When good = better than average

Don A. Moore*
Tepper School of Business
Carnegie Mellon University

Abstract

People report themselves to be above average on simple tasks and below average on difficult tasks. This paper proposes an explanation for this effect that is simpler than prior explanations. The new explanation is that people conflate relative with absolute evaluation, especially on subjective measures. The paper then presents a series of four studies that test this conflation explanation. These tests distinguish conflation from other explanations, such as differential weighting and selecting the wrong referent. The results suggest that conflation occurs at the response stage during which people attempt to disambiguate subjective response scales in order to choose an answer. This is because conflation has little effect on objective measures, which would be equally affected if the conflation occurred at encoding.

Keywords: comparative judgment, solo comparison, subjective measures, better-than-average.

1 Introduction

There is an inconsistency in research findings on comparative judgment. Both better-than-average (BTA) and worse-than-average (WTA) effects tend to be stronger in direct measures than in indirect measures of comparative judgment. Given the robustness, durability, and profound consequences of biases in comparative judgment (Dunning, Heath, & Suls, 2004; Malmendier & Tate, 2004; Odean, 1998; Weinstein & Lachendro, 1982), this inconsistency deserves investigation.

1.1 Differences between direct and indirect measures of comparative judgment

Direct measures of comparison ask people to explicitly compare two things — usually comparing themselves with others. For instance, Moore and Kim (2003, Experiment 3) had their participants take a 10-item trivia quiz that was either very easy (mean score = 87% correct) or very difficult (mean score = 15% correct). A direct comparative measure asked "How do you expect to score relative to others?" and participants responded on a scale ranging from 1 (*well below average*) to 7 (*well above average*). Those who took the easy quiz expected that they would score above average, while takers of the difficult quiz expected to score below average.[1]

Indirect measures ask people to evaluate the target and the referent (to which the target is being compared) separately using the same absolute standard. For instance, Moore and Kim (2003) also asked their participants to estimate their own and others' scores on the trivia quiz. The indirect measure of comparison is calculated by subtracting estimated performance for self minus others. Those who had taken the easy quiz guessed that their own scores would be higher than those of others, whereas those who had taken the difficult quiz guessed that their own scores would be lower.

The curious fact is that both BTA and WTA effects are stronger in direct measures of comparison than in indirect measures of comparison (Helweg-Larsen & Shepperd, 2001; Klar & Giladi, 1997; Otten & van der Pligt, 1996). In other words, task difficulty has a bigger effect on indirect than on direct comparative evaluations. Moore and Kim's manipulation of task difficulty had significantly larger effect on the direct measure than had on the indirect measure. To be precise, Moore and Kim's (2003) manipulation of task difficulty accounted for 20% of the variance in direct measure. By contrast, the difficulty manipulation accounted for only 8% of the variance

*This material is based upon work supported by the National Science Foundation under Grant No. 0451736 and by a Berkman faculty development grant from Carnegie Mellon University. Thanks to Barbara Mellers for inspiration on the first experiment and Bill Klein for helpful comments on the manuscript. Sam Swift, Wemi Peters, and Erin Morgan helped collect the data. Address: CMU/Tepper, 5000 Forbes Avenue, Pittsburgh, PA 15213. Email: don.moore@alumni.carleton.edu.

[1] These findings would seem to contradict the so-called "hard-easy" effect on overconfidence, which shows that people are most likely to overestimate their performances on difficult tasks (Erev, Wallsten, & Budescu, 1994). In fact, the two effects are compatible (Larrick, Burson, & Soll, 2007). Moore and Small (2007) offer an explanation that helps reconcile them with an explanation that shows that they arise due to the same underlying psychological processes (see also Moore & Healy, 2007).

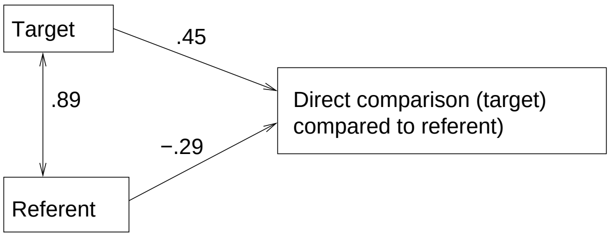

Figure 1: Path analysis appearing to show greater weighting of the target (self) than the referent (others) in direct comparative judgment (data from Moore & Kim, 2003, Experiment 3).

in the indirect measure. The most popular explanation for this difference is differential weighting.

1.2 The differential weighting explanation

The differential weighting explanation holds that people weight the target and the referent differently when making comparative judgments. That is, the self (or the target of comparison) is weighed more heavily than is the other (or the referent). The referent may be neglected for several reasons, including the fact that information about the self is generally more accessible, more vivid, or more reliable than information about others (Kruger, Windschitl, Burrus, Fessel, & Chambers, in press). As a result, when the target's performance is good it is rated above average, and when it is bad it is rated below average. This theory holds that stronger BTA effects (on easy tasks) and WTA effects (on hard tasks) on direct than on indirect judgments has deep psychological origins. It is the result of differences in the accessibility, salience, or reliability of knowledge of the target and referent, and is not simply the result of the response scale used to elicit people's beliefs.

What is the evidence for the differential weighting explanation? A number of studies have presented the results of path analyses (Chambers, Windschitl, & Suls, 2003; Giladi & Klar, 2002; Klar, 2002; Klar & Giladi, 1997; Kruger, 1999; Windschitl, Kruger, & Simms, 2003). These analyses utilize three variables: (1) absolute evaluation of a target individual, (2) absolute evaluation of the referent group (or a representative member of the group), and (3) a direct comparative judgment of the individual relative to the group. Results typically demonstrate that the direct comparative judgment is highly correlated with the individual's own absolute performance but only weakly (often insignificantly) negatively correlated with the absolute performance of the reference group—consistent with differential weighting. Figure 1 shows a path analysis using Moore and Kim's (2003, Experiment 3) data. Direct comparative judgments ought to weight the target and the referent equally and oppositely, but the target appears to be weighted more heavily (.45) than is the referent (–.29).

Differential weighting may be an accurate description of the results of the path analyses. However, it is not necessarily an accurate description of the underlying psychological processes in comparative judgment. An important flaw in these path analyses suggests that the result may not be diagnostic of actual differential weighting by the person making the comparison. If people are making comparisons sensibly then there *should* be less variance in estimates of the group average than in estimates of individual performance. After all, if everyone correctly estimated the group average, then there should be no between-person variance in estimates of the group average and it would therefore be non-predictive of comparative judgments in path analyses.

1.3 Subjective vs. objective measures

But there is an additional concern regarding the way in which path analyses are often conducted: Absolute judgments of target and referent are often measured on subjective verbally-anchored scales. For example, Klar and Giliadi (1997) had their participants rate photographs using an absolute scale that ran from 1 (*very unattractive*) to 9 (*very attractive*). Comparative judgments of those same photographs were made using a scale that ran from –3 (*much less attractive than the average student*) to +3 (*much more attractive than the average student*). Biernat's work on shifting standards has demonstrated that responses on such scales are sensitive to the relevant comparison group (Biernat, 2003; Biernat & Manis, 1994; see also Mussweiler & Strack, 2000). An American woman who measures 5 feet 9 inches would be more likely to describe herself as tall than would a man of the same height. In this case, as in innumerable others, evaluation depends crucially on the context of relevant comparison others.

Ratings on subjective response scales are unlikely to be pure measures of either absolute or relative assessment (Burson & Klayman, 2005). Measures intended to tap absolute performance will, at least in part, be measures of relative performance as well. Imagine a simple test on which everyone does well. If one person received a score of 80% and the other 99 test-takers all scored above 90%, when asked, "How well did you do on the test?" the person who got 80% is unlikely to rate himself an 8 on a 10-point scale if that scale is anchored with verbal labels such as "*very poorly*" and "*very well.*" Similarly, when relative performance is measured using subjective response scales, we should expect that judgments may be influenced by absolute performance. It should be no surprise that the target's rated absolute performance correlates highly with relative performance when they are both measured on subjective response scales. To some extent, they are measuring the same thing.

This brings us to the second explanation for why BTA and WTA effects are stronger in direct than in indirect measures of comparative judgment: conflation. Conflation is the error of treating two distinct concepts as if they were one. People routinely conflate absolute and relative evaluation with each other when making comparative judgments (see Giladi & Klar, 2002). That is, when people are asked to compare themselves with others, their comparative judgments are contaminated by their absolute judgments of their own performances (Klar & Giladi, 1999). Subjectively anchored response scales force participants to construe the scale in order to map their own private knowledge onto the response scale. Idiosyncratic construals open these subjective scales to influence (or contamination) by other considerations (Biernat, 2003; Biernat, Manis, & Kobrynowicz, 1997; Dunning, Meyerowitz, & Holzberg, 1989). After having done well at a task, people are more likely to rate themselves as being above average, even if it is a simple task on which everyone should be expected to do well.

The new hypothesis tested in this paper is that this effect is not due to some sort of profound differential weighting of self over others — rather, this effect is a mundane conflation caused by vague questions and subjective response scales. The new part of this explanation is the idea that the way comparative judgment is measured matters. Unlike the differential weighting explanation, which predicts equivalent BTA and WTA effects across both subjective and objective measures of comparative judgment, the conflation hypothesis predicts that BTA and WTA effects will shrink or disappear when comparative judgments are measured using unambiguous objective response formats. The implication would be that people *can* make more accurate estimates of comparative judgment, but that experimenters often fail to ask the question in ways most likely to elicit an uncontaminated comparative judgment.

This is not an obscure technical issue of measurement — it is important for two reasons. First, subjective verbally-anchored scales are perhaps the single most commonly used measure in psychological research, including work on comparative judgment. If such measures elicit systematically biased responses, the implications may be profound with respect to both the reinterpretation of prior findings and the optimal design of future studies. Second, it would illuminate the psychological processes at work in comparative judgment. If people can make accurate comparative judgments when provided with unambiguous response scales it shows that conflation is occurring at the response stage during which people's mental representations are translated into behavioral responses. If, instead, the conflation of relative and absolute evaluation were occurring during encoding, then it would appear in people's responses to all sorts of comparative judgments, regardless of the response format of the question, because there would be no unconflated evaluation to retrieve.

1.4 The present studies

Without a doubt, BTA and WTA effects are multiply determined. It is not the goal of this paper to show that conflation is the *only* cause of these effects, only that conflation is *a* contributing cause and distinct from other causes. This simple demonstration has important repercussions. It suggests that prior research has often overestimated the size of both BTA and WTA effects through the use of vague measures. It also suggests a straightforward methodological solution to this problem: clearer objective measures.

The four experiments presented here examine the conflation explanation by testing its specific predictions and by eliminating as many other explanations for BTA and WTA effects as possible. I use both subjective response scales and also clearer objective measures. I find that BTA effects (on easy tasks) and WTA effects (on difficult tasks) weaken with more objective measures. Experiment 1 replicates the BTA and WTA effects shown elsewhere, and seeks to eradicate them through experimental manipulations that provide participants with clear and unambiguous information about the performances of themselves and others. Full information about others' performances should rule out explanations based on greater regressiveness in estimates of others, because these "differential information" theories assume errors in people's estimations of others. However, the conflation explanation predicts the persistence of BTA and WTA effects, even in the presence of full information, but especially on subjective measures of comparative judgment.

Experiments 2, 3, and 4 also provide participants with full information about target and referents, but in order to rule out the role of egocentrism, participants are only asked to compare other individuals to each other. Finally, research has shown that BTA and WTA effects are stronger when the referent is a group rather than an individual (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Hoorens & Buunk, 1993; Klar, Medding, & Sarel, 1996; Perloff & Fetzer, 1986). In order to rule out this influence, Experiments 3 and 4 ask participants to compare two individuals whose performances are known. The fact that the effect persists, even in this context, but only on subjective direct measures, is explained better by conflation than by other theories.

The research presented in this paper contributes to theory and research on several different dimensions. First, the four studies I report are the first to put the conflation explanation for BTA and WTA effects to the test by systematically comparing judgments varying in their subjectivity. The key prediction, confirmed in all four studies, is that BTA effects on easy tasks and WTA effects on hard tasks are stronger for subjective than objective direct comparative judgments. Second, the four studies presented here do another thing that prior research has not: present participants with excellent information about performance by target and referent. This is important for ruling out other explanations for BTA and WTA effects. Third, Experiments 2, 3, and 4 take tests of the conflation explanation to their logical extreme by minimizing egocentric motives and having people compare two targets about which they have complete information.

2 Experiment 1: The effect of information

Experiment 1 varied task difficulty and crossed this manipulation with a manipulation of feedback. The manipulation of difficulty is crucial, as it is in all the studies presented in this paper, for providing variation on an absolute scale (performance) that is uncorrelated with variation relative to others. The optimal conditions under which to study conflation between absolute and relative performance is when they vary independently, and we can therefore measure their independent effects on perceptions of performance.

The manipulation of feedback allows me to test whether having better information about one's own and others' performances influences the magnitude of BTA and WTA effects, and whether the way in which comparative judgments are measured moderates changes in the effects' sizes. This issue is important, given the critical role that information has been shown to play in BTA and WTA effects (Kruger et al., in press; Moore & Cain, 2007). The experiment includes four different measures of relative self-evaluation of varying ambiguity.

Ambiguous response scales rely on participants to infer the meaning of the scale (see Gannon & Ostrom, 1996; Schwarz, 1999; Schwarz & Hippler, 1995; Schwarz, Hippler, Deutsch, & Strack, 1985). For instance, Chambers, Windschitl, and Suls (2003) asked their participants, "Compared to the average student of the same age and sex, how likely is it that you will win free tickets to a hockey game?" and invited them to respond on an 11-point scale (–5 = *much less than the average student* to +5 = *much more likely than the average student*). Would a 20% greater likelihood correspond to a +4 or a +5 on the scale? There is no right answer to this question, and so each individual must construe its meaning independently. Evidence clearly shows that this subjective construal process is vulnerable to influence from outside contextual forces (Biernat et al., 1997; Schwarz, 2001; Schwarz, Bless et al., 1991; Schwarz, Grayson, & Knaeuper, 1998; Schwarz, Knaeuper, Hippler, Noelle-Neumann, & Clark, 1991). The specific possibility tested in this paper is that the construal process is vulnerable to conflation between absolute performance and comparative judgments.

Objective measures, by contrast, have a correct answer and the rules for determining that answer are common knowledge. Svenson (1981) asked his participants to give themselves a percentile ranking relative to all other participants in the experiment with respect to their driving abilities. Assuming a common definition of what constitutes driving abilities and how to measure them, respondents' self-reported percentile ranks can be compared to their actual percentile ranks.

This raises an important distinction between ambiguity regarding what is being measured and ambiguity about how to measure it. Because Svenson did not tell his participants how exactly to evaluate driving abilities, it was unclear exactly what they were supposed to be measuring. Studies have persuasively shown that the greater the ambiguity in what is being measured, the greater the resulting biases in comparative judgments (Burson & Klayman, 2005; Dunning et al., 1989; Klein, 2001; Klein & Buckingham, 2002). This paper, by contrast, focuses on a task for which there is very little ambiguity about what is being measured: the number of questions that people get right on a ten-item trivia quiz. Instead, I vary the ambiguity of the questions I use to measure it. The greater size of BTA and WTA effects using ambiguous subjective measures of comparative judgment reveal the greater interference of conflation.

Conflation predicts that greater subjectivity of the response scale will increase both BTA and WTA effects. The differential weighting and differential information explanations are both silent on the question of the influence of the response scale.

2.1 Method

Participants. One hundred thirty-nine undergraduate students at Carnegie Mellon University agreed to participate in exchange for pay. Participants had a mean age of 22 years (*SD* = 5.23) and 61 percent of them were male.

Design. The experiment had a 2 (quiz difficulty) X 3 (feedback) between-subjects design. Quiz difficulty was manipulated between subjects using two 10-item trivia quizzes plus an eleventh tiebreaker question (see Appendix). Feedback was also manipulated between subjects: after taking the quiz, 45 participants (roughly one third of the sample) received only absolute feedback about themselves (e.g., "You answered 9 out of 10 questions on the trivia quiz correctly"). Another 46 participants received both absolute feedback about themselves and about others: in addition to learning how many items they had gotten correct, they were told exactly how ten other individuals had performed. They were given a table with one row for each of the ten others, and a score (out of 10) listed on each row, in addition to that person's answer on the tiebreaker question. The mean and standard deviation of these 10 scores roughly matched the entire sample of 128 individuals who had previously taken the quiz (see Moore & Kim, 2003, Experiment 3). The remaining 48 participants received only relative feedback, in the form of a percentile score (e.g., "You scored better than 60 percent of other test-takers"). In all conditions, feedback was truthful.

Procedure. First, each participant completed a trivia quiz. Their completed quizzes were then taken from them and scored. They were then given feedback as described above. Following feedback, participants were informed of the competitive nature of the task and the opportunity to bet:

You may bet as much of your $3 payment as you wish on beating another person on the trivia test. You will win the bet if your score on the test you have just taken is better than that of your opponent.

The opponent was drawn randomly from the list of 10 others that some participants saw. Participants' bets provided a behavioral measure of their beliefs about relative performance. After participants specified how much they wanted to bet (from $0 to 3), they were asked each of the following questions:

1. "*How many of the 10 trivia questions did you answer correctly?*"

2. "*How many of the 10 questions do you predict that your opponent will answer correctly?*"

3. "*How do you expect that you will score relative to all the other people taking the same test as you?*" (marks at 1-*well below average*, 4-*average*, and 7-*well above average*)

4. "*What percentage of the group has scores below yours? (If you expect your score will be the very best, then put 100. If you expect you will score exactly in the middle, put 50. If you expect your score will be the lowest, put 0.)*"

The first two of these questions were objective measures of absolute performance. The difference between what people thought they were going to score and what they thought their opponent was going to score served as the indirect comparative judgment in the data analyses. The third question is a subjective direct comparative measure. The fourth is an objective direct comparative measure.

At the conclusion of the experiment, participants were matched randomly with one of the ten opponents in order to resolve the bet and determine payments. The tiebreaker question was used to resolve tied scores. Participants were then paid, thanked, debriefed, and dismissed.

2.2 Results

Manipulation check. Scores on the simple quiz were indeed higher (M = 8.94, SD = 1.14) than were scores on the difficult quiz (M = 1.97, SD = 1.37), F (1, 133) = 1046.66, $p < .001$, $\eta^2 = .89$.

Direct vs. indirect comparisons. The results replicate the standard finding that BTA and WTA effects are stronger on direct than on indirect comparative measures. Using the 7-point comparative rating scale, the mean rating in the simple condition is 4.96 (SD = 1.06) but the mean rating in the difficult condition is 3.30 (SD = 1.59). When ratings are subject to a 2 (difficulty) X 3 (feedback) between-subjects ANOVA, only the effect of difficulty emerges as significant, F (1, 133) = 53.0, $p < .001$, $\eta^2 = .29$. By contrast, when the indirect comparison (estimated score for self minus estimated score for other) is subject to the same ANOVA, the effect size of difficulty appears to be smaller, F (1, 133) = 26.56, $p < .001$, $\eta^2 = .17$.

In order to compare the effect of difficulty on these two measures, I had to standardize them by converting them to z-scores. I then submitted them to a 2 (difficulty) X 3 (feedback) X (2) (format) mixed ANOVA with repeated measures on question format. If the difficulty X measure interaction is significant, that suggests that the measures differ. Indeed, the effect size difference shows up as a marginally significant interaction effect between elicitation format (subjective comparison scale vs. indirect comparison) and difficulty, F (1, 133) = 3.07, $p = .06$, $\eta^2 = .03$.

To examine differences across all four comparative measures (subjective rating scale, bet on winning, self-reported percentile rank, and indirect comparison), I converted them all to z-scores and subject them to a 2 X 3 X (4) mixed ANOVA. The four measures of judgment served as within-subjects factors. The results again reveal a measure X difficulty interaction, F (3, 399) = 4.10, $p = .007$, $\eta^2 = .03$. This interaction effect describes the fact that the effect of quiz difficulty on comparative judgments was stronger for subjective than for objective measures. See Table 1. Consistent with the conflation explanation, the subjective 1 to 7 rating scale showed the strongest effect of test difficulty ($\eta^2 = .29$), whereas more objective measures (such as estimated percentile rank) showed weaker effects ($\eta^2 = .14$). This specific comparison was tested in a 2 X 3 X (2) mixed ANOVA using only these

Table 1: Results for the four different measures of comparative judgment, Experiment 1. The fourth column shows the effect size of the difference between simple and difficult conditions, as well as the significance of the t-test comparing the two conditions. Regression results predicting indirect comparative judgment for the four different measures of comparative judgment appear in the fifth and sixth columns. The seventh and eighth columns show correlations with actual performance, both relative and absolute.

	Means (SDs):		Simple vs. difficult	Regression results:		Correlations with:	
Self-reported comparative judgment	Simple	Difficult	effect size (η^2)	β Self	β Other	Actual percentile	Actual score
Subjective relative self-rating (1–7 scale)	4.96 (1.06)	3.30 (1.59)	.29***	1.47***	−0.93***	.46***	.66***
Bet	\$1.64 (1.02)	\$1.01 (1.00)	.10***	1.40***	−1.10**	.50***	.45***
Percentile rank (objective measure)	58.8 (24.6)	39.7 (27.5)	.14***	1.34***	−0.97***	.54***	.51***
Indirect comparison	0.59 (1.34)	−0.77 (1.77)	.17***	2.07+	−1.69+	0.47***	0.54***

* $p < .05$, *** $p < .001$, + Independent variables perfectly account for dependent variable.

two direct measures. This analysis again produces the expected measure X difficulty interaction, *F* (1, 133) = 9.90, *p* = .002, η^2 = .07. Note that it is not the case that direct comparative measures show stronger effects of difficulty than indirect measures — this is only true for subjective direct measures.

The last two columns in Table 1 show the correlations between participants' self-reported comparative judgments and their actual performances. Subjective measures appear to be somewhat more strongly associated with absolute performance than are more objective measures. This effect appears modest, but it is worth noting that the subjective relative self-rating is actually more strongly correlated with actual absolute score (*r* = .66, *p* < .001) than it is with actual percentile rank (*r* = .46, *p* < .001), and these two correlations are significantly different, *t* (136) = 2.86, *p* < .001, η^2 = .06. This is the only comparative judgment measure for which this is the case.

Of course, the results of the 2 X 3 X (4) mixed ANOVA also reveal a main between-subjects effect of difficulty, *F* (1, 133) = 36.65, *p* < .001, η^2 = .22, since participants who took the simple quiz generally believed themselves to be more above-average (*M* = .40, *SE* = .09) than did participants who took the difficult quiz (*M* = -.40, *SE* = .09). The results also reveal a marginally significant main between-subjects effect of feedback, *F* (2, 133) = 2.76, *p* = .067, η^2 = .04.[2] No other effects in the 2 X 3 X (4) ANOVA emerge as statistically significant, all *F*s < 1.3, all *p*s > .21. Remarkably, the feedback X difficulty interaction is not significant, indicating that clearer feedback did not reduce the strength of BTA and WTA effects. Since other studies have documented persuasive evidence that clear feedback about others can, at least sometimes, reduce BTA and WTA effects, I can only conclude that Experiment 1's feedback manipulation was not strong enough do so.

Tests of differential weighting. In the interests of comparing the results of Experiment 1 to prior evidence of differential weighting, I used the four measures of comparative judgment as dependent variables in four different path analyses, using participants' absolute evaluations of self and other as the independent variables. The β weights resulting from these analyses are shown in Table 1.

2.3 Discussion

The results are as expected: BTA and WTA effects were stronger on subjective than objective measures, even in the presence of feedback. Objective measures showed

[2] This effect describes the fact that participants rated themselves more above average when they got relative feedback (*M* = .20, *SE* = .11) than when they got absolute feedback about self and others (*M* = -.03, *SE* = .12) or only absolute feedback about self (*M* = -.18, SE = .12). The higher comparative evaluation in the condition where participants received the clearest comparative information (the relative feedback condition) is attributable to the fact that participants in the present experiment, surprisingly, performed better than did those to whom they were asked to compare themselves. Participants in the comparison group got lower scores (*M* = 4.95, *SD* = 3.77) than did participants in Experiment 1 (*M* = 5.43, *SD* = 3.72), *F (1, 263) = 9.48, p = .002,* η^2 = .04. The two samples were drawn from the same population of students and there is no obvious explanation for this difference.

the smallest effect of task difficulty, whereas subjectively anchored scales showed conflation between absolute and relative evaluation.

The first experiment asked participants to compare themselves with others. It leaves open the question of whether the conflation effects documented might be due to egocentrism. Naturally, one's own performance (both relative and absolute) has consequences for self-evaluation and ego threat. These consequences may cloud the interpretation of the effects of feedback on self-judgment. Furthermore, participants' experience of ease or difficulty on the test might more easily influence their assessment of how well they did on a subjective scale than on an objective scale. In order to rule out these potentially complicating factors, Experiment 2 sought to replicate the conflation effect in a context where the self was not relevant.

3 Experiment 2: Ten targets

Experiment 2 again tests the hypothesis that assessments of relative and absolute performance are more frequently mixed up on subjective than on objective response scales. Again, conflation predicts that the response scale should matter, whereas theories of differential weighting and differential information are silent on the matter and offer no such prediction.

3.1 Method

Participants. Participants were 64 individuals who participated in exchange for being paid. Participants had a mean age of 24 years (*SD* = 7.5 years) and 61 percent of them were male. None of the participants in this experiment had participated in the first experiment.

Procedure. Participants were shown one of two ten-question trivia quizzes (see Appendix A) and the scores of ten actual people who had taken that quiz. The experimental manipulation varied the difficulty of the quiz. The ten scores were chosen from a broader sample of individuals who had taken the quiz as a part of a separate study. The ten scores were selected to be representative of the group, such that their means and standard deviations were approximately the same as the broader sample. Scores on the simple quiz had a mean of 8.6 (*SD* = 1.3) out of 10. Scores on the difficult quiz had a mean of 1.5 (*SD* = 1.3).

Participants were asked to rate the relative performance of each of the ten quiz-takers. First, they were reminded of the first quiz-taker's score and were asked, "*How did Person 1 do on the trivia quiz relative to the whole group?*" They were given a subjective response scale that ran from 1 (*well below average*) to 7 (*well above average*). The midpoint (4) was labeled *average*. Second, for each of the ten quiz-takers, participants were also asked to estimate that person's percentile ranking: "*What percentage of the group had scores below Person 1's score?*"

3.2 Results and discussion

Participants' subjective ratings and percentile rankings of the ten target individuals were standardized by converting them to z-scores. The ten ratings and the ten percentile rankings were then each averaged and subject to a 2 X (2) mixed ANOVA. The first factor was the between-subjects manipulation of quiz difficulty. The second factor was the within-subjects manipulation of the clarity of the response scale. The results reveal a significant effect of quiz difficulty, F (1, 62) = 20.36, $p < .001$, $\eta^2 = .25$. Participants evaluating targets who had taken the simple quiz rated them more above average (*M* = .12, *SE* = .04) than did participants evaluating the targets who had taken the difficult quiz (*M* = –.12, *SE* = .04).

More importantly, this main effect is qualified by a significant interaction between measure and difficulty, F (1, 62) = 9.11, p = .004, $\eta^2 = .13$. Consistent with the conflation explanation, on the subjective scale those who had taken the easy quiz got higher ratings (*M* = 4.50, *SD* = .62) than did those who had taken the difficult quiz (*M* = 3.74, *SD* = .44). Indeed, ratings given to those who had taken the simple quiz were significantly above the midpoint rating of 4, t (30) = 4.49, $p < .001$, $\eta^2 = .40$, whereas those ratings of those who had taken the difficult quiz were significantly below 4, t (32) = 3.34, $p < .01$, $\eta^2 = .26$. By contrast, participants gave similar percentile rankings to targets who had taken the simple (*M* = 42.5, *SD* = 11.82) and difficult quizzes (*M* = 39.06, *SD* = 5.85), F (1, 62) = 2.22, $p = .14$, $\eta^2 = .04$.

Another way of comparing the two types of dependent measures is to ask which one shares more variance with absolute performance. The ten targets' actual absolute scores are more strongly correlated with participants' ratings on the subjective comparative scale ($r = .48$) than with participants' estimated percentile rankings ($r = .33$), and these two correlations are significantly different, t (61) = 2.26, $p < .05$, $\eta^2 = .08$. Both measures are highly correlated with targets' actual relative performance (r = .84 and .81, respectively), and these correlations are not significantly different from each other, t (61) = .70, *ns*.

The differing results for the two dependent measures are striking because both questions asked roughly the same thing.[3] The primary difference between the two was

[3] There is a difference between ratings relative to average and percentile ranks that I ought to note. Although the majority of individuals will be above average in a negatively skewed distribution (as is the case in the simple condition) and the majority will be below average in positively skewed distribution (as is the case in the difficult condition), the subjective rating measure asked participants to report *how much* above or below average each individual was. If participants are using the sub-

ambiguity of the response scales. When the points on the response scale were labeled with subjectively interpretable words, participants' judgments were influenced by both absolute and the relative performance, despite the fact that the question specifically asked them to assess comparative standing. When the response scale was specified with greater precision, this effect is reduced, as shown by the non-significance of the effect of quiz difficulty on percentile rankings.

Can differential weighting explain these results? The clearest theoretical articulation of differential weighting hypothesis has been offered by Windschitl, Rose, Stalkfleet, and Smith (2007). They formalize differential weighting in a manner quite consistent with Giladi and Klar's (2002) LOGE explanation. The LOGE explanation holds that the BTA and WTA effects are attributable to the fact that even when people are asked to compare an individual member of a select group to that group (and make a LOcal comparison), they fail to ignore the broader and more representative (GEneral) sample. Instead, they consider both the local and the general samples and compare the individual to a standard that is a compromise between the local and the general samples. The BTA and WTA effects from the first two experiments are consistent with the LOGE explanation. The feature of the present data that the LOGE theory cannot account for is that effect of question format. The LOGE theory would not predict the difference between objective and the subjective measures of comparative judgment that we observe, because the LOGE theory is silent on the issue of elicitation formats.

4 Experiment 3: Two targets

A number of studies have found that better-than-average effects are strongest when people compare themselves to a group average, and are reduced or eliminated when they compare themselves to an individual (Alicke et al., 1995; Hoorens & Buunk, 1993; Klar et al., 1996; Perloff & Fetzer, 1986). In order to eliminate this individual-to-group comparison issue as an alternative explanation for the findings of Experiment 2, participants were asked to compare two individuals. In addition, in order to address concerns about asking the same question more than once, participants only made one comparative judgment for each target individual.

The third experiment again tests the hypothesis that assessments of absolute and relative performance are conflated on subjective response scales. Again, this hypothesis distinguishes the conflation explanation from theories of differential weighting and differential information, which are silent on the issue of the influence of the response scale.

4.1 Method

Participants. Participants were 66 individuals who participated in exchange for payment. None of the participants in this experiment had participated in either of the first two experiments. Participants had a mean age of 22 years (*SD* = 2.1 years) and 59 percent of them were male.

Procedure. The procedure was similar to Experiment 2. However, instead of seeing ten quiz-takers, each participant only saw two. Again, these two were selected to be roughly representative of all those who had taken the quizzes previously. Those participants who saw the easy quiz then evaluated scores of 8 and 9. Those who saw the difficult quiz evaluated scores of 1 and 2. For each target, participants were asked a single question: "*How did Person 1 do on the trivia quiz relative to Person 2?*" All participants responded on a subjective scale that ran from 1 (*much worse*) to 7 (*much better*). The midpoint (4) was labeled *same*.

4.2 Results and discussion

Participants' ratings of the two target individuals were averaged and subjected to a one-way ANOVA on difficulty. The results reveal the predicted effect of test difficulty, F (1, 64) = 7.41, p = .008, η^2 = .10. Participants who were rating those who had taken the simple test gave higher ratings (M = 4.14, SD = .38) than did participants who were rating those who had taken the difficult test (M = 3.91, SD = .29). Ratings given to those who had taken the simple quiz were marginally above the rating that would indicate an average performance (4), t (32) = 1.79, p = .08, η^2 = .09, whereas those who had taken the difficult quiz were rated below average, t (32) = 2.06, $p < .05$, η^2 = .12.

The results of Experiment 3 show that even when they compare two individuals to each other — about whom they have complete information — people make the error of being influenced by the target's absolute performance when making comparative evaluations. These results cannot be explained by problems comparing an individual to a group or by the LOGE account. They cannot be explained by differential information since participants had perfect information about both target and referent. Differential weighting is an unlikely explanation, given that participants did not have much of a reason to focus on one or the other target. The best explanation for these results is that participants confuse absolute with relative evaluation on subjective rating scales.

jective scale similarly for all ten targets, the amounts above and below average ought to be exactly equal, leading to the average difference from average to be zero within any given population.

5 Experiment 4: Two targets, lots more questions

I have not yet addressed the important issue of whether the conflation effect occurs at the encoding or response stage. If conflation occurs when people encode information mentally, then they will not be able to retrieve and report unconflated evaluations. The evidence from the first three experiments, demonstrating differences between subjective and objective measures, suggests that conflation occurs during response. If the conflation effect disappears in objective measures of comparative evaluation, the clear implication is that conflation occurs in responding, not at encoding, and that conflation is a direct result of confusion over how exactly to construe the meaning of the question or response scale. However, we ought to be concerned that the first three experiments did not systematically vary the order in which participants were asked subjective and objective questions. It is possible that the conflation effect occurs when subjective questions follow objective questions, as they did in the first two experiments.

In order to address these concerns, Experiment 4 systematically varies question order. In addition, Experiment 4 standardizes question format. Relative evaluations are elicited using the same question ("*How did Person A do on the trivia quiz, compared to Person B* ?") and the same 11-point response scale; the difference between objective and subjective measures is exclusively whether the points on the scale are labeled with words or with numbers (see Budescu & Wallsten, 1987; Budescu, Weinberg, & Wallsten, 1988; Wallsten, Budescu, & Erev, 1988). Likewise, absolute evaluations are elicited using the same question ("*How did Person A do on the trivia quiz?*"), varying only whether the 7-point response scale was labeled with words or numbers. Note that absolute and relative evaluations are elicited using different scales in order to address the possibility that conflation is facilitated by similarity in their scales.

5.1 Method

Participants. Participants were 114 individuals who participated in exchange for course credit.

Procedure. The procedure was similar to Experiment 3. Participants saw either an easy or a difficult trivia quiz, and then two quiz performances that were selected to be roughly representative of all those who had previously taken that quiz. Those participants who saw the easy quiz then evaluated scores of 8 and 9. Those who saw the difficult quiz evaluated scores of 1 and 2. For each target person, participants answered two questions:

1. "*How did Person A do on the trivia quiz?*"
2. "*How did Person A do on the trivia quiz, compared to Person B?*"

They answered each of these questions twice with different response scales. To answer the first question, participants were provided with an 11-point response scale that ran from 0 to 10. The response scales had labels at the endpoints and at the midpoint. One of the times they answered this question, the scale had labels of *very badly*, *moderately*, and *very well.* The other time they answered this question, the scale had labels of *zero correct*, *five correct,* and *ten correct.*

To answer the second question, "*How did Person A do on the trivia quiz, compared to Person B?*" participants were provided with a response scale that ran from 1 to 7. One of the times they answered this question, the scale had labels of *much worse*, *same*, and *much better.* The other time they answered this question, the scale had labels of *3 points worse*, *same score*, and *3 points better.*

In sum, then, participants evaluated each target score in absolute terms and in comparison to the other score. And they made each of these judgments twice, once using a subjective verbally-anchored response scale and once using a more objective numerically-anchored response scale.

Design. The experiment had a 2 X 2 X 2 X 2 between-subjects design. The first factor varied whether the two targets' scores came from the simple or the difficult quiz. The other three experimental factors varied order. The first order manipulation varied whether participants first evaluated the lower of the two target scores (i.e., 1 or 8) or the higher of the two (i.e., 2 or 9). The second order manipulation varied whether, for each target, participants first answered the two questions with subjective response scales or the two with objective response scales. The third order manipulation varied whether participants first assessed absolute performance or comparative performance.

5.2 Results and discussion

Comparative evaluations. In order to test the hypothesis that, when all performances were low (because the test was difficult), targets would be rated below average, but only on subjective measures, I averaged the comparative ratings for targets A and B separately for the two subjective and the two objective measures of comparative evaluation. These two averages were then subjected to a 2 X 2 X 2 X 2 X (2) mixed ANOVA with repeated measures on the last factor. The first four factors are the between-subjects manipulations and the last factor is measure (subjective vs. objective). The results reveal a main between-subjects effect of difficulty, F (1, 98) = 12.21, p = .001, η^2 = .11. This effect describes that tendency for participants to rate easy test scores as better than each other (*M* = 4.16, *SD* = .30) more so than difficult test scores (*M* = 3.97, *SD* = .30). In other words, when they were rating easy test scores, target A was more likely to be rated as better than target B, and target B was more likely to be rated as better than target A, than they were when participants were rating difficult test scores. Of course, if people were making these comparative ratings sensibly, the mean should come out to be 4 for both the subjective and objective scales.

However, this main effect is qualified by the expected difficulty X measure interaction effect, *F* (1, 98) = 10.05, *p* = .002, η^2 = .09. This interaction effect arises because, when they were using the subjective rating scale, participants who saw two performances from the easy quiz were more likely to rate them has better than each other (*M* = 4.27, *SD* = .50) than were those who saw two performances from the difficult quiz (*M* = 3.96, *SD* = .43), *F* (1, 112) = 12.79, *p* = .001, η^2 = .10. On the objective rating scale, however, participants did not differ in their comparative evaluations of those who had taken the easy (*M* = 4.03, *SD* = .24) and the difficult (*M* = 3.97, *SD* = .14) quizzes, *F* (1, 112) = 2.00, *p* = .16, η^2 = .02.

No other effects emerged as statistically significant in this 2 X 2 X 2 X 2 X (2) ANOVA on comparative performance ratings, all *F*s < 3.5, all *p*s > .06.

The conflation explanation posits that comparative evaluations on a subjective scale are easily confused with absolute individual evaluations. Consistent with this notion, the two targets' actual *absolute* scores are more strongly correlated with participants' ratings of *relative* performance on the *subjective* scale (*r* = .24) than on the *objective* scale (*r* = .15), *t* (111) = 2.49, *p* < .05, η^2 = .06. Moreover, the targets' actual *relative* scores are more strongly correlated with participants' ratings of *relative* performance on the *objective* scale (*r* = .91) than on the *subjective* scale (*r* = .78), *t* (111) = 8.07, *p* < .001, η^2 = .37.

Absolute evaluations. In order to examine the effect of question subjectivity on absolute evaluations, I averaged absolute ratings for targets A and B separately for the two subjective and the two objective measures of absolute evaluation. These two averages were then subjected to a 2 X 2 X 2 X 2 X (2) mixed ANOVA with repeated measures on the last factor. Again, the last factor is measure (subjective vs. objective). Naturally, the results reveal a main effect for difficulty, as scores from the easy tests received higher absolute ratings (*M* = 8.20, *SE* = .09) than did scores from the difficult test (*M* = 2.08, *SE* = .09), *F* (1, 98) = 2160, *p* < .001, η^2 = .96. The main within-subjects effect of subjectivity also emerged as significant, because both quiz scores received more positive evaluations on the subjective scale (*M* = 5.28, *SD* = 2.98) than on the objective scale (*M* = 4.93, *SD* = 3.50), *F* (1, 98) = 5.32, *p* = .023, η^2 = .05.

However, both these main effects are qualified by the subjectivity X difficulty interaction predicted by the conflation account, *F* (1, 98) = 43.99, *p* < .001, η^2 = .31. This interaction results from the fact that while for easy quiz scores, the subjective rating scale led to *lower* ratings (*M* = 7.94, *SD* = 1.11) than did the objective scale (*M* = 8.46, *SD* = .21), *t* (55) = -3.71, *p* < .001, η^2 = .20, for difficult quiz performances the subjective rating scale led to *higher* ratings (*M* = 2.71, *SD* = 1.66) than did the objective scale (*M* = 1.51, *SD* = .07), *t* (57) = 5.47, *p* < .001, η^2 = .34. In other words, comparative and absolute evaluations were more similar to each other when made using subjective scales than they were when objective scales were used for the identical judgment.

This two-way interaction is qualified by a significant three-way interaction between subjectivity, difficulty, and subjective/objective order, *F* (1, 98) = 7.06, *p* = .009, η^2 = .07. Contrast tests reveal that this interaction results from the fact that the subjectivity X difficulty interaction, although consistent in its pattern across order conditions, was significantly stronger when the subjective questions came first, *F* (1, 54) = 31.19, *p* < .001, η^2 = .37, than when the objective questions came first, *F* (1, 56) = 14.49, *p* < .001, η^2 = .21. The obvious explanation for this finding is that when the objective measures came first, they tended to discipline the subsequent subjective measures. However, when the subjective measures came first, they were most easily conflated with other standards of evaluation.

No other effects emerged as significant in the 2 X 2 X 2 X 2 X (2) ANOVA on absolute performance ratings, all *F*s < 3.32, all *p*s > .07.

Consistent with the conflation explanation, the targets' actual *relative* scores are marginally more strongly correlated with participants' ratings of *absolute* performance on the *subjective* scale (*r* = .20) than on the *objective* scale (*r* = .14), but the difference between these correlations does not attain statistical significance, *t* (111) = 1.82, *p* < .10, η^2 = .03. Moreover, targets' actual *absolute* scores are more strongly correlated with participants' ratings of *absolute* performance on the *objective* scale (*r* = .99) than on the *subjective* scale (*r* = .87), *t* (111) = 48.65, *p* < .001, η^2 = .98. These results suggest that objectively labeled scales produce more accurate measures of both relative and absolute evaluation than do subjectively labeled scales.

Conflation in indirect comparative judgments. So why doesn't conflation also affect indirect comparative judgments? The answer is that, although conflation does affect absolute evaluation on subjective response scales, it affects evaluations of target and referent similarly, and these effects cancel each other out when one is subtracted from the other to compute the indirect comparison.

6 General discussion

The four experiments presented here each finds support for the key predictions of the conflation explanation put forth in this paper. The experiments offer a series of successively more stringent tests that rule out a number of alternative explanations for BTA and WTA effects. In the end, conflation is the most viable explanation for the persistent BTA and WTA effects observed on subjective response scales across all four experiments. Specifically, the evidence suggests that conflation occurs during the response process during which people attempt disambiguate subjective response scales so that they can translate their knowledge into a rating on the scale. Had conflation occurred at encoding, then people would not be able to retrieve an unconflated judgment to provide on objective measures.

The results demonstrate that BTA and WTA biases manifest themselves most strongly on subjective measures of relative evaluation. The effects weaken on clearer, more objective measures, suggesting that people can make less biased judgments when provided clearer response formats. It is also obvious that BTA and WTA effects grew weaker across the four experiments, strongly suggesting that the effect has a number of causes in addition to conflation. The last two experiments most effectively ruled out explanations other than conflation, and they also produced the smallest effects. However, conflation was at work in those prior studies, helping to increase the size of both better-than-average and worse-than-average effects.

6.1 Conflation's broader consequences

There are important implications for these findings outside the experimental laboratory. There are many contexts in which comparative judgments have important consequences and in which ambiguous subjective rating scales are the norm. For example, it is routine in the corporate world for promotions and raises to be contingent on performance reviews in which the most important criterion is a manager's evaluation on a verbally-anchored scale, using endpoints with labels like *unacceptable performance* and *exceptional performance*. Biernat has shown the surprising consequences that implicit references groups can have on evaluation using subjective response scales (Biernat, 2003; Biernat & Vescio, 2002). For instance, when two athletes of different genders perform similarly, the woman is rated more positively than the man because the interpretation of the scale on which each player is rated is disproportionately defined by stereotypes of the groups from which the candidate comes. A woman can seem particularly impressive compared to the stereotype of women as unathletic. However, when it comes to allocating limited resources such as selecting team members, men are nevertheless more likely to be selected (Biernat & Vescio, 2002).

Baron (1997) has pointed out that the conflation of absolute with relative evaluation is reflected in a number of other mistakes made by people who ought to know better:

> Newspapers often tell us that "inflation increased by 2.9%" when they mean that prices increased by that much. The literature on risk effects of pollutants and pharmaceuticals commonly report relative risk, the ratio of the risk with the agent to the risk without it, rather than the difference. Yet, the difference between the two, not their ratio, is most relevant for decision making: if the risk is miniscule, a high relative risk still means very little. (p. 302).

Various other research findings show the ways in which people conflate relative with absolute evaluations. Sometimes, people attend to absolute numbers when relative proportions are more meaningful. For example, Denes-Raj and Epstein (1994) found that people prefer to bet on an urn with 9 winning chips out of 100 rather than an urn with 1 winning chip out of 10. Similarly, people are more suspicious of sex discrimination on the grading of an exam when the man who got the highest grade was the only male among 10 test-takers than when he was one of the 10 males in a group of 100 test-takers (Miller, Turnbull, & McFarland, 1990).

Other findings highlight the ways in which people attend to relative performance or to proportions when absolute counts are more meaningful. For example, Klein has found that willingness to change driving habits is more strongly influenced when people are told that their risk of accident is 20% above or below average than when they are told that their lifetime risk of being in an accident is 30% or 60% (Klein, 1997, 2002). In a related vein, people are more willing to contribute to the search for a cure that will cure 90% of sufferers than one that will cure 9% of sufferers, holding constant the number of people cured (Baron, 1997). One result of such thinking may be that governments spend a far greater amount of money for each human life saved on risks affecting few people — such as chemical spills or radiation exposure — than on risks that affect many people — such as road safety (McDaniels, 1988). This same reasoning has been implicated in the perceived futility of working to reduce world hunger: Even though helping a large number of people is relatively easy and inexpensive, it can only be a "drop in the bucket" compared to the total problem (Unger, 1996). Clearly, the conflation of relative and absolute quantities can have profound consequences.

6.2 Concluding comments

The present study offers a cautionary note to psychologists who employ subjectively labeled scales in their research. Subjective scales have many virtues, the greatest of which may be that they can be used even without common scaling rules because experimenters do not have to specify a mapping of the numbered scale on to objectively verifiable quantities. We can ask participants, "*How angry do you feel?*" and let respondents determine, by their own idiosyncratic definitions, whether they were "*somewhat angry*" or "*extremely angry*." Even if different people interpret the scale differently, the resulting responses are psychometrically meaningful and are useful for inferring more objective absolute evaluations (Dawes, 1977; Stewart, Brown, & Chater, 2005; Thurstone, 1927).

But subjective interpretation is also a great weakness of subjective response scales. Since respondents decide how to assign their subjective perceptions to numbers on a scale, their responses are influenced by a wide variety of contextual factors (Schwarz, 1999; Schwarz & Hippler, 1995). As long as these contextual factors impinge randomly and similarly on all experimental conditions, they do not pose any problems for internal validity. However, as the findings of the present paper highlight, there is the potential for contextual influences to be confounded with the independent variables we are interested in studying.

The key contribution of the conflation explanation is to point out the crucial importance of the clarity of the dependent measures. The results presented here show that objective comparative measures reduce the size of BTA and WTA effects. Whatever role differential weighting has that is distinct from conflation, it is clear that it is smaller than previously assumed. It would appear that ambiguous measures, which promote conflation, may be responsible for a good deal of the better-than-average and worse-than-average effects observed heretofore.

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.

Baron, J. (1997). Confusion of relative and absolute risk in valuation. *Journal of Risk and Uncertainty, 14*, 301–309.

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58*(12), 1019–1027.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66*, 5–20.

Biernat, M., Manis, M., & Kobrynowicz, D. (1997). Simultaneous assimilation and contrast effects in judgments of self and others. *Journal of Personality and Social Psychology, 73*, 254–269.

Biernat, M., & Vescio, T. K. (2002). She swings, she hits, she's great, she's benched: Implications of gender-based shifting standards for judgment and behavior. *Personality and Social Psychology Bulletin, 28*, 66–77.

Budescu, D. V., & Wallsten, T. S. (1987). Subjective estimation of precise and vague uncertainties. In G. Wright & P. Ayton (Eds.), *Judgmental forecasting* (pp. 63–82). Oxford, England: John Wiley & Sons.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 281–294.

Burson, K. A., & Klayman, J. (2005). *Judgments of performance: The relative, the absolute, and the in-between*. Ann Arbor: Unpublished manuscript. Available at SSRN: http://ssrn.com/abstract=894129

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 and Social Psychology Bulletin, 29*(11), 1343–1356.

Dawes, R. M. (1977). Suppose we measured height with rating scales instead of rulers. *Applied Psychological Measurement, 1*, 267–273.

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology, 66*, 819–829.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and business. *Psychological Science in the Public Interest, 5*, 69–106.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082–1090.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*, 519–527.

Gannon, K. M., & Ostrom, T. M. (1996). How meaning is given to rating scales: The effects of response language on category activation. *Journal of Experimental Social Psychology, 32*, 337–360.

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.

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 and Social Psychology Review, 5*, 74–95.

Hoorens, V., & Buunk, B. P. (1993). Social comparison of health risks: Locus of control, the person-positivity bias, and unrealistic optimism. *Journal of Applied Social Psychology, 23*, 291–302.

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., & Giladi, E. E. (1999). Are most people happier than their peers, or are they just happy? *Personality and Social Psychology Bulletin, 25*, 585–594.

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

Klein, W. M. P. (1997). Objective standards are not enough: Affective, self-evaluative, and behavioral responses to social comparison information. *Journal of Personality and Social Psychology, 72*, 763–774.

Klein, W. M. P. (2001). Post hoc construction of self-performance and other performance in self-serving social comparison. *Personality and Social Psychology Bulletin, 27*, 744–754.

Klein, W. M. P. (2002). Comparative risk estimates relative to the average peer predict behavioral intentions and concern about absolute risk. *Risk Decision and Policy, 7*, 193–202.

Klein, W. M. P., & Buckingham, J. T. (2002). Self-other biases in judgments of ambiguous performance and corresponding ability. *Psychologica Belgica, 42*, 43–64.

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., Windschitl, P. D., Burrus, J., Fessel, F., & Chambers, J. R. (in press). The rational side of egocentrism in social comparisons. *Journal of Experimental Social Psychology*.

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior & Human Decision Processes, 102*, 76–94.

Malmendier, U., & Tate, G. (2004). *CEO overconfidence and corporate investment*. Cambridge, MA. http://www.nber.org/papers/w10807.pdf

McDaniels, T. L. (1988). Comparing expressed and revealed risk preferences for risk reduction: Different hazards and different question frames. *Risk Analysis, 8*, 593–604.

Miller, D. T., Turnbull, W., & McFarland, C. (1990). Counterfactual thinking and social perception: thinking about what might have been. *Advances in Experimental Social Psychology, 23*, 305–331.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior & Human Decision Processes, 103*, 197–213.

Moore, D. A., & Healy, P. J. (2007). *The trouble with overconfidence*. Pittsburgh: Tepper Working Paper 2007-E17.

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

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative social judgment: On being both better and worse than we think we are. *Journal of Personality and Social Psychology, 92*, 972–989.

Moore, D. A., & Small, D. A. (in press). When it's rational for the majority to believe that they are better than average. In J. I. Krueger (Ed.), *Rationality and social responsibility: Essays in honor of Robyn M. Dawes*. Mahwah, NJ: Erlbaum.

Mussweiler, T., & Strack, F. (2000). The "relative self": Informational and judgmental consequences of comparative self-evaluation. *Journal of Personality and Social Psychology, 79*, 23–38.

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *Journal of Finance, 53*, 1887–1934.

Otten, W., & van der Pligt, J. (1996). Context effects in the measurement of comparative optimism in probability judgments. *Journal of Social and 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.

Rose, J. P., & Windschitl, P. D. (in press). How egocentric optimism change in response to feedback in repeated competitions. *Organizational Behavior & Human Decision Processes*.

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

Schwarz, N. (2001). Feelings as information: Implications for affective influences on information processing. In L. L. Martin & G. L. Clore (Eds.), *Theories of mood and cognition: A user's handbook* (pp. 159–176). Mahwah, NJ: Erlbaum.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*, 195–202.

Schwarz, N., Grayson, C. E., & Knaeuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research, 10*, 177–183.

Schwarz, N., & Hippler, H. J. (1995). Subsequent questions may influence answers to preceding question s in mail surveys. *Public Opinion Quarterly, 59*, 93–97.

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49*, 388–395.

Schwarz, N., Knaeuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570–582.

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

Svenson, O. (1981). Are we less risky and more skillful than our fellow drivers? *Acta Psychologica, 47*, 143–151.

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

Unger, P. (1996). *Living high and letting die: Our illusion of innocence*. Oxford: Oxford University Press.

Wallsten, T. S., Budescu, D. V., & Erev, I. (1988). Understanding and using linguistic uncertainties. *Acta Psychologica, 68*(1–3), 39–52.

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

Windschitl, P. D., Kruger, J., & Simms, E. (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.

Windschitl, P. D., Rose, J. P., Stalkfleet, M. T., & Smith, A. R. (2007). *Are people excessive or judicious in their egocentrism? A modeling approach to understanding bias and accuracy in people's optimism within competitive contexts*. Unpublished manuscript.

Appendix

Trivia questions used in the simple and difficult trivia quizzes.

	Simple	Difficult
1.	How many inches are there in a foot?	Which creature has the largest eyes in the world?
2.	What is the name of Pittsburgh's professional hockey team?	How many verses are there in the Greek national anthem?
3.	Which species of whale grows the largest?	What company produced the first color television sold to the public?
4.	Who is the president of the United States?	How many bathrooms are there in the White House (the residence of the U.S. President)?
5.	Harrisburg is the capital of what U.S. state?	Which monarch ruled Great Britain the longest?
6.	What was the first name of the Carnegie who founded the Carnegie Institute of Technology?	The word "planet" comes from the Greek word meaning what?
7.	How many states are there in the United States?	What is the name of the traditional currency of Italy (before the Euro)?
8.	What continent is Afghanistan in?	What is Avogadro's number?
9.	What country occupies an entire continent?	Who played Dorothy in "The Wizard of Oz"?
10.	Paris is the capital of what country?	Who wrote the musical "The Yeoman of the Guard"?
Tiebreaker question:	How many people live in Pennsylvania?	

Answers: Simple: (1) 12 (2) Penguins (3) Blue (4) George W. Bush (5) Pennsylvania (6) Andrew (7) 50 (8) Asia (9) Australia (10) France. Difficult: (1) Giant squid (2) 158 (3) RCA (4) 32 (5) Queen Victoria (6) wanderer (7) Lira (8) 6.02 X 10^{23} (9) Judy Garland (10) Gilbert & Sullivan. Tiebreaker: 12,281,054.