Journal of Experimental Psychology: General
2008, Vol. 137, No. 1, 182–200

0096-3445/08/$12.00 DOI: 10.1037/0096-3445.137.1.182

Direct-Comparison Judgments: When and Why Above- and Below-Average Effects Reverse

Paul D. Windschitl, Daniel Conybeare, and Zlatan Krizan
University of Iowa

Above-average and below-average effects appear to be common and consistent across a variety of judgment domains. For example, several studies show that individual items from a high- (low-) quality set tend to be rated as better (worse) than the other items in the set (e.g., E. E. Giladi & Y. Klar, 2002). Experiments in this article demonstrate reversals of these effects. A novel account is supported, which describes how the timing of the denotation of the to-be-judged item influences attention and ultimately affects the size or direction of comparative biases. The authors discuss how this timing account is relevant for many types of referent-dependent judgments (e.g., probability judgments, resource allocations) and how it intersects with various accounts of comparative bias (focalism, generalized-group, compromise between local and general standards [LOGE]).

Keywords: comparative judgment, referent-dependent judgments, above-average effect, nonselective superiority bias, focalism

How attractive is this hotel compared with the others with vacancies?

How reliable is the Civic relative to the Corolla and Protégé?

How mathematically talented is Erin compared with her classmates?

How good is this application relative to those from the other applicants?

How clever am I compared with my coworkers?

These types of questions elicit direct-comparison judgments. Each question specifies the referent or referents against which a target entity must be compared. Direct-comparison questions are frequently confronted in everyday life, and they underlie a wide range of decisions and behaviors (e.g., how much to pay for a Civic, whether to admit an applicant).

Research in multiple domains has shown that people's direct-comparison judgments are susceptible to a pair of biases that can generally be referred to as above- and below-average effects (e.g., Burson, Larrick, & Klayman, 2006; Chambers & Windschitl, 2004; Giladi & Klar, 2002; Klar & Giladi, 1997; Kruger, 1999; for work on closely related biases see Chambers, Windschitl, & Suls, 2003; Moore & Kim, 2003; Posavac, Brakus, Jain, & Cronley, 2006; Posavac, Sanbonmatsu, & Ho, 2002; Ross & Sicoly, 1979; Shepperd, Helweg-Larsen, & Ortega, 2003; Windschitl, Kruger, & Simms, 2003; Weinstein, 1980; Weinstein & Lachendro, 1982). For example, Kruger (1999) found that when people are asked to make direct-comparison judgments involving how skillful they are relative to others, they tend to report being above average among their peers when the skill in question is relatively easy (e.g., operating a computer mouse) but below average among their peers when the skill in question is relatively hard (e.g., sculpting human figures from clay). In research on direct-comparison judgments regarding the likelihood of experiencing events, people have tended to report being less likely than others to experience an event when that event is generally rare (e.g., going blind, being involved in a boating accident) but more likely than others to experience an event when the event is generally common (e.g., tearing a hole in your clothes, receiving a speeding ticket) (see Chambers et al., 2003; Kruger & Burrus, 2004; see also Klar, 1996; Weinstein & Lachendro, 1982).

The present article concerns related but more generic versions of above- and below-average effects. These more generic versions were documented by Giladi and Klar (2002) in their research on member-to-group comparisons (see also Klar, 2002; Klar & Giladi, 1997). Participants in their studies were exposed to groups of items that were either generally high or generally low on a specific dimension, and then they rated how a randomly selected member from the group compared with the others from that group. For example, in one study, participants smelled either a set of six pleasant soaps or a set of six unpleasant soaps before being asked to judge how a soap that was randomly selected from the set compared with the average of the group of others from that set. For participants in the pleasant soaps condition, the randomly selected soap tended to be rated as better than the remaining group from that set. For participants in the unpleasant soaps condition, the randomly selected soap tended to be rated as worse than the remaining group from that set. Giladi and Klar (2002) referred to these effects as *nonselective superiority* and *nonselective inferiority biases*, respectively, but we use the broader terms of *above-average* and *below-average* effects. Analogous above- and below-average effects were also found for randomly selected targets from desirable and undesirable sets of songs, healthy and unhealthy

Paul D. Windschitl, Daniel Conybeare, and Zlatan Krizan, Department of Psychology, University of Iowa.

This work was supported by Grant SES 03-19243 to Paul D. Windschitl from the National Science Foundation. We thank Jerry Suls and Jason Rose for their comments regarding this article.

Correspondence concerning this article should be addressed to Paul D. Windschitl, Department of Psychology, University of Iowa, Iowa City, IA 52242. E-mail: paul-windschitl@uiowa.edu

foods, high- and low-prestige occupations, attractive and unattractive faces, and liked and disliked people (see Giladi & Klar, 2002; Klar, 2002; Klar & Giladi, 1997; Suls, Krizan, Chambers, & Mortensen, 2007).

An important clue as to what causes the above- and below-average effects on direct-comparison judgments concerns the results of the indirect method of assessing people's judgments. In the indirect method, participants are asked separate absolute-evaluation questions about (a) the randomly selected member/item and (b) the remaining group of members/items. The researcher then compares these two judgments to determine whether there is a comparative bias. Within experiments detecting robust above- and below-average effects in direct-comparison judgments, the indirect method does not routinely reveal such effects. Although some previous studies have produced the same above- and below-average effects on indirect comparisons as on direct comparisons (e.g., Study 2 of Klar & Giladi, 1997), most studies have shown either no effects on indirect comparisons (e.g., Study 3 of Klar, 2002) or sometimes even a reversed effect (e.g., Study 6 of Klar, 2002; see also Krizan & Suls, 2007). Hence, a comprehensive explanation of comparative bias in member-to-group judgments must explain how above- and below-average effects can arise in direct-comparison judgments, even when people's separate judgments of the member and the remaining group do not reveal a comparative bias.

Differential Weighting

The key to resolving this apparent discrepancy is the notion of differential weighting. That is, when people make a direct-comparison judgment in a member-to-group paradigm, their evaluations of the member garner more weight than do their evaluations of the remaining group, even though the two evaluations should have equal weight. This notion of differential weighting has been applied either implicitly or explicitly in explanations of comparative bias not only for member-to-group judgments (see Klar, 2002; Klar & Giladi, 1997) but also for a variety of other types of direct-comparison judgments (e.g., Chambers et al., 2003; Krizan & Windschitl, 2007; Kruger, 1999; Kruger & Burrus, 2003; Moore & Kim, 2003; Tversky, 1977; Windschitl et al., 2003). However, for many readers, the process by which differential weighting leads to above- or below-average effects might be a mysterious one; how can weighting one evaluation more than another lead to a comparative bias even though the items being evaluated would receive the same evaluation if those evaluations were solicited separately (in the indirect method)?

There are many possible ways in which differential weighting can occur (see Chambers & Windschitl, 2004), but we provide examples of two interrelated ways, and we briefly explain how these ways can ultimately yield comparative biases. First, differential weighting in a comparative judgment can occur when people simply attend to one item's attributes more than the attributes of other relevant items. If we assume people make their comparative judgment on the basis of a limited (or nonexhaustive) consideration of the items' attributes, comparative bias can result. Consider a case in which a person is asked to compare the relative attractiveness of three cities (Seattle, Boston, and Miami), and assume that when asked about these cities separately, he or she would give the cities the same very positive evaluations. If, while considering the items' attributes prior to making a comparative judgment, the person attends to Seattle's attributes more so than to Boston's or to Miami's, then at the point of judgment he or she may have considered numerous positive attributes about Seattle but only a few positive attributes about the other two cities. Assuming that this proportion (the proportion of positive attributes considered for Seattle relative to those considered for the other cities) drives the comparative judgment, Seattle would be judged to be more attractive than the other cities. A second and related way in which differential weighting can influence comparative judgment is when people tend to translate their general evaluation of one item into their comparative assessment, even though they should be comparing their evaluations of that item against their evaluations of other items that were specified by the comparative question. This hypothesized process is a key component of Giladi and Klar's (2002) *compromise between local and general standards* (LOGE) approach to explaining biases in member-to-group comparisons. We discuss the LOGE approach in more detail later. However, in terms of our three cities example, the LOGE approach would suggest that when the person generates a comparative assessment of Seattle, he or she might dwell primarily on whether Seattle is better than a general standard (all known cities) rather than on how Seattle fares against the more appropriate local standard (Miami and Boston).

The above paragraph describes two types of differential weighting and applies them to an individual example. However, what would cause systematic cases of differential weighting to occur, such that above-average and below-average effects would be observed at a group level? There are, in fact, many systematic reasons why people would weight some items more than others. In reviewing these reasons, Chambers and Windschitl (2004) organized them into three types of accounts—*egocentrism accounts, focalism accounts,* and *generalized-group accounts*. Egocentrism accounts are not directly relevant to this article because they address comparisons between the self and others. Focalism accounts assume that the target entity (also known as the *focal entity*) carries more weight in the comparison judgment because of the very fact that the question soliciting the judgment specified that entity as the target rather than as a referent (see Chambers et al., 2003; Chambers & Windschitl, 2004; Eiser et al., 2001; Hodges, Bruininks, & Ivy, 2002; Hoorens, 1995; Moore & Kim, 2003; Otten & Van der Pligt, 1996; Suls et al., 2007; Tversky, 1977; Windschitl et al., 2003; see also Schkade & Kahneman, 1998; Wilson, Wheatley, Meyers, Gilbert, & Axsom, 2000). Focalism accounts assume that a question such as *How tall is A compared with B?* is psychologically different from *How tall is B compared with A?* This notion is closely related to Tversky's (1977) point regarding similarity judgments: The statement *A is like B* is not psychologically equivalent to the statement *B is like A*. One version of a focalism account assumes that people simply direct more attention to the target entity than to referent entities perhaps because evidence regarding the target seems directly related to the judgment, whereas evidence related to the referent seems less directly related (see, e.g., Windschitl, 2000). Another version of a focalism account might assume that people give more weight to the target because their attention is directed first to the target, and, although they later take into consideration the referent entity, this consideration or adjustment process is incomplete (Chambers & Windschitl, 2004).

Generalized-group accounts are distinct from focalism accounts. Generalized-group accounts assume that whenever a single entity

(e.g., an item or person from a group) is compared with a group or generalized representation of a group (e.g., "other items," "the average student"), the single entity will carry more weight (see Chambers & Windschitl, 2004; Windschitl et al., 2003). One version of a generalized-group account might suggest that evaluating a single item is easier than evaluating a group or a generalized representation, causing people to direct their attention more toward the easier of the two types of evaluations (i.e., more toward the attributes of the single item; see Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Klar & Giladi, 1997). Another related version suggests that people have more confidence in their assessments of a single entity than of a group, so they place disproportionate weight on the former evaluation (Chambers & Windschitl, 2004).

Despite their differences, both the focalism accounts and the generalized-group accounts predict that if people are asked to judge how randomly selected singletons from a generally favorable set of items compare with the rest of the set, their responses will be biased upward, producing an above-average effect among a group of participants. The account also predicts systematic below-average effects, if the singletons are drawn from a generally unfavorable set.

The Present Work

The present research began with an unexpected finding from an experiment involving a member-to-group judgment paradigm (Experiment 1). Namely, we found robust evidence for a complete reversal of the usual above-average effects. Participants had judged singletons from attractive sets to be less attractive than the rest of the set. On the face of it, this finding threatened the validity of the usual focalism and generalized-group accounts.

We conducted six other experiments to search for the critical sources of the reversal and to ultimately provide evidence for our updated theoretical account of why comparative biases occur in their standard form and in their reversed form. Experiment 2 tested whether the reversal occurred because we had, in designing Experiment 1, neutralized focalism, leaving only generalized-group mechanisms as possible causes of a standard above-average effect. Experiment 3 tested whether the a priori groupings of the focal and referent items triggered the reversal of the above-average effect. Having ruled out some seemingly plausible causes of the reversal, we conducted Experiments 4–7 to directly test our new account as to why comparative biases occur in their standard form and under what conditions they would reverse.

According to this account, the timing with which an item is identified as the focal item plays a key role in determining whether the usual differential weighting effects of the focalism and generalized-group accounts will be strong enough to produce standard above- and below-average effects. Furthermore, we suggest that the influence of this timing factor is mediated by the extent of attention to the focal item prior to judgment. Experiment 4 tested whether standard comparative biases would be observed under the conditions specified by this account. Experiment 5 used a process-tracing method to examine whether the late identification of the focal item did indeed lead to greater attention to the focal item just prior to the comparative judgment. Experiment 6 tested whether attention to an item influences comparative judgments as predicted. Finally, Experiment 7 tested with new stimuli whether the timing factor reliably influences the magnitude of the comparative biases.

This set of experiments establishes not only that the timing factor can be a robust influence on the magnitude of above- and below-average effects but also that the timing factor is critical in causing the usual direction of these phenomena to be completely reversed. As we describe in the General Discussion, the conditions for these reversals are ecologically plausible and are therefore quite important to understand. As a whole, these studies provide a broader understanding of when, what type of, and why biases in member-to-group comparisons will occur.

Experiment 1

At the time that we initiated Experiment 1, it was intended largely as a pilot study for our version of the member-to-group paradigm. However, the results of this "pilot study" became the main impetus for the rest of the studies described in this article. Our initial goals for the experiment were threefold.

First, we wanted to establish and test a member-to-group judgment paradigm that could be efficiently executed on computers. The paradigm required that participants make member-to-group judgments based on pictures of hotel rooms, vacation destinations, and sofas.

Second, we wanted to test for above-average effects under conditions in which focalism explanations were made irrelevant. Hence, when participants were asked to make comparative judgments, the question soliciting those judgments did not specify one item as a target and the others as referent items. Instead, participants were simply asked to compare items against each other. Because the focalism accounts—as defined earlier—are relevant to occasions when the wording of a question deems one item as the to-be-judged item, the question wording that we used in this study made focalism irrelevant.

Our third goal was largely unrelated to the main thrust of this article. However, this goal influenced the development of the procedures and materials for this experiment and is therefore important to describe. The goal was to test whether the choices people make when facing a member-to-group decision suggest the same preferences as their comparative judgments suggest. We set up hypothetical-choice scenarios in which participants chose between taking a specified individual item or taking some other item from the group at random.

The experiment involved three stages. In the first stage, participants made one choice regarding each of six sets—attractive hotel rooms, fair hotel rooms, attractive vacation destinations, fair vacation destinations, attractive sofas, and fair sofas. In the next stage, they made one direct comparative judgment regarding each set. In the final stage, they made absolute evaluations of each singleton from each set and of each group from each set (allowing us to assess indirect comparisons).

Method

Stimulus sets. To create the stimulus sets for this experiment, we gathered photographs of hotel rooms, vacation destinations, and sofas that we believed undergraduate students would view as either attractive or merely fair. On the basis of our intuitions and informal screening from undergraduate assistants, we created six sets

(each with five items): attractive hotel rooms, fair hotel rooms, attractive vacation destinations, fair vacation destinations, attractive sofas, and fair sofas. As revealed by participants' absolute judgments (reported below), our intuitions about whether items would be viewed as attractive or fair were generally accurate. Although we constructed sets of fair items, we did not create sets of very unattractive items because part of this experiment involved having people hypothetically chose between free items, and we thought it would be overly contrived to ask people to pick between free unattractive items.

Participants and design. The participants were 40 students from elementary psychology classes at the University of Iowa, who participated to fulfill a research exposure component of the course. For this experiment and the other experiments in the article, set quality (attractive or fair) was manipulated as a within-subject factor. Counterbalancing factors were between subjects.

Procedure. The experiment involved three stages in the following order: the choice stage, comparison-judgment stage, and absolute-judgment stage.

In the choice stage, participants made one choice for each of the six sets, presented in a random order. As an example, Figure 1 shows the computer screen that some participants saw when making a choice regarding attractive sofas. For this sofa set, the participants were told to imagine they were eligible for a free sofa but that they had to choose between Option A or Option B. One of these options specified a single sofa (which we call the *singleton*), and the other option grouped the remaining sofas (the *foursome*) and indicated that one of these sofas would be selected at random. The assignment of an item to the singleton status was fully counterbalanced, as was the assignment of the singleton to the Option A box or to the Option B box. The choice scale contained four response values: *definitely pick A*, *probably pick A*, *probably pick B*, and *definitely pick B*.

In the comparison-judgment stage, participants made a direct-comparison judgment regarding each of the six sets, presented in a random order. Figure 2 shows a relevant example. For each participant, the same item that appeared as a singleton in the choice stage also appeared as the singleton in the comparison-judgment stage. The comparative-judgment question did not specify the singleton as being a target item. For example, for sofas, the question asked, *Regarding your personal preferences for sofas, how do the sofas in the two sets compare to each other overall?* (The term *sets* in that question referred not to the unattractive or fair sets, but to the sets that were concurrently appearing in the separate boxes on the screen—one of which was a singleton and one of which was a foursome.) The anchors on the 7-point scales (which were counterbalanced) also did not designate a target. Instead, they merely stated, *The sofa(s) in A(B) is (are) much more desirable than those (the one) in B(A)* (see Figure 2).

In the absolute-judgment stage, participants made an absolute judgment about each singleton and each foursome for the six sets, again in a random order. For each screen, either the singleton or the foursome from a set appeared, and participants responded to (in the case of sofas) the following: *Regarding your personal preferences for sofas, how would you rate the sofa(s) overall?* The 7-point response scale ranged from *very undesirable* to *very desirable*.

Results

Manipulation check. Before proceeding to the key results, we note that for each of the three categories—hotel rooms, vacation destinations, and sofas—the items that we gathered for use in the attractive sets received generally high desirability ratings on the 7-point absolute scales. Moreover, the average means for these attractive rooms ($M = 5.95$, $SD = 1.04$), destinations ($M = 5.93$,

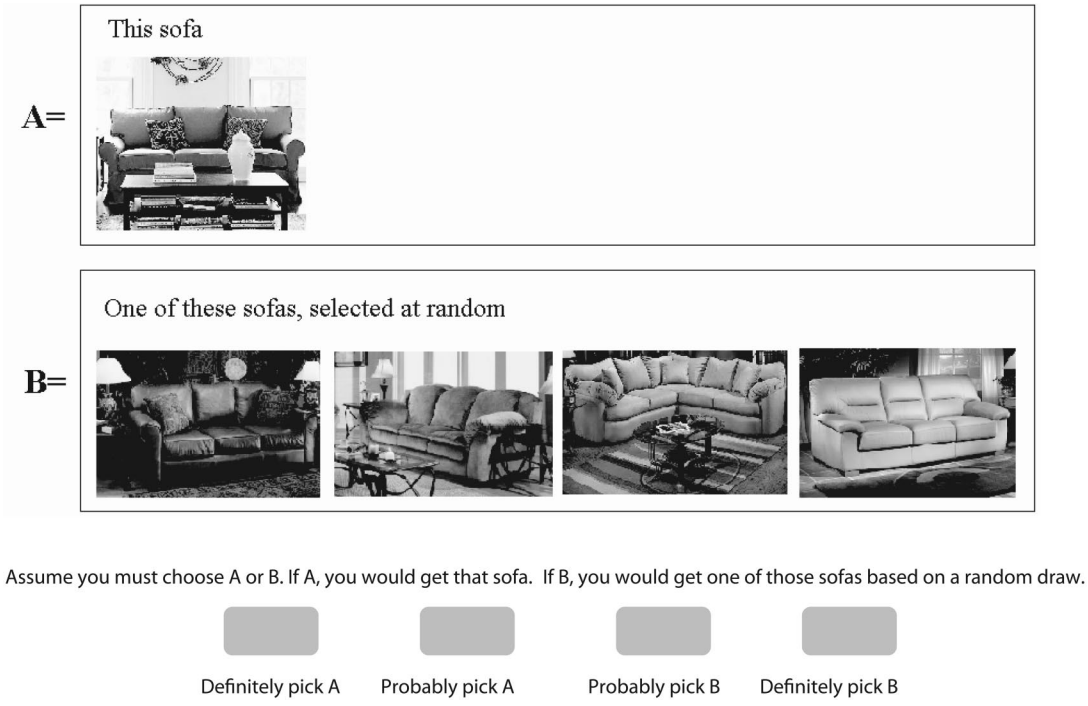

Figure 1. An example of how the set of attractive sofas appeared during the choice stage of Experiment 1.

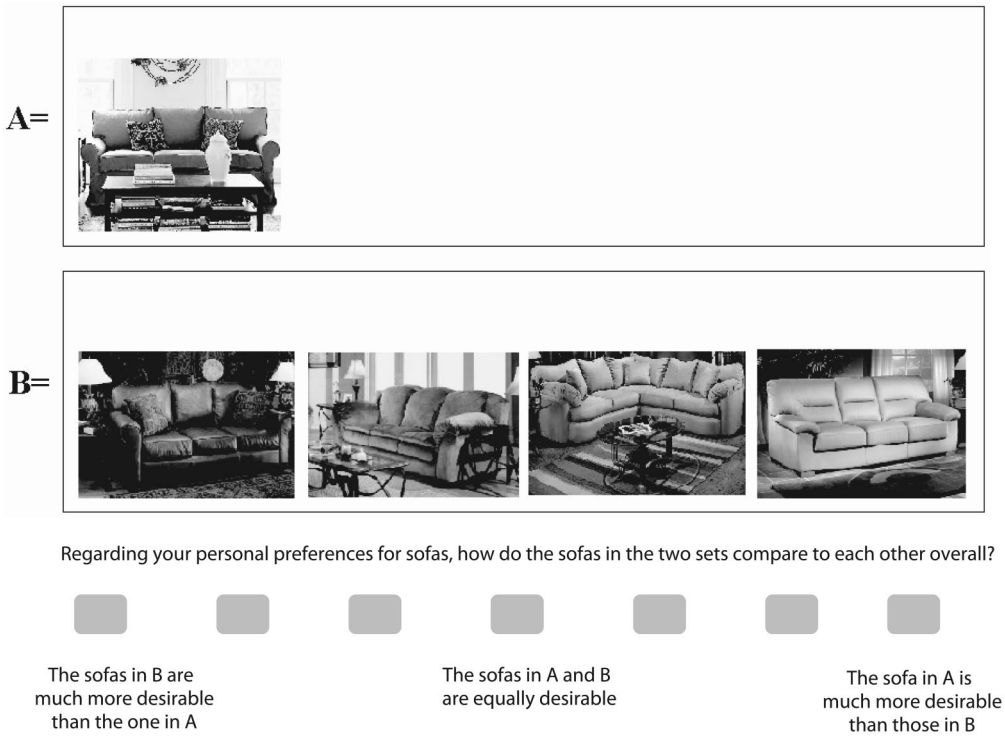

Figure 2. An example of how the set of attractive sofas appeared during the direct-comparison stage of Experiment 1.

SD = 1.07), and sofas (*M* = 5.85, *SD* = 1.12) were significantly higher in pairwise comparisons than were the means for fair rooms (*M* = 2.90, *SD* = 1.30), destinations (*M* = 4.05, *SD* = 1.47), and sofas (*M* = 2.58, *SD* = 1.17; all *p*s < .001). Of these six means, only the mean for fair destinations was not significantly different from the scale midpoint ($\alpha = .05$).

Direct-comparison judgments. The key prediction for the experiment was that the results from the direct-comparison judgments would replicate the above-average effects that have been repeatedly found in previous experiments. That is, we expected that when participants were making comparison judgments regarding attractive sets, they would tend to judge the singleton to be above average—that is, better than the foursome. We also expected that comparison judgments regarding fair sets would show significantly less bias or none at all. To our surprise, however, when participants were making comparison judgments regarding attractive sets, they exhibited the opposite of the above-average effect. That is, they judged the singleton to be worse than average—that is, worse than the foursome. Comparison judgments in the fair groups revealed, as expected, no significant bias.

To conduct the analyses that led to these conclusions, we first coded the direct-comparison judgments from −3 to +3, such that positive values reflected a preference for the singleton over the foursome, and zero reflected indifference. Preliminary analyses that treated category (hotels, sofas, and destinations) and set quality (attractive or fair) as repeated measures indicated that the key results were similar across the three categories—thereby allowing us to simplify our analyses by collapsing across the categories. Table 1 displays means for the comparative judgments and *p* values from *t* tests. As indicated by Table 1, the comparative judgments in the attractive condition were significantly less than zero, $t(39) = 3.66$, $p < .001$, and they were significantly different from those in the fair condition, $t(39) = 2.70$, $p = .01$, which were not significantly different from zero, $t(39) = 0.36$, $p > .10$.

Choices. Did people's choices suggest the same preferences as their direct-comparison judgments? Yes—when making a choice from the attractive sets, participants typically elected to take a random selection from the foursome rather than taking the singleton. On these attractive sets, 85% of participants' responses indicated they would *definitely* or *probably* select the foursome, whereas 15% indicated they would *definitely* or *probably* select the

Table 1
Mean Judgments Regarding Attractive and Fair Sets in Experiment 1

Judgment type	Attractive		Fair		Difference
	M	*SD*	*M*	*SD*	
Direct	−0.54***	0.94	0.04	0.73	$p = .01$
Indirect	−0.65***	0.71	−0.08	0.83	$p < .01$

Note. Direct-comparison judgments were scored from −3 to +3, such that positive values would reflect a preference for the singleton over the foursome. Indirect comparisons are the difference scores between the absolute ratings of the singleton and the foursome; therefore, positive values also reflect a preference for the singleton.
*** $p < .001$.

singleton. When we scored the responses from 1 (*definitely pick the foursome*) to 4 (*definitely pick the singleton*), the mean response across all attractive sets was 1.64 ($SD = 0.58$). This mean was significantly below the indifference point of 2.5, $t(39) = 9.33$, $p < .001$, and it was significantly lower than the relevant mean for the fair sets ($M = 2.47$, $SD = 0.54$), $t(39) = 6.20$, $p < .001$. The mean for the fair sets was not significantly different from the indifference point, $t(39) = 0.29$, $p > .10$, which indicated that people's tendency to "take a chance" on the foursome (i.e., pick the foursome over the singleton) was unique to attractive sets.

Indirect comparisons. The Appendix displays the mean absolute judgments given to singletons and foursomes from the attractive and unattractive sets. As follows from what was discussed in the *Manipulation check* section, participants' absolute ratings for both singletons and foursomes were significantly higher for attractive sets than for unattractive sets. More important, Table 1 displays the indirect-comparison values, which are derived by subtracting people's judgments for foursomes from their judgments for singletons. As indicated in Table 1, these indirect-comparison values are significantly below zero for the attractive sets, and they are significantly different from the values in the fair sets, which are not significantly different from zero. In other words, regarding attractive sets, people gave significantly higher ratings to the foursomes than to the singletons. No such bias was observed regarding the fair sets.

Discussion

On all three dependent measures, people's responses suggested that they preferred the items in the foursome over the singleton. That this pattern was detected on indirect comparisons was not completely unexpected, as some previous studies have produced this pattern on indirect comparisons (e.g., Study 6 of Klar, 2002). There are at least two plausible reasons why the people would tend to give higher ratings to a group of very attractive items than to a similarly attractive singleton. When formulating answers to the absolute question about the foursome (e.g., *Regarding your personal preferences for sofas, how would you rate the sofas overall?*), perhaps people tended to focus most on the best of the attractive foursome, resulting in a higher than warranted impression of that group. Alternatively, perhaps the number of attractive members of the group somehow influenced evaluations of the group as a whole, even though the number of attractive members of the group was not relevant to the absolute question. This possibility is generally consistent with recent research on the group-size effect, which has shown that risk and height judgments for the "average" member of a group tend to increase as a roughly logarithmic function of the size of the group (Price, 2001; Price, Smith, & Lench, 2006; see also Betsch, Kaufmann, Lindow, Plessner, & Hoffmann, 2006). Determining exactly why the attractive foursomes received higher ratings than the singletons lies beyond the scope of this article. More important for the present work is simply the fact that the foursome did indeed receive better ratings than the singleton.

The more surprising element of the results was the fact that—for attractive sets—the direct comparisons also favored the foursome over the singleton. In virtually all studies of member-to-group judgments, the opposite has been true. In those previous studies, evaluations of the singleton had much more weight in determining people's direct comparisons than did evaluations of the group. Hence, when a target item from an attractive group in those studies was judged, it tended to be judged as better than the rest of the group because the main determinant of that judgment was whether the item was generally attractive, not whether the remaining items were generally attractive (see, e.g., Giladi & Klar, 2002).

In the present experiment, it appears that the evaluations of the singleton did not receive inordinate weight in the judgment process. If they had received inordinate weight, direct comparisons would not have produced the same pattern of results as were produced by the indirect comparisons; the direct comparisons would have revealed either a preference for the singleton or at least substantially less of a preference for the foursome (relative to what was observed in indirect comparisons).

The contrast between the results of Experiment 1 and the results of previous studies of member-to-group judgments raises an obvious question: What were the critical differences between our member-to-group paradigm and the member-to-group paradigm used by Giladi & Klar (2002)? The more specific question is why the latter paradigm caused severe differential weighting favoring the singleton, whereas ours did not. Recall that in the introduction we listed two types of accounts that could explain why differential weighting occurs in generic member-to-group comparisons: focalism accounts and generalized-group accounts. In Experiment 1, we purposefully neutralized the focalism account; when participants were asked to make comparative judgments, the question soliciting those judgments did not specify one item as a target item and the others as referent items. We expected this to reduce the degree of differential weighting and above-average effects, but not to eliminate them, because the mechanisms of the generalized-group accounts were still viable. However, perhaps the generalized-group mechanisms were weaker than expected, and the neutralization of focalism had a more severe influence than we had expected, resulting in an elimination of the weighting bias that favored the singletons in previous studies (thereby allowing the direction of the above-average effect to flip in Experiment 1). We addressed this possibility in Experiment 2 by systematically manipulating whether focalism was neutralized.

Experiment 2

There were three main goals for Experiment 2, which used the same materials and general procedures as Experiment 1. First, given the surprise and importance of finding a complete reversal of the usual above- and below-average effects in member-to-group comparisons, we wanted to see if these findings would replicate. Second, we sought to explicitly test the role of focalism by manipulating whether the questions soliciting the direct-comparison judgments specified a target or whether they were completely neutral or balanced. Third, we wanted to rule out an explanation for Experiment 1 that we thought was possible but unlikely. Namely, participants in Experiment 1 always completed the choice measures before moving on to the direct-comparison questions. It seemed possible that some aspect of the choice task or the participants' choices themselves might have had a carryover effect on their direct-comparison judgments. For example, if they chose the foursome option for the sofas, perhaps that selection influenced how they later judged the sofas—in a direction that supported their choice. To remove this possibility in Experiment 2,

we moved the choice measures to the last stage of the experiment. This ensured that, at the point participants were making comparison judgments, they did not anticipate that they would be making any choices among the items in the sets.

Method

Participants and design. The participants were 40 students from elementary psychology classes at the University of Iowa, who participated to fulfill a research exposure component of the course. Set quality (attractive or fair), question wording (balanced wording or focal wording), and counterbalancing factors were manipulated.

Materials and procedure. The materials and procedures were the same as those from Experiment 1, with the following two exceptions. First, of the three stages of the experiment (comparison judgment, absolute judgment, and choice), the choice stage was always last. Half of the participants completed the comparison-judgment stage before the absolute-judgment stage, and half completed those stages in the reverse order. This ordering manipulation did not produce significant main effects or interactions on any of the key dependent variables and therefore will not be discussed further. Second, participants were randomly assigned to either a balanced-wording condition or a focal-wording condition. The balanced-wording condition was a straight replication of Experiment 1. In the focal-wording condition, the direct-comparison questions and scale anchors specified a target (i.e., focal item) and a referent. For example, a question that specified the *A* sofa as the focal item was *Regarding your personal preferences for sofas, how does the sofa in A compare with the sofas in B?* The left, middle, and right anchors on the 7-point scale for this question read, *A is much less desirable than those in B*, *A is equally desirable to those in B*, and *A is much more desirable than those in B*. Whether the focal item appeared in Box A or Box B was counterbalanced and this determined whether the *A* or *B* was in the focal position for the question and scale anchors.

Results

Direct-comparison judgments. As was the case for Experiment 1, we coded the direct-comparison judgments from −3 to + 3, such that positive values reflected a preference for the singleton over the foursome. Table 2 displays means for the comparison judgments and *p* values from relevant *t* tests. The comparison judgments from the balanced-wording condition replicated the results from Experiment 1. That is, the judgments in the attractive condition were significantly less than zero, indicating a preference for the foursome. These judgments were also significantly different from those in the fair condition, which were not significantly different from zero (see Table 2 for relevant *p* values).

A similar but slightly weaker pattern of results was found for the focal-wording condition. The judgments in the attractive condition were again significantly less than zero, indicating a preference for the foursome. However, these judgments did not significantly differ from those in the fair condition. A mixed-model ANOVA including wording (balanced vs. focal) and set quality (attractive vs. fair) as factors revealed a main effect for set quality, $F(1, 38) = 8.34$, $p < .01$, but nonsignificant effects for the wording factor, $F(1, 38) = 1.86$, $p = .18$, and for the interaction, $F(1, 38) = 1.02$, $p = .32$.

Table 2
Mean Judgments Regarding Attractive and Fair Sets by Wording Condition in Experiment 2

Wording condition and judgment type	Attractive		Fair		Difference
	M	*SD*	*M*	*SD*	
Balanced					
Direct	−1.02***	1.14	−0.05	0.84	$p < .05$
Indirect	−0.78***	0.84	−0.05	0.61	$p < .01$
Focal					
Direct	−0.52*	1.12	−0.05	0.74	$p = .18$
Indirect	−0.45**	0.58	−0.26	0.71	$p = .29$

Note. All direct-comparison judgments were scored from −3 to +3, such that positive values would reflect a preference for the singleton/target over the foursome/referents. Indirect comparisons are the difference scores between the absolute ratings of the singleton and the foursome; therefore, positive values also reflect a preference for the singleton.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Indirect comparisons. Table 2 also displays the mean indirect-comparison values derived from participants' absolute judgments. The results from the indirect-comparison values generally replicate those of Experiment 1—suggesting a preference for the foursome over the singleton among attractive sets but not among fair sets (see Table 2 for *p* values). A mixed-model ANOVA on the indirect comparison values revealed a significant main effect for set quality, $F(1, 38) = 9.09$, $p < .01$, but nonsignificant effects for the wording factor ($F < 1$) and for the interaction, $F(1, 38) = 3.27$, $p = .08$. These results of the ANOVA confirm that the preference for the foursome over the singleton was significantly stronger among attractive sets than among fair sets.

Choices. Did people's choices suggest the same preferences as their direct- and indirect-comparison judgments? Yes—when making a choice from the attractive sets, participants indicated that they would *definitely* or *probably* select the foursome at a rate of 79%. The comparable rate for fair sets was only 54%. We also scored the responses from 1 (*definitely pick the foursome*) to 4 (*definitely pick the singleton*) and submitted the composite values to a Wording × Set Quality ANOVA. The main effect for set quality was significant, $F(1, 38) = 13.04$, $p < .001$, but the main effect for wording and the effect for the interaction were not significant (Fs < 1). The mean response across all attractive sets ($M = 1.78$, $SD = 0.70$) was significantly lower than the indifference point of 2.5, $t(39) = 6.49$, $p < .001$, but the mean response across all fair sets ($M = 2.42$, $SD = 0.77$) was not different from the indifference point. To summarize the findings of the choice measures, participants tended to choose the foursome over the singleton, but only regarding attractive sets, which is the same pattern observed in Experiment 1.

Discussion

The key finding from Experiment 2 was that the reversal of the standard above-average effects in member-to-group comparisons was replicated. That is, in direct-comparison judgments regarding attractive sets, participants again preferred the foursome over the singleton. The fact that this effect was replicated, even when choice measures followed the comparison judgments, rules out

accounts involving carryover effects. The effect was also found even when the direct-comparison question specified the singleton as a target item and the foursome as referent items. Hence, focalism alone (i.e., the notion that the wording of a comparison question can cause differential weighting favoring the item specified as the target) was clearly not enough to produce standard above-average effects in our member-to-group paradigm.

Experiment 3

Upon failing to produce standard above-average effects in the focal-wording conditions of Experiment 2 (and in fact replicating a reversed effect), we began to suspect that the key difference between our paradigm and that of Giladi and Klar (2002) concerned the way in which the items were grouped for presentation to the participants. In previous experiments on member-to-group comparisons, the focal item was clearly part of a larger set. Participants in those experiments knew that the target item was "singled out" merely because it was randomly selected from the set. However, in our Experiments 1 and 2, the target item/singleton was not depicted as part of the set. The first time the singleton appeared, it was already separate from the foursome. Hence, whereas previous research required participants to make member-to-group comparisons, our procedures may have essentially required participants to make set-to-set comparisons, with one set having an *N* of 1 and the other having an *N* of 4. Perhaps our framing the comparison as being between two sets made it easier for people to avoid differentially weighting the singleton and the foursome in their comparative judgment. More specifically, perhaps the generalized-group account for above-average effects was effectively made irrelevant in our version of the paradigm because the foursome could be easily viewed as a distinct and well-defined set that was as important as the other "set" (i.e., the singleton). If the task was framed more as a member-to-group comparison than as a set-to-set comparison, perhaps robust differential weighting would occur, thereby yielding the standard above-average effects. We tested this possibility by slightly modifying our paradigm in Experiment 3. Namely, when the items from a set were displayed, all the items within the set appeared in one row, suggesting a common a priori grouping. Rather than being located apart from the other items (as in Experiments 1–2), the focal item was now within the row and simply marked by an *X*.

Method

Participants and design. The participants were 40 students from elementary psychology classes at the University of Iowa, who participated to fulfill a research exposure component of the course. Set quality (attractive or fair) and counterbalancing factors were manipulated.

Materials and procedure. The materials and procedures were the same as those from Experiment 2, with the following exceptions. There were two stages of the experiment, with the comparison-judgment stage always preceding the absolute-judgment stage. In the comparison-judgment stage, the computer displayed the five items from a group in a randomly ordered row with one item (counterbalanced across participants) marked with an *X* above it (see Figure 3, for an example). The comparison question treated the marked item as a target item to be judged. For example, the comparison question for the sofa sets was as follows: *Regarding your personal preferences for sofas, how desirable is the sofa marked with an 'X' compared to the other sofas in the group?* The left, middle, and right anchors on the 7-point scale read, *It is much less desirable than the rest*, *It is equally desirable to the rest*, and *It is much more desirable than the rest*.

Results

Direct-comparison judgments. The modified methodology produced largely the same patterns as those observed in Experiments 1 and 2, albeit somewhat weaker ones. The mean direct-comparison judgment regarding the focal item from attractive sets ($M = -0.35$, $SD = 0.82$) was significantly below zero, which indicates a preference for the group of referents, $t(39) = 2.68$, $p = .01$. The mean direct-comparison judgment regarding the focal item from fair sets ($M = -0.22$, $SD = 0.85$) was not significantly different from the mean for attractive sets, $t(39) = 0.70$, $p = .49$, and it was not significantly below zero, $t(39) = 1.57$, $p = .13$.

Indirect comparisons. The mean values for the indirect comparisons regarding the attractive sets ($M = -0.74$, $SD = 0.79$)

Figure 3. An example of how the set of attractive sofas appeared during the direct-comparison stage of Experiment 3.

were also significantly below zero, indicating a preference for the group of referents, $t(39) = 5.86$, $p < .001$. The mean values for the fair sets ($M = -0.22$, $SD = 0.82$) were not significantly below zero, $t(39) = 1.69$, $p = .10$.

Discussion

Our framing of the participants' task as a member-to-group comparison rather than a set-to-set comparison continued to yield a reversal of the standard above-average effects. This finding does not bode well for the use of a generalized-group account to explain why previous member-to-group paradigms (e.g., Giladi & Klar, 2002) have produced standard above-average effects, whereas the present paradigm has produced the reverse. A generalized-group account would appear to make the same predictions for the two paradigms, but the results clearly differ between the paradigms. The same can be said for the focalism account.

Experiment 4

Experiments 2 and 3 ruled out carryover effects, focal wording, and the a priori grouping of items as critical reasons for differences in the comparative biases found in our paradigm versus that of Giladi and Klar (2002). Having narrowed the list of plausible reasons for the differences, we hypothesized that the timing with which a person learns which item is focal has a critical influence on the person's attention to the items and ultimately on the magnitude and/or direction of comparative bias. In the standard paradigm used by Giladi and Klar (2002), the randomly selected target item is denoted late—that is, only after participants have already inspected the full set of items. For example, in their study involving soaps, participants inspected each of the six soaps before selecting a slip of paper to determine which soap (the target soap) was to be inspected again before rating it. In our paradigm, the target item is denoted early—that is, from the first time the items are seen. Even in Experiment 3, the target item was denoted with an *X* when the items first appeared.

We hypothesized that this timing factor has a systematic influence on the order in which information about the various items is processed by respondents. In a late-denotation paradigm (e.g., the soap paradigm), the participants would first process each item without knowledge of which will be the target. Then, after the target item is denoted, the participants would focus primarily on that item, given their new knowledge that this is the item that has been singled out for the comparison judgment. Whereas participants might briefly consider the referents, we suspect that such considerations of the referents would be rather cursory because the participant had already inspected each one. In short, we suspect that because the target item is revealed late, it enjoys somewhat of a recency effect (in terms of heightened attention or weight) when people make their comparative judgment (see Houston, Sherman, & Baker, 1989). Regarding the early denotation paradigm, such as in our Experiments 1–3, we suspect the recency effects to be less strong or relevant. The processing of the target item might not typically be reserved for last (i.e., right before judgment). Instead, the target item might be processed early or perhaps within a "standard" sequence (e.g., it might be the third item that is processed when it appears in the third from-the-left location on the screen). Hence, if the weighting of the target and the referents is based on attention (or recency of attention) at the time of the comparison judgment (see Taylor & Fiske, 1975), then the weighting of the target and referents might be more balanced in an early denotation paradigm than in a late-denotation paradigm. In the latter paradigm, the delayed revelation of the target triggers processes that result in enough differential weighting favoring the target to yield the standard above-average effects.

We tested this account in the next several experiments. In Experiment 4, we modified the way in which the target item was revealed. Instead of denoting the target item immediately, the target item was denoted after a short delay (12 s after the items appeared but before the comparative judgment was posed). If the timing factor does have a critical influence on differential weighting, then we should detect above-average effects in the standard direction in this experiment.

In Experiment 4, we also tested the role of another possible factor as to why previous paradigms have produced standard above-average effects whereas Experiments 1–3 did not. This factor concerned the fact that in Experiments 1–3, the target and the referents all remained in a prominently visible location while the participants made their comparative judgments. Alternatively, in other member-to-group paradigms, the referent items are not displayed quite so prominently. For example, in the soap paradigm (Giladi & Klar, 2002), when the target soap is selected, it is lifted by the participant from a table while the referent soaps remain on the table. Results from one member-to-group study suggest that the visual prominence of the referents does not substantially diminish the standard above-average effects (Klar, 2002). However, it did seem intuitively plausible that our failure to replicate a standard above-average effect might have been due, in part, to the fact that our target and referent stimuli in Experiments 1–3 were always given equal prominence on a computer screen (see also Suls et al., 2007). To test this possibility in Experiment 4, we manipulated whether the referent items stayed on screen for the entire judgment phase or whether the referent items faded off the screen after participants had a chance to process them.

Method

Participants and design. The participants were 87 students from elementary psychology classes at the University of Iowa, who participated to fulfill a research exposure component of the course. Set quality (attractive or fair), referent status (fading or nonfading referents), and counterbalancing factors were manipulated.

Materials and procedure. The materials, dependent measures, and procedures were the same as those from Experiment 3 with the following exceptions. In the direct-comparison stage, an introductory screen for each set informed participants that they would be seeing several items and that they should study the items carefully. The items from a set (none marked with an *X*) appeared on screen for 12 s before a *Continue* button appeared along with the following instructions: *Please feel free to continue looking at the items. When you are ready to start the random selection process, click the button below*. When the *Continue* button was clicked, an *X* appeared above the target. In a fading-referent condition, the images of the referents (but not the target) disappeared as the target was denoted. In a nonfading-referent condition, the images of the referents (and target) remained unchanged as the target was de-

noted. After a second *Continue* button was clicked, the direct-comparison question appeared. After all direct-comparison judgments were made, participants then completed the absolute-judgment stage of the experiment.

Results

Direct-comparison judgments. Table 3 displays cell means and *p* values from relevant *t* tests. Unlike Experiments 1–3, the direct-comparison judgments in Experiment 4 produced a significant above-average effect in the standard direction. The mean direct-comparison judgment regarding the focal item from attractive sets ($M = 0.22$, $SD = 0.68$) was significantly above zero, indicating a preference for the target over the referents, $t(86) = 3.07$, $p < .01$. In addition, the mean from fair sets ($M = -0.27$, $SD = 0.75$) was significantly below zero, which indicates a standard below-average effect, $t(86) = 3.33$, $p < .01$. A mixed-model ANOVA including referent status (fading vs. nonfading) and set quality (attractive vs. fair) as factors revealed a main effect for set quality, $F(1, 85) = 19.40$, $p < .001$, but revealed nonsignificant effects for the referent status and for the interaction (Fs < 1).

Indirect comparisons. As expected, the results of the indirect measures are quite consistent with those of Experiments 1–3. This, of course, means that they are generally inconsistent with the results of the direct-comparison measures (see Table 3). The mean values for the indirect comparisons regarding the attractive sets ($M = -0.69$, $SD = 0.68$) were significantly below zero, indicating a preference for the group of referents, $t(87) = 9.48$, $p < .001$. The mean values for the fair sets ($M = -0.09$, $SD = 0.73$) were not significantly below zero, $t(87) = 1.23$, $p > .10$. A mixed-model ANOVA, including referent status and set quality as factors, revealed a main effect for set quality, $F(1, 85) = 29.64$, $p < .001$. The interaction was not significant ($F < 1$). The main effect for referent status was significant, but we interpret this as spurious because that manipulation was located in the comparison judgment stage of the experiment, $F(1, 85) = 4.32$, $p = .04$.

Discussion

It appears that the timing change introduced in Experiment 4—denoting the target after a delay—was enough to produce standard above-average effects on direct-comparison measures rather than the reversed above-average effects that were observed in Experiments 1–3. In fact, the timing change even produced a standard below-average effect for fair sets.[1] These effects were roughly equivalent in magnitude and were not significantly moderated by whether the referent items disappeared or stayed on screen when the target was denoted.

The flip in the direction of the effects is consistent with our hypothesis that the weighting of a target and referents in a direct-comparison judgment can be biased by where the respondent's attention is focused immediately prior to making a comparative judgment. In Experiments 1–3, the bulk of the target processing likely preceded or was intermixed with the processing of the referents. In Experiment 4, however, participants were not aware of which item was the target until they had already inspected all the items. Therefore, the target item likely drew considerable attention and processing immediately before the comparison-judgment was made. The resulting differential weighting was strong enough to overcome the fact that in absolute judgments regarding attractive sets, participants' ratings of the referents were more positive than their ratings of the targets. That is, the direct-comparison judgments regarding attractive targets tended to be "above average" not because targets looked better than referents when judged separately (quite the contrary), but because when an attractive target was judged "against" attractive referents, the attractiveness of the target drove the judgment more so than did the attractiveness of the referents.

Table 3
Mean Judgments Regarding Attractive and Fair Sets by Nonfading or Fading Referent Condition in Experiment 4

Nonfading or fading condition and judgment type	Attractive		Fair		
	M	*SD*	*M*	*SD*	Difference
Nonfading					
Direct	0.19	0.75	-0.27^{*}	0.71	$p < .01$
Indirect	-0.58^{***}	0.65	-0.21	0.75	$p < .05$
Fading					
Direct	0.26^{**}	0.60	-0.27^{*}	0.80	$p < .01$
Indirect	-0.81^{***}	0.70	0.02	0.69	$p < .001$

Note. All direct-comparison judgments were scored from −3 to +3, such that positive values would reflect a preference for the singleton/target over the foursome/referents. Indirect comparisons are the difference scores between the absolute ratings of the singleton and the foursome; therefore, positive values also reflect a preference for the singleton.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Experiment 5

As just noted, the results of Experiment 4 are consistent with our proposal that the timing of when the target is revealed influences the degree of attention directed to the target (immediately prior to the comparative judgment), which thereby influences the weight that the target item receives and the direction/magnitude of the comparative bias. To provide a more direct test of this proposed causal chain, we conducted a set of two experiments (for discussion of experimental-causal-chain designs, see Spencer, Zanna, & Fong, 2005). Experiment 5 was designed to directly test whether the timing of when the target is revealed does, in fact, influence respondents' degree of attention to the target immediately prior to making the comparative judgment. Experiment 6 was designed to directly test whether forcing participants to attend to the target immediately prior to a comparative judgment does, in fact, influence the magnitude of the comparative bias.

For Experiment 5, we used a process-tracing method to directly compare the order in which people processed the target and referents in two separate conditions—one in which the denotation of the target was early (as in Experiments 1–3) and one in which the

[1] Upon closer inspection, the standard below-average effects were found for the fair sofa set and the fair hotel set. As reported in the manipulation-check section of Experiment 1, the stimuli for these two sets were generally perceived as undesirable (rated below the midpoint on scales ranging from *very undesirable* to *very desirable*). Therefore, it is not surprising that the comparative ratings from these two sets in Experiment 4 account for the standard below-average effects, whereas the below-average effect did not exist for the fair destinations, which were not generally perceived as undesirable.

denotation of the target was delayed (as in Experiment 4). The stimuli in this experiment were identical to those of the previous experiments except that each item on a given screen was occluded by a box until the participant clicked that box and gained a view of the item (only until the next box was clicked). The computer recorded the location of the clicks, thereby providing a tracing of what items the participants viewed and in what order.

Method

Participants and design. The participants were 82 students from elementary psychology classes at the University of Iowa, who participated to fulfill a research exposure component of the course. In addition to manipulating set quality (attractive or fair), we also manipulated the timing of the denotation of the focal item (early or delayed). The latter manipulation was between subjects.

Materials and procedure. The materials, dependent measures, and procedures were similar to those from the nonfading referent condition of Experiment 4, except as noted in this section. The critical difference involved the presence of boxes and buttons that needed to be clicked by participants to view items and to proceed through the computer program. Each item on a given screen was occluded by a box until the participant clicked that box and gained a view of the item (only until the next box was clicked).

In the early condition, for each set, participants saw a row of labeled boxes (e.g., *Sofa 1*, *Sofa 2*), one of which had an *X* above it. Participants were told that they could view the items as many times as they wanted before clicking on a button that read, *When you are done viewing [items] and are ready to make your judgment, click here*. In the delayed condition, for each set, participants saw a row of labeled boxes (e.g., *Sofa 1*, *Sofa 2*), none of which had an *X* above it. Participants were told that they could view the items as many times as they wanted before clicking on a button that read, *When ready for the random selection process, click here*. After this button was pressed, an *X* marked the target item, and participants were free to view more boxes until they clicked on a button that read, *When you are done viewing [items] and are ready to make your judgment, click here*.

In both conditions, only direct-comparison judgments were solicited, and the questions and scales were identical to those used in Experiment 3 and 4. However, to simplify the computer programming of the tracing methodology, the items that were denoted as targets were the same for all participants (not fully counterbalanced as in the previous experiments). The position of the target differed across the six sets of items (with the second position used twice).

Results

Process tracing findings. Table 4 contains a variety of statistics that help to provide an overall picture of the numbers and types of items that people viewed. However, the prediction being tested in this experiment was rather precise. We predicted that the delay in the denotation of the focal item would lead to an increase in the attention that the focal item would receive immediately before participants made their comparative judgments. Therefore, for each participant, we calculated the proportion of times that the focal item was the last item they viewed before clicking the button to start the rating task. This was done separately for attractive sets and for fair sets. These proportions were then submitted to a mixed-model ANOVA with timing and set quality as factors. As expected, the main effect for timing was significant, $F(1, 80) = 12.77$, $p < .001$, and the other main effect and interaction were not significant (Fs < 1). Collapsing across set quality, the focal item was the last item viewed at a rate of 62.7% in the delayed condition but only at 43.3% in the early condition. Hence, the delay did result in a greater likelihood that the focal item would be processed just prior to the comparative judgment.

Direct-comparison judgments. Recall that in this experiment, we did not fully counterbalance the assignment of items to the

Table 4
Process Tracing Statistics for the Early and Delayed Conditions of Experiment 5

	Early		Delayed		
Index	*M*	*SD*	*M*	*SD*	Difference
Number of focal boxes opened per set	1.73	0.51	2.33	0.51	$p < .001$
Number of boxes opened per set	6.86	1.49	8.46	2.22	$p < .001$
% of opened boxes that were focal	24.84	4.29	28.12	3.62	$p < .001$
Number of focal boxes opened per set (before target was denoted)			1.32	0.36	
Number of boxes opened per set (before target was denoted)			6.51	1.55	
% of opened boxes that were focal (before target was denoted)			20.22	1.80	
Number of focal boxes opened per set (after target was denoted)			1.00	0.40	
Number of boxes opened per set (after target was denoted)			1.97	1.40	
% of opened boxes that were focal (after target was denoted)			65.17[a]	26.21	
% of times that the first opened box was focal	31.67	30.62	18.25	7.18	$p < .01$
% of times that the last opened box was focal	43.3	23.5	62.7	25.5	$p < .001$

Note. For all means in this table, we first calculated the relevant value for each subject, then averaged across subjects. In preliminary analyses of all the above variables, the quality of a set did not significantly interact with the timing manipulation. Therefore, to simplify this table, we collapsed across attractive and fair sets. Repeat visits to a box (nonconsecutive) were included in the count. For example, if a participant opened Box 1, then Box 4, then Box 1, the return visit to Box 1 was included in our count.

[a] The value of 65.17% may seem somewhat inconsistent with the two values above it, but it is not calculated from those values. It is the mean of the proportions calculated per participant. The apparent discrepancy is due to the fact that participants opened different numbers of boxes, and those who opened the most boxes tended to open smaller proportions of focal boxes. The median of the percentage of focal boxes that were opened by participants (postselection) was 60%.

focal "position." Instead, we randomly selected a focal item for each set, and the selected item was seen as the focal item by all participants.

Therefore, main-effect tests of whether ratings of the focal item fell above, at, or below the midpoint of the direct-comparison scale are not very consequential because deviations from zero may simply reflect the random qualities of our small number of selected focal items—not a systematic judgment bias. Nevertheless, it is somewhat instructive to test for the interaction between timing and set quality. As expected, the interaction was significant, $F(1, 80) = 10.22$, $p < .01$. Specifically, the difference in mean (1.19) between the focal ratings in the attractive sets ($M = 4.44$, $SD = 0.68$) and the fair sets ($M = 3.25$, $SD = 0.81$) was greater in the delayed condition than the difference (0.51) between the focal ratings in the attractive sets ($M = 4.09$, $SD = 0.87$) and in the fair sets ($M = 3.58$, $SD = 0.53$) in the early condition. This pattern is consistent with the idea that a delayed denotation of the focal item leads to disproportional weighting of that item (and a greater chance of standard above- and below-average effects).

Experiment 6

Whereas Experiment 5 tested the first component of our proposed causal chain (that a delayed revelation of the target enhances attention to the target immediately prior to the comparative judgment), Experiment 6 tested the second component (that this enhanced attention to the target influences the direction/magnitude of comparative bias). To test this second component, we directly manipulated the sequence with which participants saw the items in a set, prior to making their comparative judgment. Participants in the target-last group saw the full set of items and then just the target. Participants in the referents-last group saw the full set of items, then just the target, and then just the referents. Both groups then made a comparative judgment about the target relative to the referents. We predicted that participants in the former group would exhibit standard comparative biases (same as Experiment 4). More important, we expected that these biases would be significantly weaker (or would flip) in the latter condition, because participants in that condition were essentially forced to attend to the referents—not to the target—immediately prior to making their comparative judgment.

Method

Participants and design. The participants were 90 students from the University of Iowa, who were fulfilling a research exposure component of the elementary psychology course. Set quality (attractive or fair) and presentation order (target last or referents last) were manipulated, with the latter as a between-subjects factor.

Materials and procedure. The pictorial stimuli were identical to those used in Experiments 4. The procedures differed slightly as a function of the presentation order manipulation.

In the target-last condition, the procedures were similar to those of the fading referent condition from Experiment 4. For each set, participants first saw the full set of items for at least 12 s, at which point a *Continue* button also appeared. After clicking the *Continue* button, the target was marked with an *X* and remained on screen as the referents faded away. After 3 s, another *Continue* button appeared. Clicking this *Continue* button triggered the appearance of the comparative question (with the target remaining on screen). The comparative question asked participants to rate how desirable the item marked with an *X* was compared with the other items in the group.

In the referents-last condition, the procedures were the same except that after the participants viewed the target, they clicked a *Continue* button that resulted in the removal of the target and the reappearance of the referents. After 3 s, another *Continue* button appeared. Clicking this *Continue* button triggered the appearance of the comparative question (with the referents remaining on screen). The comparative question asked participants to rate how desirable the item that had been marked with an *X* was compared with the other items in the group. After all direct-comparison judgments were made, participants then completed the absolute-judgment stage of the experiment.

Results

Direct-comparison judgments. Table 5 displays cell means and *p* values from relevant *t* tests. As expected, within the target-last condition, we found a significant above-average effect for attractive sets in the standard direction. The below-average effect for fair sets was also in the standard direction but was not quite significant ($p = .11$). More important, these comparative-bias effects were essentially absent in the referents-last condition. Between-groups *t* tests revealed that the above-average effect was significantly smaller when the referents rather than the targets were viewed last, $t(88) = 2.29$, $p < .05$. The same pattern was true for below-average effects, but the trend was not significant. A mixed-model ANOVA including set quality (attractive vs. fair) and order (target-last vs. referents-last) revealed, as expected, a significant interaction, $F(1, 88) = 4.42$, $p < .05$. In short, these results confirm that the standard comparative biases were reduced in magnitude (albeit not reversed) when the referents, rather than the targets, were viewed last.

Indirect comparisons. The results of the indirect measures were generally consistent with those of Experiments 1–4 (see Table 5). One unexpected result was a significant below-average effect for the fair sets. However, this finding does not seem consequential for our main conclusions from this experiment.

Table 5
Mean Judgments Regarding Attractive and Fair Sets in Experiment 6 by Target-Last or Referents-Last Condition

Condition and judgment type	Attractive		Fair		
	M	*SD*	*M*	*SD*	Difference
Target-last					
Direct	0.38**	0.93	−0.23	0.94	$p < .01$
Indirect	−0.66***	0.72	−0.16	0.75	$p < .01$
Referents-last					
Direct	−0.04	0.82	−0.06	0.87	$p = .95$
Indirect	−0.36**	0.87	−0.23*	0.61	$p = .42$

Note. All direct-comparison judgments were scored from −3 to +3, such that positive values would reflect a preference for the singleton/target over the foursome/referents. Indirect comparisons are the difference scores between the absolute ratings of the singleton and the foursome.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Discussion

The results of Experiment 6 clearly demonstrate that drawing people's attention to the target rather than to referents immediately prior to their making a comparison judgment will impact the magnitude of comparative bias. These findings are generally consistent with related research showing that the order in which people encounter multidimensional stimuli from a given category will influence their evaluative judgments about those stimuli (Houston et al., 1989; see also Bruine de Bruin & Keren, 2003). Coupled together, Experiments 5 and 6 provide strong evidence for the causal chain we proposed to explain the role of timing in comparative bias. Namely, the timing of when the target is revealed influences the degree of attention directed to the target (immediately prior to the comparative judgment), which thereby influences the magnitude of the comparative bias.

Experiment 7

The pattern of results across Experiments 1–6 reveals critical information about the role of timing in whether a comparative bias will take its standard form (observed in Experiment 4 and within the target-last condition of Experiment 6) or whether it will take a reversed form (observed in Experiments 1–3). However, all of these experiments have involved the same six sets of stimuli. Although this consistent use of stimuli makes cross-experiment comparisons more informative, it also places a limit on conclusions as to whether timing would be a significant moderator of comparative biases observed with other sets of stimuli. Therefore, in Experiment 7, we used new stimuli (in addition to our old stimuli) to again test the moderating role of timing. For the new stimuli, we used three pairs of sets that contained items that were social in nature and two pairs that contained items that were nonsocial (e.g., houses). We also diversified the dimensions being judged. Whereas in Experiments 1–5, the dimension being judged was always the desirability of the items, desirability was the relevant dimension for only the nonsocial sets in Experiment 7. The relevant dimensions for the three social pairs of sets were athleticism, morality, and prestige.

Another important feature of Experiment 7 was that we manipulated the timing factor (whether the target was identified early or after a delay) within the experiment itself. This timing factor differed between Experiment 4 and our earlier experiments, but it was not experimentally manipulated within Experiment 4.

Finally, Experiment 7 provided one more test of the possibility that standard above-average effects would be significantly stronger if the referents disappeared rather than remained when the target was specified. Participants were randomly assigned to one of three cells—one in which target items were denoted early, one in which the denotation was delayed, and one in which the delayed denotation was also accompanied by the disappearance of the referents.

Method

Stimulus sets. We gathered five photographs for each of the 10 new stimulus sets. The photographs for the 5 "superior" sets contained items (i.e., objects or people) that are generally perceived as high on the relevant dimension. The photographs for the 5 "inferior" sets contained items that are generally perceived as low on the relevant dimension. The resulting 10 sets were as follows: athletic or unathletic individuals, people with spotless or spotted moral reputations, high- or low-prestige occupations, mansions or dilapidated houses, and new sports cars or old clunker cars. For example, the people displayed in the athletic group included Lance Armstrong and Michael Jordan, whereas the unathletic group included Dick Cheney and Woody Allen. The people displayed in the spotless morals group included Nelson Mandela and Mother Teresa, whereas the spotted morals group included Kenneth Lay and Richard Nixon. In addition to using the 10 new stimulus sets, we also had participants respond to the 6 "old" sets that were used in Experiments 1–6.

Participants and design. The participants were 206 students from elementary psychology classes at the University of Iowa. Set quality (superior or inferior), presentation procedure (early/nonfading, delayed/nonfading, or delayed/fading), and counterbalancing factors were manipulated.

Dependent measures. The dependent measures and scale formats were identical to those used in Experiments 3 and 4, with the following exception: Whereas the direct-comparison questions about the sets of houses, cars, hotels, sofas, and vacation destinations continued to ask about desirability (e.g., *How desirable is the car marked with an X compared with the other cars in the group?*), the questions for the other sets asked about other dimensions (*How athletic is the person. . .? How moral is the person. . .? How prestigious is the occupation. . .?*). The absolute questions contained analogous changes.

Procedure. All participants made direct-comparison judgments regarding the target for each of the 16 sets in random order. An introductory screen for each set informed participants that they would be seeing several items and that they should study the items carefully. For participants in the early/nonfading cell of the design, the target was marked with an *X* when the items from a set first appeared. After 12 s, a *Continue* button appeared, which, when clicked, resulted in the display of the comparison question (below the items). The referents did not fade away when the *Continue* button was clicked. For participants in the delayed/fading and delayed/nonfading cells of the design, the target was not marked with an *X* when the items from a set first appeared. Just as in Experiment 4, a *Continue* button and instructions appeared after 12 s. When the button was clicked, an *X* appeared above the target. In the delayed/nonfading cell, the referents remained unchanged as the *X* appeared. In the delayed/fading cell, the referents disappeared from the screen as the *X* appeared. After a second *Continue* button was clicked, the direct-comparison question appeared. The procedures for the absolute judgments, which were made last, were identical to those of Experiment 4.

Results

Direct-comparison judgments. Preliminary analyses indicated that the old and new stimuli did not yield significantly different patterns of results, so we report overall results (i.e., including old and new sets). Table 6 displays cell means and *p* values from relevant *t* tests. A mixed-model ANOVA including set quality (superior, inferior) and presentation procedure (early/nonfading, delayed/nonfading, delayed/fading) as factors revealed a main effect for set quality, $F(1, 203) = 33.89$, $p < .001$, and a nonsignificant main effect for presentation procedure, $F(1, 203) = 1.09$,

Table 6
Mean Judgments Regarding Superior and Inferior Sets by the Between-Subjects Conditions of Experiment 7

Condition and judgment type	Superior		Inferior		
	M	*SD*	*M*	*SD*	Difference
Early/nonfading					
Direct	0.03	0.41	−0.10	0.44	$p = .10$
Indirect	−0.46***	0.42	0.10*	0.41	$p < .001$
Delayed/nonfading					
Direct	0.18**	0.55	−0.09	0.47	$p < .01$
Indirect	−0.44***	0.38	0.02	0.36	$p < .001$
Delayed/fading					
Direct	0.24***	0.50	−0.27***	0.59	$p < .001$
Indirect	−0.49***	0.46	0.07	0.47	$p < .001$

Note. All direct-comparison judgments were scored from −3 to +3, such that positive values would reflect a preference for the singleton/target over the foursome/referents. Indirect comparisons are the difference scores between the absolute ratings of the singleton and the foursome; therefore, positive values would also reflect a preference for the singleton.
* $p < .05$. ** $p < .01$. *** $p < .001$.

$p = .34$. More important, there was a significant interaction, $F(1, 203) = 5.05$, $p < .05$. As indicated in Table 6, within the early/nonfading condition, the mean direct-comparison responses were not significantly different between the superior and inferior sets. However, within the delayed/nonfading condition and within the delayed/fading condition, these means were significantly different. More specifically, within both of these conditions, there was a significant tendency for participants to rate targets from superior groups as above average (i.e., better than others in the same group). Within the delayed/fading condition, there was also a significant tendency to rate the targets from inferior groups as below average. This tendency was not significant within the delay/nonfading condition.

A Set Quality × Presentation Procedure ANOVA that included only the delayed/nonfading and delayed/fading cells (excluding the early/nonfading cell) revealed an interaction that was not quite statistically significant, $F(1, 133) = 3.41$, $p = .07$. This borderline interaction is intriguing, but it does not provide much confidence that the fading (vs. nonfading) of referents has a reliable influence on comparative bias, especially because the same analysis produced a clear null effect in Experiment 4.

Indirect comparisons. The overall results of the indirect measures are consistent with our previous experiments. A mixed-model ANOVA including set quality and presentation procedure as factors revealed a significant main effect for set quality, $F(1, 203) = 145.24$, $p < .001$, but a nonsignificant procedure effect and a nonsignificant interaction. As can be seen in Table 6, within each presentation procedure cell, the mean indirect-comparison value for superior sets was significantly below zero, which indicates that the group/foursome received better ratings than did the target/singleton. For inferior sets, all three means were directionally above zero, with one mean being significantly greater than zero.

Discussion

Experiment 7 involved new sets of stimuli. Many of the new sets (three pairs out of the five new sets) were social in nature, and the relevant dimensions for these sets were athleticism, prestige, and morality rather than desirability (which was used in Experiments 1–5). However, the patterns of bias did not differ substantially between old and new sets of stimuli, and the results were again consistent with our hypothesis about the influence of the timing factor. When the target item was denoted after a delay (regardless of whether the referents stayed or faded away), there was significant evidence for standard comparative biases in direct-comparison judgments. This occurred even though indirect-comparison measures continued to show that peoples' ratings of a singleton from a superior set (but not from an inferior set) tended to be lower than their ratings of a foursome from that set. This overall pattern across direct and indirect measures in the delayed denotation conditions again implicates differential weighting as a critical factor in producing standard comparative biases in direct-comparison judgment. However, when the target item was denoted early—from the first appearance of a set—the comparative biases were not significant in the direct-comparison judgments. Although this suggests that differential weighting was not as strong in this condition, it does not preclude some influence of differential weighting. In that early/nonfading condition, people's indirect comparisons revealed markedly higher ratings for the foursomes from superior sets than for the singletons from the superior sets. If people had used their absolute evaluations in an equally weighted fashion to generate their direct-comparison judgments, they would have reported the opposite of the standard above-average effect. Therefore, the null effects on the direct-comparison measures in the early/nonfading condition are likely due to the fact that even people in this cell based their comparative judgments more on the singletons and their attributes than on the foursomes and their attributes.

General Discussion

The present experiments reveal a complex yet coherent pattern of results that are critical for understanding how people make comparison judgments. Interpreting the results requires attention to the distinction between indirect comparisons and direct comparisons. In Experiments 1, 2, 3, 4, 6, and 7, the indirect comparisons revealed the same type of results; when participants were asked about stimuli from an attractive or superior group, their ratings of foursomes were significantly higher than their ratings of singletons. Whereas some previous studies have produced this type of pattern on indirect comparisons (e.g., Study 6 of Klar, 2002), we believe this set of studies constitutes the first occasion on which several studies from a given project have consistently produced this finding. The precise reasons for why stimuli from an attractive/superior set are rated higher as a foursome than as a singleton are far from clear (see discussion after Experiment 1). This finding is clearly worthy of additional research attention, especially with respect to how it might mesh or conflict with related findings on product evaluations (e.g., Hsee & Leclerc, 1998) and within social psychology (e.g., McConnell, Sherman, & Hamilton, 1997; Susskind, Maurer, Thakkar, Hamilton, & Sherman, 1999).

More important in the present work was the goal of understanding how people make direct-comparison judgments. The results across the experiments revealed that whether a singleton from an attractive/superior set is rated more attractive than a foursome in a direct comparison depends on how and when the singleton is

denoted as a focal item (if in fact it is denoted as focal). In Experiment 1, the singleton was not denoted as a focal item; participants were simply asked to compare the singleton and foursome against each other. Under these procedures, participants tended to rate the foursomes from attractive sets as more attractive than singletons. Whereas this finding is consistent with the findings from the indirect measures, it constitutes a complete reversal of the standard above-average effects that have previously been found for direct-comparison judgments (e.g., Giladi & Klar, 2002; Klar, 2002; Klar & Giladi, 1997; Suls et al., 2007). In Experiment 2, a condition was added in which the wording of the comparison question identified the singleton as the focal item, but this wording continued to result in a reversal of the standard above-average effect. Experiment 3 ruled out the possibility that the reversal was due to the way in which we had grouped the items on the screen for the participants (such that the foursome appeared as one entity and separate from the singleton). Even when the five stimuli from a set were presented together (suggesting a common a priori grouping) and the singleton was marked with an *X*, participants continued to show a reversal of the standard above-average effect. Having ruled out some seemingly plausible causes of the reversal (carryover effects, focal wording, and the a priori grouping of items), we conducted Experiments 4–7 to directly test our new account as to why comparative biases occur in their standard form and under what conditions they would reverse. The account assumed that the timing with which the focal item is denoted is a critical factor. More specifically, the timing of the denotation of the focal item influences the order in which people attend to the focal and referent items, which thereby influences the extent to which the focal item receives inordinate attention immediately prior to when the comparative judgment is made. Experiment 4 showed that, as hypothesized, when a focal item was denoted after participants had already viewed a full set, robust above-average effects were finally detected (i.e., participants rated focal singletons from attractive sets as more desirable than the referent foursomes). Experiment 5 used a process-tracing methodology to verify that the delayed denotation of a focal item does tend to increase the attention that a focal item receives immediately prior to a comparative judgment. The results from Experiment 6 demonstrated that these shifts in attention can be consequential for determining the magnitude/direction of comparative bias. Finally, in Experiment 7 we confirmed the basic findings and conclusions from Experiment 4, but we did so with new sets of stimuli (some social and some nonsocial) that were rated on different dimensions (e.g., prestige, morality). When the focal item was denoted after a delay, standard comparative biases were detected on direct measures. However, when the focal item was denoted immediately from the onset of a stimulus set, no comparative biases were detected on direct measures.

More broadly, then, whereas previous research on member-to-group comparisons has consistently found that members from attractive/superior groups tend to be rated as above the average relative to the group, the present findings for direct-comparison judgments reveal that this tendency is far from universal. Not only do the findings illustrate when the above-average effect (or its reversal) will be found, but also they provide new information about relevance of causal factors that underlie the standard effects. From previous work on member-to-group comparisons, one might assume that the mechanisms specified by the focalism or generalized-group accounts (or some combination) are typically adequate to produce enough differential weighting to yield standard above- and below-average effects. However, although a typical generalized-group account would have predicted a standard above-average effect in Experiment 1, the reverse was found. For Experiments 2 and 3, both focalism accounts and generalized-group accounts would have predicted standard above-average effects, but the reverse was also found in these studies. Clearly, the mechanisms specified by these accounts were, by themselves, too weak to cause enough differential weighting (favoring the singleton) to thereby yield a standard above-average effect.[2] The timing factor, which has not been identified in previous work on member-to-group comparisons, was a critical factor for turning a reversed above-average effect (Experiments 1–3) into a standard above-average effect (Experiments 4 and 7).

We are not suggesting that focalism and generalized-group accounts are invalid, only that they are substantially incomplete for explaining when standard above-average effects will occur. Our timing account, as well as the focalism accounts and the generalized-group accounts, all specify conditions under which a singleton/focal item receives disproportionate weight in a judgment process. Therefore, under the right conditions (such as those in Experiments 4 and 7) it is possible for the mechanisms of these accounts to all work in concert to produce standard above-average effects. The influences described by these accounts might be additive—with the influence of timing having a substantial independent role. However, the influences described by these accounts might be interactive, such that the mechanisms of the focalism and generalized-group accounts are not as strong when the singleton/focal item is denoted initially rather than late. For example, a typical generalized-group account would suggest that when a judge makes a comparative judgment about a focal item from an attractive set, he or she will find it difficult to evaluate the group of referents and, therefore, might direct disproportionate attention to the singleton/focal item. This difficulty (and the disproportional weighting that would ultimately result) might be minimal for the judge when the focal item is denoted immediately. However, the difficulty might be substantial when the focal item is denoted late—because when the judge first evaluates the items, he or she does not even know which items will be grouped together as the referents. Evaluating them as a group of referents might require a partial reprocessing of each one. To the extent that the judge forgoes this reprocessing and primarily attends to the singleton, the differential weighting would be substantial. This example illustrates one way in which existing accounts of differential weighting could be interactive. The larger question of whether the influences of the timing, focalism, and generalized-group accounts tend to be additive or interactive cannot be answered with the present data.

[2] It is important to recall that in making general (not comparative) evaluations of the stimuli from attractive groups, participants evaluated foursomes more highly than they evaluated singletons. Hence, when participants were formulating direct comparisons, attractive foursomes essentially started with an evaluation advantage over attractive singletons (in terms of the valance of the evaluation). For direct comparisons to show a standard above-average effect, the disproportionate weighting of the singleton/focal item would need to be rather extreme; equal weighting would presumably still result in a pattern that resembled that for indirect judgments (i.e., a reversed above-average effect).

Nevertheless, what remains clear from the present studies is that the mechanisms described in our timing account can be crucial determinants of the magnitude and direction of comparative biases.

The Role of Timing in Previous Research and the LOGE Approach

Given the results of the present experiments, it appears that a delayed denotation of the target item might have played an important but unidentified role in previous experiments that showed standard above- and below-average effects in member-to-group comparisons. In all of the five experiments from Giladi and Klar (2002) that demonstrated above- or below-average effects, the focal item was identified after participants had already examined the full set of items. For example, in the initial study involving soaps, participants first smelled the set of soaps, then they drew a slip of paper that denoted the focal soap, and then they smelled that focal soap one last time before making a direct comparison rating. In a slightly different paradigm involving ratings of food, participants first wrote a list of six unhealthy or six healthy foods, and then they rated how healthy a food from a randomly selected row (e.g., the food listed in the forth row) is relative to the other listed foods. If participants in the soap study knew which soap was focal from the beginning, or if participants in the food study were shown a list of healthy foods with one premarked, we expect that there would have been significantly less differential weighting and therefore significantly smaller (if not absent or reversed) above- and below-average effects.

Earlier, we briefly discussed Giladi and Klar's (2002) LOGE approach to understanding above- and below-average effects in member-to-group comparisons. According to the LOGE approach, although participants who are asked to make a member-to-group comparison should evaluate the target member relative to the local standard (i.e., the other members of the group), they infelicitously evaluate the target relative to a standard that is a compromise between the local standard and the general standard. A *general standard* is the one that is typically used to generate an answer to "absolute" questions such as "How pleasant is this soap?" (Helson, 1964). Hence, a general standard is based on the entirety of a person's past and concurrent experiences with the category. For participants in a pleasant soaps condition of Giladi and Klar's experiment, the general standard would tend to be lower than the local standard (because the general standard is a function of all soaps in memory, some of which are quite bad). As a result, the target soaps will tend to be better than the compromise standard, yielding "above-average" responses. For participants in an unpleasant soaps condition, the general standard would tend to be higher than the local standard, yielding "below-average" responses regarding the target soaps.

Although the LOGE approach clearly takes a different perspective on explaining above- and below-average effects, it is nevertheless fully compatible with the notion of differential weighting. In fact, the LOGE approach cannot explain the above- and below-average effects without differential weighting. The approach assumes that making an absolute evaluation of a target involves a comparison between the target and the general standard and making an absolute evaluation of the referents involves a comparison between the referents and the same general standard. If people gave equal weight to both of these absolute evaluations when making a direct-comparison judgment, then a participant's use of general standards for evaluating the target and the referents would fully cancel out. The direct comparison would not at all reflect how the target compares with the general standard—only how it compares with the referents. No above- or below-average effect would occur. However, when there is differential weighting in which the absolute evaluation of the target has more weight than the absolute evaluation of the referents, then the cancellation is not complete. Therefore, the direct comparison would partially reflect how the target compares with the general standard. This is essentially what Giladi and Klar (2002) referred to as the infelicitous influence of the general standard, and it can cause above- and below-average effects.

Hence, the LOGE approach is fully compatible with the general notion of differential weighting. Its articulation of how above- and below-average effects can develop includes differential weighting. With that said, it is important to recognize that LOGE is also moderately permissive regarding possible answers to the question of why differential weighting occurs within a given experimental context. Although the authors of the LOGE approach have previously discussed accounts akin to focalism and the generalized-group accounts for producing differential weighting (see Klar, 2002; Klar & Giladi, 1997), the LOGE model does not formally commit itself to any one causal account of differential weighting. Depending on the experimental context, the differential weighting involved in the LOGE approach could arise because of target/referent differences or single/group differences (and even self/other differences if the self was one of the entities being judged).

More important (given the present findings), the discovery of the role of timing in above- and below-average effects does not require any formal change to the LOGE approach if the LOGE approach remains permissive as to precisely what makes differential weighting occur within specific instances of comparative judgment. In other words, the fact that the focalism and generalized-group accounts must now be interpreted vis-à-vis the timing factor does not invalidate any aspect of the LOGE approach. Differential weighting and the effects explained by LOGE can simply be augmented or reduced depending on when the target item is denoted within a paradigm or judgment setting.

Implications for Related Fields

A main lesson from the present work is that the differential weighting that leads to comparative biases may vary dramatically as a function of when a person becomes aware of which item from a group is the focal item/singleton. This lesson is important not only for direct-comparison judgments but also potentially for any field in which referent dependent judgments or allocations are critical. We use the term *referent-dependent* to describe any type of judgment or allocation for which there are referents that—from a prescriptive perspective—must be considered by the respondent. For example, comparative optimism judgments are one form of referent-dependent judgments. Perhaps the most widely known finding from the comparative optimism literature is that people tend to report that they are less likely than others to experience various negative life events (e.g., Burger and Palmer, 1992; Helweg-Larsen and Shepperd, 2001; Middleton, Harris, & Surman, 1996; Shepperd, Carroll, Grace, & Terry, 2002; Weinstein, 1980). Whereas such findings could be attributable to motivated

influences or to nonmotivated egocentrism (see, e.g., Alicke et al., 1995; Klar et al., 1996; Perloff & Fetzer, 1986; Regan, Snyder, & Kassin, 1995; Rothman, Klein, & Weinstein, 1996), recent research has also provided evidence consistent with the notion that focalism or generalized-group mechanisms could also contribute to comparative optimism effects (Chambers et al., 2003; Kruger & Burrus, 2004). It is also not uncommon for surveys to solicit comparative optimism judgments about events that people are not directly involved in, making egocentrism irrelevant but leaving focalism and generalized-group mechanisms quite relevant (e.g., *Compared with these other cities, how likely is it that Los Angeles would be the site of the next major terrorist attack?*). Hence, the timing factor might be influential in the overall amount of differential weighting and comparative bias that is exhibited with such comparative optimism judgments.

A closely related type of referent-dependent judgment is a traditional probability judgment (e.g., *What is the probability that Los Angeles will be the site of the next major terrorist attack? . . .that Hillary Clinton will be the nominee?. . .that Option C is the correct answer?*). Although probability questions typically do not explicitly prescribe a comparison referent, there is a clear referent set implied that must be considered in order for a respondent's probability judgments to be accurate. Research indicates that even when there are only two possible outcomes for an event, focalism is influential; people's probability judgments about one outcome are driven more by the strength of evidence for that outcome than about the strength of the evidence for the only alternative outcome (e.g., Fox & Levav, 2000; Idson, Krantz, Osherson, & Bonini, 2001; Lehman, Krosnick, West, & Li, 1992; Macchi, Osherson, & Krantz, 1999; McKenzie, 1998; Moore & Kim, 2003; Windschitl, 2000; Windschitl et al., 2003; Yamagishi, 2002). Research also indicates that generalized-group mechanisms (or mechanisms closely related to them) are also important. In fact, a critical stipulation of an influential theory of probability judgment—support theory—is that the support (i.e., perceived evidence) for a group of hypotheses (or possible outcomes) tends to be less than the sum of the support for each of the individual hypotheses (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994; see also Koehler, Brenner, & Tversky, 1997). This subadditivity of support can lead to highly inflated probability judgments about a singleton—essentially because the full support for the group of referents is difficult for people to accurately assess. Despite the clear relevance of focalism and generalized-group mechanisms for probability judgments, we know of no research in which the timing of the denotation of the focal hypothesis/outcome was manipulated. If the results of the present work extend to probability judgments, then the severity of some probability biases might vary as a function of such a timing factor.

Proportion judgments, rank judgments, and allocation decisions are other forms of referent-dependent tasks for which the timing manipulation might be important. One domain in which these types of tasks have been investigated is the domain of product evaluation and more generally in contingent valuation. Comparison processes are being heavily researched because of their relevance to the evaluation process (e.g., Hsee, 1996; Hsee & Leclerc, 1998; Posavac, Sanbonmatsu, Kardes, & Fitzsimons, 2004; Wang & Wyer, 2002). Several product evaluation studies have demonstrated effects akin to above-average effects in the use of dependent variables such as ranks, direct-comparison judgments, and purchase intentions about individual products from a set (e.g., see, Kardes, Sanbonmatsu, Cronley, & Houghton, 2002; Posavac et al., 2002; Posavac et al., 2004; Posavac, Kardes, Sanbonmatsu, & Fitzsimons, 2005). In a study about how people evaluate and prioritize environmental assets (in this case, four national parks), the researchers found that participants overvalued whatever park was randomly selected for evaluation (Posavac et al., 2006). For example, when asked to indicate what amount out of every $100 the government spends on the four parks should go to the focal park, the average response was significantly greater than $25. This type of effect can be attributable to focalism and/or to generalized-group mechanisms. The timing with which the focal item is denoted in these types of studies has not been systematically investigated, but we believe that valuations of products, environmental efforts, and other entities might be significantly affected by variations in timing.

For all of the types of judgments and domains mentioned above, the timing of whether a focal item is denoted early or late is not just a theoretically relevant factor. This factor varies across real-world situations in which important direct-comparison judgments or other referent-dependent judgments are made. There are numerous real-world situations in which people are presented with a set of entities (e.g., products, people, possible outcomes) and they know from the beginning that there is one entity that they need to judge. However, there are also numerous real-world situations in which people are familiarized with a set of entities and they only later learn which one they must specifically judge. Hence, there are important applied reasons to study both the early and delayed denotation conditions in a systematic fashion. Indeed, investigating referent-dependent judgments in a delayed denotation paradigm when trying to understand real-world settings in which tasks tend to resemble early denotation paradigms (or vice versa) might leave one with a false sense of the magnitude and perhaps even the direction of bias.

References

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology, 68,* 804–825.

Betsch, T., Kaufmann, M., Lindow, F., Plessner, H., & Hoffmann, K. (2006). Different principles of information integration in implicit and explicit attitude formation. *European Journal of Social Psychology, 36,* 887–905.

Bruine de Bruin, W., & Keren, G. (2003). Order effects in sequentially judged options due to the direction of comparison. *Organizational Behavior & Human Decision Processes, 92,* 91–101.

Burger, J. M., & Palmer, M. L. (1992). Changes in and generalization of unrealistic optimism following experiences with stressful events: Reactions to the 1989 California earthquake. *Personality & Social Psychology Bulletin, 18,* 39–43.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90,* 60–77.

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above average and comparative optimism effects. *Psychological Bulletin, 130,* 813–838.

Chambers, J. R., Windschitl, P. D., & Suls, J. (2003). Egocentrism, event frequency, and comparative optimism: When what happens frequently is

more likely to happen to me. *Personality & Social Psychology Bulletin, 29,* 1343–1356.

Eiser, J. R., Pahl, S., & Prins, Y. R. A. (2001). Optimism, pessimism, and the direction of self–other comparisons. *Journal of Experimental Social Psychology, 37,* 77–84.

Fox, C. R., & Levav, J. (2000). Familiarity bias and belief reversal in relative likelihood judgment. *Organizational Behavior & Human Decision Processes, 82,* 268–292.

Giladi, E. E., & Klar, Y. (2002). When standards are wide of the mark: Nonselective superiority and inferiority biases in comparative judgments of objects and concepts. *Journal of Experimental Psychology: General, 131,* 538–551.

Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior.* New York: Harper & Row.

Helweg-Larsen, M., & Shepperd, J. A. (2001). Do moderators of the optimistic bias affect personal or target risk estimates? A review of the literature. *Personality & Social Psychology Review, 5,* 74–95.

Hodges, S. D., Bruininks, P., & Ivy, L. (2002). It's different when I do it: Feature matching in self–other comparisons. *Personality & Social Psychology Bulletin, 28,* 40–53.

Hoorens, V. (1995). Self-favoring biases, self-presentation, and the self–other asymmetry in social comparison. *Journal of Personality, 63,* 793–817.

Houston, D. A., Sherman, S. J., & Baker, S. M. (1989). The influence of unique features and direction of comparison on preferences. *Journal of Experimental Social Psychology, 25,* 121–141.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior & Human Decision Processes, 67,* 247–257.

Hsee, C. K., & Leclerc, F. (1998). Will products look more attractive when presented separately or together? *Journal of Consumer Research, 25,* 175–186.

Idson, L. C., Krantz, D. H., Osherson, D., & Bonini, N. (2001). The relation between probability and evidence judgment: An extension of support theory. *The Journal of Risk and Uncertainty, 22,* 227–249.

Kardes, F. R., Sanbonmatsu, D. M., Cronley, M. L., & Houghton, D. C. (2002). Consideration set overevaluation: When impossibly favorable ratings of a set of brands are observed. *Journal of Consumer Psychology, 12,* 353–361.

Klar, Y. (2002). Way beyond compare: Nonselective superiority and inferiority biases in judging randomly assigned group members relative to their peers. *Journal of Experimental Social Psychology, 38,* 331–351.

Klar, Y., & Giladi, E. E. (1997). No one in my group can be below the group's average: A robust positivity bias in favor of anonymous peers. *Journal of Personality and Social Psychology, 73,* 885–901.

Klar, Y., Medding, A., & Sarel, D. (1996). Nonunique invulnerability: Singular versus distributional probabilities and unrealistic optimism in comparative risk judgments. *Organizational Behavior & Human Decision Processes, 67,* 229–245.

Koehler, D. J., Brenner, L. A., & Tversky, A. (1997). The enhancement effect in probability judgment. *Journal of Behavioral Decision Making, 10,* 293–313.

Krizan, Z., & Suls, J. (2007). *Losing sight of oneself in the above-average effect: When egocentrism, focalism, and group diffuseness collide*. Manuscript submitted for publication.

Krizan, Z., & Windschitl, P. D. (2007). Team allegiance can lead to both optimistic and pessimistic predictions. *Journal of Experimental Social Psychology, 43,* 327–333.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77,* 221–232.

Kruger, J., & Burrus, J. (2004). Egocentrism and focalism in unrealistic optimism (and pessimism). *Journal of Experimental Social Psychology, 40,* 332–340.

Lehman, D. R., Krosnick, J. A., West, R. L., & Li, F. (1992). The focus of judgment effect: A question wording effect due to hypothesis confirmation bias. *Personality & Social Psychology Bulletin, 18,* 690–699.

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review, 106,* 210–214.

McConnell. A. R., Sherman, S. J., & Hamilton, D. L. (1997). Target entitativity: Implications for information processing about individual and group targets. *Journal of Personality and Social Psychology, 72,* 750–762.

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 771–792.

Middleton, W., Harris, P., & Surman, M. (1996). Give'em enough rope: Perception of health and safety risks in bungee jumpers. *Journal of Social and Clinical Psychology, 15,* 68–79.

Moore, D. A., & Kim, T. G. (2003). Myopic social prediction and the solo comparison effect. *Journal of Personality and Social Psychology, 85,* 1121–1135.

Otten, W., & Van Der Pligt, J. (1996). Context effects in the measurement of comparative optimism in probability judgments. *Journal of Social & Clinical Psychology, 15,* 80–101.

Perloff, L. S., & Fetzer, B. K. (1986). Self–other judgments and perceived vulnerability to victimization. *Journal of Personality and Social Psychology, 50,* 502–510.

Posavac, S. S., Brakus, J. J., Jain, S. P., & Cronley, M. L. (2006). Selective assessment and positivity bias in environmental valuation. *Journal of Experimental Psychology: Applied, 12,* 43–49.

Posavac, S. S., Kardes, F. R., Sanbonmatsu, D. M., & Fitzsimons, G. J. (2005). Blissful insularity: When brands are judged in isolation from competitors. *Marketing Letters, 16,* 87–97.

Posavac, S. S., Sanbonmatsu, D. M., & Ho, E. A. (2002). The effects of the selective consideration of alternatives on consumer choice and attitude-decision consistency. *Journal of Consumer Psychology, 12,* 203–213.

Posavac, S. S., Sanbonmatsu, D. M., Kardes, F. R., & Fitzsimons, G. J. (2004). The brand positivity effect: When evaluation confers preference. *Journal of Consumer Research, 31,* 643–651.

Price, P. C. (2001). A group size effect on personal risk judgments: Implications for unrealistic optimism. *Memory & Cognition, 29,* 578–586.

Price, P. C., Smith, A. R., & Lench, H. C. (2006). The effect of target group size on risk judgments and comparative optimism: The more, the riskier. *Journal of Personality and Social Psychology, 90,* 382–398.

Regan, P., Snyder, M., & Kassin, S. M. (1995). Unrealistic optimism: Self-enhancement or person positivity? *Personality & Social Psychology Bulletin, 21,* 1073–1082.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology, 37,* 322–336.

Rothman, A., Klein, W., & Weinstein, N. (1996). Absolute and relative biases in estimations of personal risk. *Journal of Applied Social Psychology, 26,* 1213–1236.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104,* 406–415.

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science, 9,* 340–346.

Shepperd, J. A., Carroll, P., Grace, J., & Terry, M. (2002). Exploring the causes of comparative optimism. *Psychologica Belgica, 42,* 65–98.

Shepperd, J. A., Helweg-Larsen, M., & Ortega, L. (2003). Are comparative risk judgments consistent across time and events? *Personality & Social Psychology Bulletin, 29,* 1169–1180.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89,* 845–851.

Suls, J., Krizan, Z., Chambers, J. R., & Mortensen, C. (2007). *The role of*

focalism and group diffuseness in comparative biases. Manuscript submitted for publication.

Susskind, J., Maurer, K., Thakkar, V., Hamilton, D. L., & Sherman, J. W. (1999). Perceiving individuals and groups: Expectancies, dispositional inferences, and causal attributions. *Journal of Personality and Social Psychology, 76,* 181–191.

Taylor, S. E., & Fiske, S. T. (1975). Point of view and perceptions of causality. *Journal of Personality and Social Psychology, 32,* 439–445.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547–567.

Wang, J., & Wyer, R. S. (2002). Comparative judgment processes: The effects of task objectives and time delay on product evaluations. *Journal of Consumer Psychology, 12,* 327–340.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39,* 806–820.

Weinstein, N. D., & Lachendro, E. (1982). Egocentrism as a source of unrealistic optimism. *Personality & Social Psychology Bulletin, 8,* 195–200.

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78,* 821–836.

Windschitl, P. D. (2000). The binary additivity of subjective probability does not indicate the binary complementarity of perceived certainty. *Organizational Behavior & Human Decision Processes, 81,* 195–225.

Windschitl, P. D., Kruger, J., & Simms, E. N. (2003). The influence of egocentrism and focalism on people's optimism in competitions: When what affects us equally affects me more. *Journal of Personality and Social Psychology, 85,* 389–408.

Yamagishi, K. (2002). Proximity, compatibility, and noncomplementarity in subjective probability. *Organizational Behavior & Human Decision Processes, 87,* 136–155.

Appendix

Mean Absolute Judgments Regarding Attractive and Fair Sets in Experiments 1, 2, 3, 4, and 7

	Attractive		Fair	
Experiment and condition	*M*	*SD*	*M*	*SD*
1				
Singleton	5.91	0.82	3.18	0.81
Foursome	6.56	0.52	3.26	0.81
2 (Balanced wording)				
Singleton	5.60	1.35	3.35	0.99
Foursome	6.38	1.20	3.40	1.06
2 (Focal wording)				
Singleton	5.88	0.77	3.08	1.18
Foursome	6.33	0.38	3.35	0.95
3				
Singleton	5.85	0.74	3.10	1.15
Foursome	6.59	0.45	3.32	0.95
4 (Nonfading referents)				
Singleton	5.68	0.83	2.99	0.88
Foursome	6.27	0.73	3.20	0.78
4 (Fading referents)				
Singleton	5.78	0.64	2.68	0.71
Foursome	6.59	0.40	2.66	0.71
7 (Early/nonfading)				
Singleton	6.19	0.46	2.46	0.57
Foursome	6.65	0.30	2.36	0.58
7 (Delayed/nonfading)				
Singleton	6.12	0.53	2.49	0.54
Foursome	6.55	0.39	2.48	0.47
7 (Delayed/fading)				
Singleton	6.12	0.47	2.55	0.73
Foursome	6.61	0.33	2.47	0.69

Note. All absolute judgments were scored from 1 to 7, with 7 reflecting a high rating regarding the relevant dimension.

Received April 18, 2007
Revision received June 27, 2007
Accepted June 28, 2007 ■