Emotion
2013, Vol. 13, No. 3, 366–374

1528-3542/13/$12.00 DOI: 10.1037/a0032041

BRIEF REPORT

The Thrill of (Absolute) Victory: Success Among Many Enhances Emotional Payoffs

Ed O'Brien and Linda Hagen
University of Michigan

The emotional value of placing in a given percentile of a competition (e.g., placing in the "top 10%") depends on how many competitors are involved. Five studies reveal that winning among larger groups is associated with more positive emotional reactions than winning among smaller groups, even when the objective chances for success are held constant. Participants thought that a runner would feel happier after placing in the top 10% in a race with many (vs. few) competitors (Experiment 1); participants who imagined placing in the top 10% of a trivia quiz predicted that they would feel happier after succeeding among many (vs. few) respondents (Experiment 2); and participants who were given randomly assigned false feedback that they placed in the top 10% of a real creativity challenge actually felt happier when the pool was described as containing many (vs. few) contestants (Experiment 3). This effect appears to be driven by participants' intuitions about the statistical law of large numbers: when people think about success among large pools, they infer that the outcome is more diagnostic of "true" abilities—that the performance must not be a fluke—compared with identical success among small pools, which provides an affective boost (Experiments 4–5).

Keywords: emotions, biases, group size, competition, probability

Imagine your manuscript has just been accepted for publication in a journal that accepts only 10% of submissions. This news undoubtedly feels good. But now suppose you discover that the journal received a record number of submissions this year. Despite the fact that the objective threshold for acceptance has remained unchanged, would your success suddenly seem more special? Might you feel happier than you would have had there been fewer submissions?

In this article, we explore the intersection between numerical reasoning and emotion, in an attempt to understand people's emotional reactions to events like the one above. Many real-life competitive contexts alert us to a given chance of success as well as to the absolute number of people involved. For example, researchers await decisions from journals, conferences, and funding agencies that announce not only the number of successful applicants but also the total tally of submissions; athletes vie for the highest ranks among a specified pool of competitors; and students and employees seek to score in top percentiles on tests and performance evaluations relative to a set number of peers.

Despite facing the same chance of success—defined here as holding constant the percentage of competitors who are considered to have succeeded—people's perceived and experienced emotions in such contexts might become amplified when they think about succeeding among larger (vs. smaller) groups. In purely statistical terms, the *law of large numbers* posits that large samples should contain distributions of values that better reflect a "real" population of possible scores (e.g., Sedlmeier & Gigerenzer, 1997; Tversky & Kahneman, 1971). To the extent that people have a sense of this property, they may infer that a larger group contains a more representative pool of competitors simply due to its size. Accordingly, winning in large groups should seem more diagnostic of "true" ability and less sensitive to random chance, which may afford emotional boosts.

Mukherjee and Hogarth (2010) provided a theoretical model in support of this claim. The authors highlight that people's skill level in a given domain varies within a population and that the larger a sample drawn from this distribution, the more likely it will include a representative range of talent. In other words, sampling error (i.e., chances that the drawn sample will consist of only individuals with the same degree of the skill, such as all low-talent individuals) necessarily decreases as pool size increases. Mukherjee and Hogarth (2010) argued that, as a result, "luck is more important in determining the winner in small samples than in large samples" (p. 746), and hence even nonexceptional competitors are able to win in small pools simply because of a fluke sample. In contrast, the likelihood that such individuals could place in a top percentile by

This article was published Online First April 29, 2013.

Ed O'Brien, Department of Psychology, University of Michigan; Linda Hagen, Ross School of Business, University of Michigan.

This research was supported by a National Science Foundation Graduate Research Fellowship awarded to the corresponding author. Alvin Chan, Taylor Rothman, and Sarah Lee assisted with data collection. We gratefully acknowledge Stéphane Côté for extremely helpful comments and suggestions.

Correspondence concerning this article should be addressed to Ed O'Brien, Department of Psychology, University of Michigan, 530 Church Street, Ann Arbor, MI 48109. E-mail: obrieneh@umich.edu

mere happenstance in large pools is (theoretically) much lower, decreasing their overall chances of winning.

This logic leads us to our hypothesis. Winners in large competitive pools should infer that it is improbable that they outperformed many others because of a mere fluke and that instead they must have done so by virtue of their (superior) skills. This inference, in turn, should boost their emotional reactions in comparison to winners in small pools, because they have more reason to count the win as diagnostic success—as attributable to "actual" ability. Providing related empirical evidence, participants in one study thought it was more difficult to succeed among large versus small pools even when their probability of success remained constant (Garcia & Tor, 2009, Study 5). Although yet to be tested in terms of emotion, it follows that victory may seem even sweeter if people actually compete and succeed among many (vs. few) competitors, because winning among many cannot be easily explained away as "illegitimate." Placing in the same percentile in the same competition could seem more thrilling simply when many others are involved.

Much research has examined the influence of numbers and statistical reasoning on people's initial evaluations of competitive contexts (see Festinger, 1954; Taylor & Lobel, 1989; Tesser, 1988). But to our knowledge, none has explored downstream consequences for emotion after competing. Five studies tested whether, all else being equal (including the threshold for success), winning among many is associated with more positive emotional reactions compared with identical success among few. The first two studies tested for this association when people imagine someone else's (Experiment 1) and their own (Experiment 2) success in hypothetical contexts; the next study measured people's actual reactions after competing (Experiment 3); and the final two studies tested whether these boosts are indeed driven by the extent to which people infer "true abilities" from the performance (Experiments 4–5). Whereas previous work has overwhelmingly focused on cognitive processes and outcomes (e.g., mental accounting or risk assessment: see Garcia & Tor, 2009; Koehler, 1996; Slovic, 1986; Tversky & Kahneman, 1980), in these studies we sought to extend such principles in a novel way to emotional experiences.

Experiment 1

Athletics

In Experiment 1, people were asked to think about a runner named Steve. They were randomly assigned to read that Steve placed in the top 10% of a race that contained few or many competitors and then estimated how happy he felt.

Method

Participants. Eighty online participants were recruited via Amazon's Mechanical Turk in exchange for \$0.10. In this and all studies, we assessed three demographic variables: age, sex, and ethnicity (M_{age} = 34.40, 61.3% female, 73.8% Caucasian). Given that our hypothesis involves numerical reasoning, we also measured education level in this and the next two studies via a multiple-choice question about participants' highest attainment ("I received a high school degree"; "I received a high school degree and some college, but didn't graduate"; "I received a high school degree and college degree"; or "I received a high school degree, college degree, and have at least some postcollege education"). In Experiment 1, 88.7% reported some college or more. The sample was restricted to U.S. residents with a 95% or higher Amazon Turk approval rating.

Procedure. Participants completed a study that ostensibly examined hypothetical reasoning. First, they were asked to read the following passage about Steve, a runner who competed in a marathon between few (20) or many (20,000) competitors:

> Steve decided to run in his city's summer marathon. In preparation, he made sure to train each morning for the month leading up to the race. This year, 20 [20,000] runners have signed up to compete, including Steve. The top 10% of runners are awarded a medal. It turns out that Steve does finish in the top 10% of among the group of 20 [20,000].

Next, participants predicted Steve's emotional response by rating how happy (−5 = *very unhappy* to +5 = *very happy*), proud (−5 = *very ashamed* to +5 = *very proud*), thrilled (−5 = *very upset* to +5 = *very thrilled*), worthy (−5 = *very worthless* to +5 = *very worthy*), and positive (−5 = *very negative* to +5 = *very positive*) he felt. These items were combined into a "Positive Feelings" index (α = .92), which served as our primary dependent measure.

Participants also rated how prestigious they thought the race was (1 = *not at all* to 6 = *very*) to account for a potential alternative explanation of the findings: the possibility that a larger race seems more important than a smaller race, which may lead people to believe that Steve would feel better about his success for reasons beyond size. For example, we hoped not to compare people's beliefs about winning the prestigious Boston Marathon (which happens to be very large) with winning an unremarkable neighborhood contest (which happens to be very small); in other words, we sought to hold constant people's differential inferences about the race itself to examine group size as the unique manipulated variable across an otherwise identical competition. Finally, participants were asked to rate the size of the race as a manipulation check (1 = *very small* to 6 = *very large*).

Results and Discussion

As expected, participants thought the race was larger when it contained 20,000 runners (M = 5.58, SD = 0.64) versus 20 (M = 2.65, SD = 1.29), $t(78) = -12.85$, $p < .001$, d = 2.88. More important, despite the same threshold for success, participants predicted that Steve would feel better after placing in the top 10% of many runners (M = 4.62, SD = 0.60) compared with placing in the top 10% of few (M = 3.87, SD = 1.91), $t(78) = -2.37$, p = .02, d = 0.53. A univariate analysis of covariance (ANCOVA) with sex, age, ethnicity, and education level added simultaneously as covariates revealed that the effect of group size on prediction of Steve's happiness remained significant when controlling for these demographic variables, $F(1, 74) = 4.78$, p = .032. Further, neither sex (p = .29), age (p = .98), ethnicity (p = .99), nor education level (p = .19) exerted a significant effect. These patterns held across all subsequent studies and, thus, they are not discussed further.

In addition, although participants did rate the race as more prestigious when it contained many runners (M = 4.83, SD = 1.01) than when it contained few (M = 2.83, SD = 1.13), $t(78)$ = −8.35, $p < .001$, the effect of group size on predictions of Steve's happiness remained significant when controlling for prestige as a covariate, $F(1, 77) = 4.69$, $p = .033$, and ratings of prestige did not exert a significant effect ($p = .51$).

These findings support the hypothesis that winning in larger groups is associated with more positive reactions than equivalent success in smaller groups. Participants thought a target person felt happier after succeeding among many than among few, even though the threshold for success was held constant. Moreover, this effect did not appear to be driven by the possibility that participants were simply comparing a very prestigious competition with a less prestigious competition; although the large competition was indeed rated as more prestigious than the small competition, ratings of prestige did not alter the effect of pool size on emotion judgment.

The next study sought to extend these findings in two important ways. First, people predicted *their own* reactions. Second, we selected a different domain to help further address any differences caused by our manipulation beyond size (the "Boston Marathon problem"). Rather than using a scenario about a fictional event—which is ambiguous as to what it is associated with, its prominence, and so forth—the next study included an explicit description of a *New York Times* contest. Thus, no matter the size of the pool, all people imagined the same event sponsored by the same well-known source.

Experiment 2

Intellect

In Experiment 2, people imagined participating in a trivia contest sponsored by the *New York Times* and predicted how they would feel if they placed in the top 10% of a small or large group of respondents.

Method

Participants. Sixty online participants ($M_{age} = 32.65$, 66.7% female, 70% Caucasian, 78.4% some college or more) were recruited via Amazon's Mechanical Turk in exchange for $0.05. The sample was restricted to U.S. residents with a 95% or higher Amazon Turk approval rating.

Procedure. As in Experiment 1, participants were randomly assigned to complete an alleged study about hypothetical reasoning. First, they were asked to imagine that they had decided to take a trivia quiz on the *New York Times* website where they competed among few (20) or many (20,000) respondents. They read the following scenario:

> The *New York Times* publishes a "Millennium" quiz on its website, a large set of tricky questions about various topics that can be solved with a bit of savvy internet research. Scores are based on a combination of the number of correct answers and answer speed. Today, 20 [20,000] respondents take the quiz, including you. Each day the website automatically calculates who scores in the top 10%, earning the title "most clever." It turns out that you are in top 10% among the group of 20 [20,000].

Participants then predicted their reactions using the same Positive Feelings index that was used in Experiment 1 ($\alpha = .95$) and rated the size of the pool as a manipulation check (1 = *very small* to 6 = *very large*).

Results and Discussion

Again, participants thought the pool was larger when it contained 20,000 respondents ($M = 4.53$, $SD = 1.38$) versus 20 ($M = 3.03$, $SD = 1.32$), $t(58) = -4.29$, $p < .001$, $d = 1.11$. Moreover, despite the same threshold for success, participants predicted they would feel better placing in the top 10% of many quiz takers ($M = 3.77$, $SD = 1.49$) versus the top 10% of few ($M = 2.33$, $SD = 2.51$), $t(58) = -2.70$, $p = .009$, $d = 0.70$.

These findings extend the results of Experiment 1 to judgments of self within a different domain. It is important to note that it seems unlikely that the effect was driven by inferences about the scenario beyond size. All participants imagined completing the same survey at their own leisure sponsored by the same well-known organization, helping to isolate the influence of our manipulation. The third study went beyond hypothetical scenarios to test whether the effect extends to people's own real-time emotional reactions. Moreover, we tested yet a different competitive domain that could further isolate the causal role of size beyond any incidental inferences about large versus small competitive contexts, as well as to help establish the generalizability of the effect across a wider variety of settings.

Experiment 3

Creativity

In Experiment 3, people actually competed in a "creative reasoning challenge." They were randomly assigned to receive false feedback that they placed in the top 10% among a small or large pool of competitors, then rated how happy they currently felt.

Method

Participants. Sixty-six online participants ($M_{age} = 31.88$, 50% female, 81.8% Caucasian, 87.9% some or more college) were recruited via Amazon's Mechanical Turk in exchange for $0.20. The sample was restricted to U.S. residents with a 95% or higher Amazon Turk approval rating.

Procedure. Participants were invited to take a "creative reasoning challenge" as part of an ongoing university project on creativity. They completed 10 remote associate tests (Mednick, 1962), in which they were asked to generate connector words among sets of target words (e.g., cream/skate/water = *ice*). The tests were taken from a validated set of remote associate problems that were found to be equally difficult based on normative comparison data (Bowden & Jung-Beeman, 2003: see Table 1). Participants were told that scores were computed from their number of attempts, spelling, response time, and correct answers. In reality, participants were randomly assigned to receive false feedback that they placed in the top 10% among our database of 20 [20,000] respondents, then reported their reaction using the Positive Feelings index from Experiments 1 and 2 ($\alpha = .95$) and rated the size of the pool as a manipulation check (1 = *very small* to 6 = *very large*).

It is important to note that because feedback was disconnected from any actual performance, participants also rated how much they believed their score (1 = *not at all* to 6 = *very much*). This item was included to help account for possible cases in which people's own sense of their performance did not match the false feedback. For example, participants who felt confident that they made many mistakes would nonetheless have been told that they scored in the top 10% of the pool, which could seem hard to believe and hence influence their emotional reactions. Indeed, participants may have varied a great deal regarding how much they believed the outcome. Accordingly, we controlled believability as a covariate in ANCOVA for all subsequent analyses—an effective statistical technique used to address false feedback manipulations (e.g., McFarlin & Blascovich, 1984).

Results and Discussion

As expected, participants thought the database was larger when it contained 20,000 competitors ($M = 4.45$, $SD = 1.36$) versus 20 ($M = 2.20$, $SD = 1.28$), $t(64) = -6.92$, $p < .001$, $d = 1.70$. Moreover, despite the same threshold for success, participants reported actually feeling better after placing in the top 10% of many people ($M = 2.65$, $SD = 1.54$) compared with the top 10% of few ($M = 2.02$, $SD = 2.13$), $F(1, 65) = 4.21$, $p = .044$, $\eta^2 = .06$.[1]

These results provide strong support for the hypothesis that success among larger pools is associated with emotional boosts. Participants reported feeling happier after succeeding among many others compared with identical success among few—in real time. This effect was unlikely driven by qualitative differences beyond group size because all participants were explicitly presented with the same task for the same university project.

Summary: Experiments 1–3

The first three studies provide converging support for our hypothesis. The emotional payoffs of success are larger when many others compete—even when the percentage of competitors who are considered to have succeeded is held constant. This effect was found in perceptions of others' reactions, in forecasts of one's own reactions, and in actual reactions to success, across a variety of competitive contexts (see Figure 1).

In the final two studies, we sought to more directly explore the proposed mechanism—that large-scale success is associated with more positive emotional reactions because people perceive the outcome as more diagnostic of true abilities. If so, then explicit measures of diagnosticity should mediate the effect (Experiment 4), and the effect should be reduced or eliminated by competitions that are clearly undiagnostic of actual skill (Experiment 5). In other words, the next studies sought to test whether the effect is indeed driven by people's intuitive understanding of the *law of large numbers*.

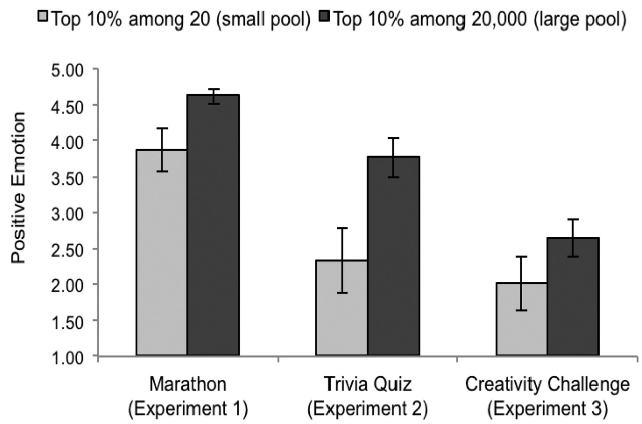

Figure 1. Results of Experiments 1–3 between conditions. Higher bars represent more positive emotion (more positive hypothetical judgments in Experiments 1–2 and more positive judgments of real-time reactions in Experiment 3). Error bars represent ±1 *SE*.

Table 1
Remote Associate Tests Used in the Creative Reasoning Challenge (Experiment 3)

Set of target words	Answer
1. cream/skate/water	ice
2. dew/comb/bee	honey
3. sleeping/bean/trash	bag
4. show/life/row	boat
5. fountain/baking/pop	soda
6. rocking/wheel/high	chair
7. night/wrist/stop	watch
8. loser/throat/spot	sore
9. cane/daddy/plum	sugar
10. cottage/Swiss/cake	cheese

Experiment 4

Real Running Ability

In Experiment 4, participants were again presented with the "Steve Marathon" scenario from the first study, but along with predicting Steve's happiness, participants rated what they thought he would infer about his true running abilities after the victory.

Method

Participants. Ninety-six undergraduates ($M_{age} = 18.70$, 67.7% female, 78.1% Caucasian) were recruited via a university subject pool for course credit.

Procedure. Participants were brought into the laboratory to complete a study related to how people think about competitions. First, they read the "Steve Marathon" scenario that was used in Experiment 1 and were randomly assigned to conditions. However, in addition to predicting Steve's emotional response on the Positive Feelings index from previous studies ($\alpha = .88$) and rating the size of the pool as a manipulation check (1 = *very small* to 6 = *very large*), they answered three questions about diagnosticity. First, participants were asked to predict what Steve might infer

[1] When simply conducting an independent samples *t* test with group size as the independent variable and emotional reaction as the dependent variable—that is, without controlling for ratings of believability—the effect of group size drops to near-marginal significance, and the pattern remains in the same direction, $t(64) = -1.35$, $p = .18$, $d = 0.34$. This result is not surprising and is conceptually consistent with the study. Unlike false feedback on a personality test, participants in a remote associates task have an objective sense of whether they are answering correctly or wildly guessing (Mednick, 1962; see also Table 1); hence, false feedback about their performance can vary greatly and should be controlled to provide a meaningful interpretation of the effect of group size.

from the outcome about his real running abilities. The low anchor was labeled as 1 (*not at all diagnostic; Steve wouldn't infer much about how good of a runner he "really" might be*). The high anchor was labeled as 11 (*very diagnostic; Steve would infer much about how good of a runner he "really" might be*). Second, they were asked to rate how predictive the outcome was of Steve's ability to win different kinds of races in the future (1 = *not at all predictive* to 11 = *very predictive*). Third, they were asked to imagine that Steve's true running ability could be measured on a meaningful scale from 0 (*has no running ability at all*) to 100 (*has the most running ability possible*) and were asked to predict the lower and upper boundaries between which his real skills might fall. We calculated the mean of these two values to serve as a single global prediction of Steve's trait running ability.

Results and Discussion

As expected, participants thought the race was larger when it contained 20,000 runners ($M = 5.22$, $SD = .90$) versus 20 ($M = 2.06$, $SD = .94$), $t(94) = -16.85$, $p < .001$, $d = 3.44$. Moreover, replicating previous studies, participants predicted that Steve would feel better after placing in the top 10% of many runners ($M = 4.04$, $SD = 1.18$) compared to placing in the top 10% of few ($M = 3.22$, $SD = 1.63$), $t(94) = -2.82$, $p = .006$, $d = 0.57$.

Importantly, there was a significant effect of group size on all three diagnosticity questions. First, participants thought that Steve's victory against many was significantly more diagnostic of his real running abilities ($M = 8.32$, $SD = 1.53$) than the same victory against few ($M = 6.57$, $SD = 1.85$), $t(94) = -5.06$, $p < .001$, $d = 1.02$. Second, participants thought that Steve's victory against many was significantly more predictive of future success ($M = 6.94$, $SD = 2.08$) than the same victory against few ($M = 5.51$, $SD = 2.03$), $t(94) = -3.36$, $p = .001$, $d = 0.69$. Third, participants who read that Steve placed in the top 10% among many inferred that he had a significantly higher average trait running ability ($M = 74.58$, $SD = 9.96$) than those who read about Steve's identical success among few ($M = 66.80$, $SD = 13.44$), $t(94) = -3.23$, $p = .002$, $d = 0.65$.

For simplicity, these three items were converted to standardized z scores and combined into a "Diagnosticity" index ($\alpha = .78$). In turn, participants who read that Steve placed in the top 10% among many inferred that his success was significantly more diagnostic according to this standardized scale ($M = 0.36$, $SD = 0.69$) than those who read about Steve's identical success among few ($M = -0.38$, $SD = 0.81$), $t(94) = -4.79$, $p < .001$, $d = 0.98$.

Finally, this aggregated index was used to test mediation. As predicted, diagnosticity mediated the effect of size on predictions of Steve's emotional reaction (see Figure 2). The results of multiple regression analyses (Baron & Kenny, 1986) were as follows. Group size (independent variable) was significantly related to positive feelings (dependent variable) and to diagnosticity (mediator). Diagnosticity was also significantly related to positive feelings, even after controlling group size. In addition, the link between group size and positive feelings became nonsignificant after controlling diagnosticity. The indirect effect of group size on positive feelings, via diagnosticity, was significant (95% bootstrap confidence interval = .03 to 0.64, which excludes the value 0; see Preacher & Hayes, 2004).

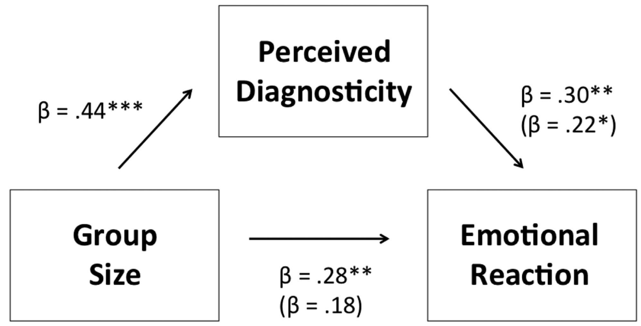

Figure 2. Results of multiple regression analyses with group size as the independent variable, diagnosticity as the mediator, and positive feelings as the dependent variable (Experiment 4). The βs in parentheses were obtained from a model that included both the independent variable and mediator as predictors of the dependent variable. * $p < .05$. ** $p < .01$. *** $p < .001$.

These findings replicate and extend the results of Experiment 1. Undergraduates in the lab thought Steve would feel happier after succeeding among many (vs. few), even though his threshold for success was held constant. Moreover, participants thought Steve would infer he was a better runner after succeeding among many, which mediated the effect of size on predicted reactions. These findings suggest people have an intuitive understanding of the law of large numbers, which has downstream consequences for emotion judgment. In the final study, we sought to build more diverse evidence for the proposed mechanism of diagnosticity by using a moderation-based strategy, helping to supplement the approach of directly measuring diagnosticity as a statistical mediator.

Experiment 5

Chili Cook-Off

In Experiment 5, people were asked to think about "Steve the Cook." They were randomly assigned to read that Steve entered a cooking challenge and placed among the top 10% of few or many competitors, and then estimated how happy he felt. However, we also manipulated the quality of judges in the competition. Some participants read that the food was evaluated by a panel of actual human judges—hence, success in this context should be viewed as having stemmed from a valid test of "true" cooking ability. Other participants read that the food was evaluated randomly by a computer—hence, success here should not be viewed as reflective of any true cooking ability. We predicted that group size would boost emotional reactions only for success within valid scenarios, in line with a mechanism that is grounded in diagnostic inferences about one's actual skill.

In addition to this primary comparison, we included a third exploratory condition in which participants read that the food was evaluated by a panel of ravenous children. Here, we were interested in exploring a competitive context that involved a more ambiguous degree of diagnosticity, as opposed to the clearly diagnostic real judges condition and the clearly nondiagnostic computers condition. Indeed, many evaluative contexts in everyday life are not as clear cut in terms of diagnostic assessment as in these two conditions (e.g., authors of peer-reviewed academic papers are unable to objectively assess the qualifications of their anonymous reviewers). We had no a priori hypothesis about whether participants would view success in the kids condition as stemming from a valid test of cooking abilities; to the extent that they did, however, these participants should be influenced by group size in line with the law of large numbers and hence produce similar results as if they had been assigned to the real judges condition.

Method

Participants. Three-hundred three undergraduates (M_{age} = 18.65, 54.1% female, 67.3% Caucasian) were recruited through a university participant pool for course credit.

Procedure. Participants were brought into the laboratory to complete a study related to how people think about competitions. The study followed a 2 (group size: small or large) × 3 (judge type: real judges, kids, or computers) randomized between-subjects design. Participants were asked to read a scenario about Steve, a cook who competed in a chili cook-off between few (20) or many (20,000) competitors. Participants in the real judges condition read the following passage:

> Steve decided to cook a family recipe for his city's "Chili Cook-Off" challenge. A large panel of judges tastes a sample of each submission throughout the weekend event and ranks them from best to worst. This year, 20 [20,000] people signed up to compete, including Steve. The top 10% of submissions are awarded a ribbon. It turns out that Steve does place in the top 10% of among the group of 20 [20,000].

Participants in the computers condition read a similar scenario, except they were told that because "all the submissions were so close this year," the judges were forced to use a computer "to rank the submissions completely at random." Participants in the exploratory kids condition also read a scenario similar to the real judges condition, except they were told that the panel was made up of "ravenous kids from local grade schools."

After reading the scenario, all participants predicted Steve's reactions using the same Positive Feelings index from prior experiments (α = .93) and responded to two manipulation checks: one for the size of the pool (1 = *very small* to 6 = *very large*) and one for the diagnosticity of the competition (1 = *not at all diagnostic; Steve wouldn't infer much about how good of a cook he "really" might be* to 11 = *very diagnostic; Steve would infer much about how good of a cook he "really" might be*).

Results and Discussion

Data were analyzed through multivariate analysis of variance (MANOVA) with group size, judge type, and the Group Size × Judge Type interaction serving as predictors. Our two manipulation checks and the Positive Feelings index served as dependent variables.

First, we report the results for our manipulation check of size. As expected, there was no main effect of judge type (p = .41), no interaction (p = .24), but a significant main effect of group size such that all "20,000 competitor" scenarios were rated as larger than all "20 competitor" scenarios, $F(1, 300) = 598.15$, $p < .001$, $\eta^2 = .67$ (all individual ps < .001).

Next, we report the results for our manipulation check of diagnosticity. Overall, one-sample t tests revealed that mean diagnosticity ratings for the real judges condition and the kids condition were significantly above the midpoint of the scale, whereas the computers condition mean was significantly below the midpoint (all ps < .001); in other words, the real judges and kids groups were seen as diagnostic, but the computers group was seen as undiagnostic. Accordingly, there was a main effect of judge type, $F(2, 301) = 100.76$, $p < .001$, $\eta^2 = .41$; a main effect of group size, $F(1, 301) = 5.97$, $p = .015$, $\eta^2 = .02$; and a significant interaction, $F(2, 301) = 3.68$, $p = .026$, $\eta^2 = .024$. Simple effects analyses revealed that larger pools were perceived as more diagnostic than smaller pools for the real judges and kids groups but not for the computers group (see Table 2).

Finally and most important to note is that in line with the hypothesis, results for our Positive Feelings index followed parallel patterns to these ratings of diagnosticity. Again, there was a main effect of judge type, $F(2, 302) = 118.70$, $p < .001$, $\eta^2 = .44$; a main effect of group size, $F(1, 302) = 9.60$, $p = .002$, $\eta^2 = .03$; and a significant interaction, $F(2, 302) = 2.63$, $p = .07$, $\eta^2 = .017$. Regarding our primary comparison, simple effects analyses confirmed that participants in the real judges group rated Steve as happier after success in large pools than small pools, but participants in the computer group showed no such effect. Regarding our exploratory comparison, given that participants rated the kids group as diagnostic, they did rate Steve as happier after success among many (vs. few) others—as expected and in line with the similarly diagnostic real judges group (see Table 2).

Taken together, these findings support the hypothesis that success in large pools feels better than identical success in small pools—but only under valid conditions. In obviously diagnostic

Table 2
Simple Effects Results for Our Manipulation Check of Perceived Diagnosticity, as Well as Judgments of Steve's Emotional Reaction, Across Each Condition (Experiment 5)

Scenario	20 competitors M (SD)	20,000 competitors M (SD)	p
"Real judges"	7.59 (1.70)$_{diagnosticity}$	8.63 (1.73)$_{diagnosticity}$	.013$_{diagnosticity}$
	3.59 (1.60)$_{emotion}$	4.39 (.75)$_{emotion}$	.004$_{emotion}$
"Computers"	4.46 (2.95)$_{diagnosticity}$	4.12 (2.76)$_{diagnosticity}$	.42$_{diagnosticity}$
	1.20 (1.43)$_{emotion}$	1.18 (1.48)$_{emotion}$	.935$_{emotion}$
"Kids"	7.27 (1.71)$_{diagnosticity}$	8.37 (1.44)$_{diagnosticity}$	.012$_{diagnosticity}$
	3.38 (1.87)$_{emotion}$	4.14 (1.26)$_{emotion}$	.01$_{emotion}$

situations (i.e., when success in a taste test is determined by actual judges), participants predicted that a target person would feel better after winning among many others compared with winning among few others. In contrast, this positive boost of size was attenuated in obviously invalid situations that afford no chance of being diagnostic to begin with (i.e., when success is determined randomly). This distinction is consistent with our hypothesis that the effect of size is driven by diagnostic inferences about true abilities, which is precisely what would be expected if people possessed an intuitive understanding about the law of large numbers.

It is interesting that in regard to our exploratory comparison, participants appeared to view a chili cook-off judged by ravenous children as a valid and diagnostic context. In turn, they thought that success among many would lead to more positive emotional reactions than success among few, adding further support for people's intuitions about the law of large numbers as a driving mechanism. Clearly, however, the objective merit of being judged in a cooking contest by a panel of hungry kids remains an open question. To the extent that this situation's perceived diagnostic validity diverges from actual diagnostic validity, people's ensuing emotional reactions might be misguided. We revisit this implication as a direction for future research in the General Discussion below.

General Discussion

Success feels good, but it feels better when many others compete—even when the percentage of competitors who are considered to have succeeded is held constant. Placing in the same percentile in the same competitive context is associated with more positive emotional reactions simply when the pool contains many (vs. few) people.

The first three studies established this effect across perceptions of hypothetical reactions as well as one's own actual reactions using a variety of competitive domains. The final two studies provided complementary evidence that these differences are indeed driven by perceived diagnosticity. People infer that a victory is more indicative of having superior true ability—that the performance must not be a fluke—after success among large groups than after identical success among small groups, leading to affective boosts. In line with this possibility, the effect is mediated by diagnosticity (Experiment 4), and it is attenuated when success occurs within clearly undiagnostic contexts (Experiment 5).

This process suggests that people exhibit an intuitive understanding of the law of large numbers, which posits that larger samples should contain distributions of values that better reflect a real population of possible scores (Sedlmeier & Gigerenzer, 1997; Tversky & Kahneman, 1971). Thus, success among larger pools is associated with more positive emotional reactions because people perceive the performance as more indicative of real superiority. On the one hand, this finding seems reasonable; after all, if winning in larger samples does theoretically reflect a more accurate assessment of real abilities, then participants were simply making a rational calculation. On the other hand, a large body of work documents people's notoriously irrational numerical reasoning, particularly in misperceiving relationships between absolute numbers relative to base probabilities (for reviews, see Kahneman & Tversky, 1982; Stanovich & West, 2008). In this sense, these findings demonstrate one salient situation where people do accurately take into account statistical properties, which in turn influences their emotion.

Alternative Explanations

Some readers might wonder if the affective boosts associated with winning in large pools are driven not by diagnostic inferences about true ability but by some other alternate reason. Two prominent possibilities are described below. We believe, however, that perceived diagnosticity is the most promising candidate, based on the current data.

First, might the boosts be driven by other qualitative differences caused by our manipulation? This possibility was captured earlier by our discussion about the Boston Marathon problem—whether imagining large or small competitive contexts invoked incidental attributes beyond size and diagnosticity that could have influence participants' emotional reactions to success. We believe this possibility was not driving the effect. First, we made explicit efforts to compare the same competitive context, isolating the influence of size itself from other characteristics of competitions. We presented participants with a wide variety of domains across 5 studies and nonetheless observed similar patterns—even for domains that seem to maintain the same qualities regardless of pool size (e.g., the *New York Times* contest in Experiment 2). Moreover, recall that Experiment 1 used the Steve Marathon scenario. As reported earlier, participants did rate the larger race as more prestigious than the smaller race, but the effect of pool size on predicted happiness remained significant when accounting for prestige. Similarly, in revisiting these data here, we found that ratings of prestige were completely unrelated to emotion in terms of mediation-based regression. Group size (independent variable) was significantly related to positive feelings (dependent variable), $\beta = .26$, $p = .02$, and also to prestige (mediator), $\beta = .69$, $p < .001$. Prestige, however, was not significantly related to positive feelings, $\beta = .13$, $p = .27$, and remained nonsignificant after controlling group size, $\beta = -.10$, $p = .51$; moreover, the link between group size and positive feelings remained significant after controlling prestige, $\beta = .33$, $p = .03$. This null effect of prestige as a statistical mediator was further reflected by bootstrapping analyses (95% bootstrap confidence interval $= -0.77$ to 0.28, which includes the value 0). In contrast, Experiment 4 found that people's diagnostic inferences about true ability did mediate the effect.

Second, might the boosts be driven by a more general feature of large-scale success? For example, people could simply possess a preference for beating an absolute greater number of others, or have an easier time imagining themselves winning or losing in varying-sized groups. In other words, mental images of victory in large groups could feel more satisfying (e.g., picturing many other losers, enhanced processing fluency) than mental images of victory in small groups. If so, people's associations between large-scale success and positive emotional reactions should carry over into nondiagnostic situations, regardless of what they can infer about true ability. This was not the case—judgments of emotional reactions were unaffected by size when participants were asked to imagine success that was randomly determined by a computer in Experiment 5. Collectively, these findings seem to render an analysis in terms of perceived diagnosticity more promising, suggesting people possess a rational understanding of the statistical law of large numbers.

Implications and Future Directions

These studies suggest fruitful follow-up research. First, on a practical level, surely not all competitive contexts should be susceptible to the effect, despite the fact that we observed similar results across a variety of domains using online and laboratory samples. Teasing apart when pool size more generally has an influence represents an important future direction. To this point, Figure 1 suggests people predict that others would feel stronger emotions after success (Experiment 1) than how they actually feel themselves (Experiment 3). This distinction is consistent with a large literature on forecasting error, which posits that people tend to overpredict the intensity and duration of their emotional reactions to various life events (for a review, see Wilson & Gilbert, 2003). Thus, one interesting possibility is that the group size effect is more pronounced for judgments of others or oneself in the nonimmediate future compared with real-time emotional reactions. Of course, given that our studies on self and other also explored different domains, differences between the two remain an open question.

Second, on a conceptual level, these findings are limited because they are confined to positive outcomes and do not shed light on the bandwidth of possible moderators. Might similar patterns be observed for negative performances (e.g., placing in the bottom 10%)? And are certain people more susceptible to the effect? None of our demographic variables influenced the results—including education in the first three studies, which may be closely tied to statistical reasoning tasks (for a review, see Nisbett, 2009; but see also West, Meserve, & Stanovich, 2012). This fact, as well as our replication of the effect in college samples in the laboratory, perhaps highlights its robustness across education. However, it also leaves open the potential for other relevant individual differences to serve as moderators. One possibility is people's need for uniqueness (Synder & Fromkin, 1980). If people regard feeling individuated from the group as more important than feeling confident about their abilities, then success among small pools (which contain a relatively smaller absolute number of other winners) should feel better than success among large pools (which contain a relatively larger number of absolute other winners). Another possibility is people's confidence about their actual skill (Kruger & Dunning, 1999). If Einstein scored in the top 10% of an intelligence test, his emotional reaction might be unaffected by pool size given that he should have expected to do well against any comparison group. Follow-up studies should explore these intriguing possibilities.

Third, future work should broaden the scope of the parameters of the scenarios that were used in the current studies (e.g., manipulating a wider range of group sizes and diagnostic features of the competitive landscape). It is interesting that these findings appear to imply that people are not susceptible to the "belief in the law of small numbers" (Tversky & Kahneman, 1971), the statistically inaccurate perception that small samples contain an equally representative range of skill as large ones, even if by a fluke. If this were the case, success in pools of 20 competitors should have seemed just as diagnostic and hence have been associated with just as positive emotional reactions, as success in pools of 20,000. Upon closer inspection, however, our findings may not rule out this belief. Diagnosticity ratings for 20 competitor scenarios were high at an absolute level, often surpassing the scale midpoint (e.g., see Table 2). In other words, judgments of low diagnosticity in these studies may have been inherently relative. Future work can shed light by testing a wider range of group sizes beyond 20. Evidence against a belief in small numbers would be more compelling if the patterns remain within objectively low diagnostic contexts.

Finally, the results of the kids condition in Experiment 5 suggest a potential split between perceived diagnosticity and actual diagnosticity. Although participants rated this scenario as highly diagnostic, the validity of having one's cooking judged by a group of ravenous kids may be debatable. Teasing apart when such perceptions represent an accurate assessment versus a problematic bias is an important avenue for future work. On the one hand, real success may be reinforced by emphasizing absolute features of positive outcomes. For example, a teacher could motivate outstanding students not just by revealing that their grades are in the "top 10% of the class" but also by highlighting the large size of the class in which they excel. On the other hand, to the extent that people misperceive their competitive context as valid (e.g., thriving in a large class of subpar students or scoring well on a clearly invalid test), successful people may claim more credit than warranted. Indeed, whenever we are rewarded for relative achievement—from winning awards and athletic contests to placing in top percentiles on standardized tests, professional performance evaluations, and public opinion polls—we may be prone to feeling unjustifiably special simply because many (vs. few) others were involved.

Conclusion

Taken together, these studies broadly reveal the need for a better understanding of how numbers and chances can influence emotional—and not just cognitive—processes. The top 10% apparently has no stable value; rather, placing in the same percentile in the same competition seems better simply when many competitors are involved. So long as the context is perceived as valid and the outcome as successful, people's intuitions about the law of large numbers can provide an emotional boost: the greater the number of others, the greater the affective payoffs.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182. doi:10.1037/0022-3514.51.6.1173

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavioral Research Methods, Instruments, & Computers, 35,* 634–639. doi:10.3758/BF03195543

Festinger, L. (1954). A theory of social comparison. *Human Relations, 7,* 117–140. doi:10.1177/001872675400700202

Garcia, S. M., & Tor, A. (2009). The *N*-effect: More competitors, less competition. *Psychological Science, 20,* 871–877. doi:10.1111/j.1467-9280.2009.02385.x

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 493–508). New York, NY: Cambridge University Press.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19,* 1–53. doi:10.1017/S0140525X00041157

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121–1134. doi:10.1037/0022-3514.77.6.1121

McFarlin, D. B., & Blascovich, J. (1984). On the Remote Associates Test (RAT) as an alternative to illusory performance feedback: A methodological note. *Basic and Applied Social Psychology, 5,* 223–229. doi:10.1207/s15324834basp0503_5

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69,* 220–232. doi:10.1037/h0048850

Mukherjee, K., & Hogarth, R. M. (2010). The *N*-effect: Possible effects of differential probabilities of success. *Psychological Science, 21,* 745–747. doi:10.1177/0956797610368812

Nisbett, R. E. (2009). *Intelligence and how to get it: Why schools and cultures count*. New York, NY: Norton.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavioral Research Methods, Instruments, & Computers, 36,* 717–731. doi:10.3758/BF03206553

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10,* 33–51. doi:10.1002/(SICI)1099-0771(199703)10:1<33::AID-BDM244>3.0.CO;2-6

Slovic, P. (1986). Informing and educating the public about risk. *Risk Analysis, 6,* 403–415. doi:10.1111/j.1539-6924.1986.tb00953.x

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94,* 672–695. doi:10.1037/0022-3514.94.4.672

Synder, C. R., & Fromkin, H. L. (1980). *Uniqueness: The human pursuit of difference*. New York, NY: Plenum.

Taylor, S. E., & Lobel, M. (1989). Social comparison activity under threat: Downward evaluation and upward contacts. *Psychological Review, 96,* 569–575. doi:10.1037/0033-295X.96.4.569

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 181–227). New York, NY: Academic Press. doi:10.1016/S0065-2601(08)60227-0

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110. doi:10.1037/h0031322

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Erlbaum.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology, 103,* 506–519. doi:10.1037/a0028857

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. *Advances in Experimental Social Psychology, 35,* 345–411. doi:10.1016/S0065-2601(03)01006-2

Received September 5, 2012
Revision received January 17, 2013
Accepted January 28, 2013 ■