9 is expensive to compute, all ANCHOR simulations use the approximate Equation 10 (Petrov, 2006). It retains only three critical pieces of information about the anchor: the time since its creation t_{life}, the time since its most recent use t_{last}, and the total number of uses n. The approximation preserves the three qualitative features of the activation dynamics: (a) sharp transient peak immediately after each use, (b) decay in the absence of use, and (c) gradual buildup of strength with frequent use. The third property drives ANCHOR's tendency for assimilative context effects under skewed stimulus distributions. The first property explains why this tendency is diminished when stimuli of different types are mixed within a block in Experiment 2.

Because each exemplar in INST is used only once, the sum in Equation 9 contains only one term. This produces the simple decay Equation 11. Thus, the activation of an exemplar in INST equals the activation of an anchor in ANCHOR that has been created on the same trial as the exemplar and has not been used ever since. The buildup of strength with frequent use is driven in INST by the accumulation of separate instances.

$$B = \ln t_{life}^{-d} = -d \ln t_{life} \tag{11}$$

The memory elements compete to match the target M on each trial. This competition is governed by two equations. Equation 12 produces *goodness scores* G_i, and the *softmax* Equation 13 converts them into selection probabilities P_i. Only one element is selected per trial.

$$G_i = -\,|M - A_i| + HB_i \tag{12}$$

$$P_i = \frac{\exp(G_i/T)}{\sum_u \exp(G_u/T)} \tag{13}$$

Each goodness score G_i is a sum of two terms: similarity $-|M - A_i|$ and history HB_i. The history weight parameter H controls the relative strength of these factors. The direction of context effects (assimilative or compensatory) in ANCHOR depends mostly on this parameter (see Figures 5 and 6). The temperature parameter T controls the stochasticity of the softmax selection. Values close to zero produce deterministic choice, whereas large values result in nearly random sampling.

Equations 12 and 13 follow the ACT–R notational convention (Anderson & Lebiere, 1998). The influential Generalized Context Model (GCM, Nosofsky, 1986) uses a different notation in which the exponentiation is incorporated into the definition of similarities. ACT–R's temperature T is the inverse of GCM's sensitivity parameter c. The two formulations are mathematically equivalent, except that GCM treats exemplar strengths as free parameters whereas the base-level activations in INST are grounded in the rational analysis of memory (Anderson & Milson, 1989).

The winning anchor (in ANCHOR) or instance (in INST) represents the "first guess" about the classification of the current stimulus. It provides a reference point for the correction mechanism, which is the same in both models. The target magnitude M is compared to the magnitude A of the element retrieved from memory:

$$D = M - A \tag{14}$$

The size and magnitude of the discrepancy D then determines the correction. There are five possible increments:

[11] Only the line-length data are modeled here. The psychophysics of the motion and texture stimuli in Experiment 2 is beyond the scope of this article.

$I \in \{-2, -1, 0, 1, 2\}$. The decision rule is based on four criteria described by two free parameters: $\{-3c^-, -c^-, c^+, 3c^+\}$. The cutoffs are multiplied by the category width $W = 0.040$ magnitude units (40 pixels). For example, an increment $I = +1$ is made when $c^+W < D \leq 3c^+W$. The corrected response R is clipped at 1 or 7 if necessary:

$$R = R_A + I \qquad \text{clipped between } R_{min} \text{ and } R_{max} \tag{15}$$

An ideal observer would use $c^+ = c^- = 0.5$ (Petrov & Anderson, 2005). Thresholds greater than 0.5 produce conservative correction strategies consistent with the sequential and anchoring effects in the data (Petrov & Anderson, 2005). The upward and downward corrections need not be symmetric. In particular, when $c^+ < c^-$, the average response levels (ARLs) settle above the midpoint of the response scale. This allows the models to account for the systematic upward trend in the ARLs from Experiment 1.

Importantly, the two models differ in the mechanism that learns the locations L_i of the memory elements. ANCHOR uses the *competitive learning rule* in Equation 16. The new anchor location $L_{i^*}^{(t+1)}$ is a linear combination of the old location $L_{i^*}^{(t)}$ and the target magnitude $M^{(t)}$ on trial t. The learning rate α is fixed to 0.3 based on previous research (Petrov & Anderson, 2005). Exactly one anchor, with index i^*, is updated on each trial. If there is feedback, this is it; otherwise the system's own response designates the anchor for update.

$$L_{i^*}^{(t+1)} = (1 - \alpha)L_{i^*}^{(t)} + \alpha M^{(t)} \tag{16}$$

INST does not use competitive learning. Instead, the target magnitude M on each trial is stored as a separate exemplar. Because of this, INST always predicts assimilative context effects under skewed stimulus distributions:

$$P_J = \frac{\sum_{j \epsilon J} \exp(G_j/T)}{\sum_{u \epsilon U} \exp(G_u/T)} \tag{17}$$

The probability P_J to retrieve an instance of category J is the sum of the individual retrieval probabilities of all members $j \epsilon J$ of that category (Equation 13). The sum in the denominator is over the total memory pool U. Clearly, every member j makes a positive contribution to P_J. Categories with many members thus exert stronger gravitational fields than do categories with few members.

Appendix B
Parameter Search

Five parameters were allowed to vary to optimize the fits in Figure 4. The best-fitting values are reported in Table B1. The default values from the original ANCHOR publication (Petrov & Anderson, 2005) are also listed for comparison. Three other constants are not listed because they are not treated as free parameters here. The perceptual noise coefficient $k_p = 0.04$ is constrained by the Weber fraction (Petrov & Anderson, 2005). The activation decay rate $d = 0.5$ is constrained by the ACT–R architecture (Anderson & Lebiere, 1998). The learning rate $\alpha = 0.3$ in Equation 16 does not apply to the INST model. To equate the number of free parameters, it was not allowed to vary in ANCHOR either.

Table B1
Free parameters of the ANCHOR and INST models. The Default column lists the values used to generate Figures 5 and 6. The two rightmost columns report the best-fitting values used to generate Figure 4.

Parameter	Default	ANCHOR	INST
History weight H (Eq. 12)	varies	0.040	0.050
Memory noise k_m (Eq. 8)	0.070	0.083	0.050
Temperature T (Eq. 13)	0.040	0.032	0.030
Correction cutoff c^-	0.80	0.60	0.60
Correction cutoff c^+	0.80	0.45	0.42

The two models were fitted using a combination of sequential quadratic programming[12] and grid search of the parameter space. The objective was to minimize the root mean squared error (RMSE) between the model ARLs and the group-level data in Figure 5. Because the quadratic algorithm had poor convergence with respect to parameters c^+, c^-, and H, they were explored on a grid. The algorithm then minimized the RMSE with respect to k_m and T. The best-fitting values are reported in Table B1. Note the asymmetric correction thresholds. The minimal $RMSE$ was 0.173 for INST and 0.154 for ANCHOR. The latter fit could have been improved further if the learning rate α were allowed to vary ($RMSE = 0.138$ for $\alpha = 0.40$).

Appendix C
Simulation Experiment Method

The simulations that generated Figures 5 and 6 used stimulus sequences conforming to the design of Experiment 1. Each sequence consisted of 17 blocks of 28 trials each. The uniform blocks contained 4 presentations of each stimulus. The low (positively skewed) blocks contained 7, 6, 5, ..., 1 presentations of Stimuli 1, 2, 3, ..., 7, respectively. The high blocks were skewed in the opposite (negative) direction. The order of presentation within each block was randomized. There were six types of stimulus sequences (or groups). Five of those (U1, L1, H1, L2, and H2) were the same as in Experiment 1. A no-feedback-first, uniform control (U2) was added for completeness.

The simulation was organized in batches. Each batch ran a given model with given parameters on 250 replications of each sequence type. Informal explorations indicated that the pattern of context effects depended mostly on the history

[12] The *fmincon* function in Matlab's Optimization Toolbox (The MathWorks, 2004). Transcripts of all model-fitting sessions are available at `http://alexpetrov.com/proj/anchor/`

weight parameter H. Reasonable variations of the other parameters did not introduce any qualitatively new patterns. All simulations reported in Figures 5 and 6 were produced with default values for all parameters except H. The defaults are from the original ANCHOR publication (Petrov & Anderson, 2005) and are listed in the Default column in Table B1. The specific H values are reported on the corresponding figure panels. Each run was initialized with 7 perfectly placed memory elements, one per response category.

The sequence of stimulus-response pairs for each run was converted to an ARL profile in the same way as the data from Experiment 1. The 250 replications in each group were then averaged together. Each panel on Figures 5 reports the 6 mean ARL profiles generated with a particular parameter setting. Figure 6 combines the profiles according to Equation 2.