Correspondence Bias in Performance Evaluation: Why Grade Inflation Works

Don A. Moore[1], Samuel A. Swift[1], Zachariah S. Sharek[1], and Francesca Gino[2]

Abstract

Performance (such as a course grade) is a joint function of an individual's ability (such as intelligence) and the situation (such as the instructor's grading leniency). Prior research has documented a human bias toward dispositional inference, which ascribes performance to individual ability, even when it is better explained through situational influences on performance. It is hypothesized here that this tendency leads admissions decisions to favor students coming from institutions with lenient grading because those students have their high grades mistaken for evidence of high ability. Three experiments show that those who obtain high scores simply because of lenient grading are favored in selection. These results have implications for research on attribution because they provide a more stringent test of the correspondence bias and allow for a more precise measure of its size. Implications for university admissions and personnel selection decisions are also discussed.

Who is likely to be the more ambitious and hard-working graduate student—the one with a 3.6 grade point average (GPA) from a school where the average GPA is 3.4 or the one with a 3.3 from an institution where the average GPA is 2.8? This sort of difficult attribution problem is crucial to all types of personnel selection decisions, from admitting applicants to picking teammates (Staw, Bell, & Clausen, 1986; Staw, Sutton, & Pelled, 1994). The question we pose in this article is whether those making the selections can adequately adjust for the difficulty of success when making inferences about what performance signals about abilities.

We begin with the fact that undergraduate institutions vary in their grading standards, even schools that are otherwise similar in selectivity and student quality (Attewell, 2001; Bagues, Sylos Labini, & Zinovyeva, 2008; Goldman & Widawski, 1976). This basic fact raises the question of whether those who use information about grades to assess students (such as future employers or graduate schools) use that information appropriately. Do people appropriately adjust their interpretation of grades based on the leniency of grading? Research findings on the psychology of attribution give us reason to doubt that they do.

Biases in the Attribution Process

The problem of assessing the informative value of academic grades is a special case of a more general problem: how to infer the qualities of the individual (such as intellectual abilities) from behavior or outcomes (GPA) while subtracting out the influence of the situation (leniency of grading). The solution to this problem is provided by Kurt Lewin's (1951) attributional equation: Behavior = f(Disposition, Situation). In other words, behavior is a joint function of both the individual's disposition and the influence of the situation. We capitalize on Gilbert's (1994) suggestion that the Lewinian equation often takes the specific form: Disposition = Behavior – Situation. In this case, Academic Performance = Grades – Grading Leniency.[1] However, research suggests that people do not apply this simple formula perfectly. One of the most common biases in the attribution process is the tendency to ascribe too little influence to the situation and too much to the individual's disposition (Davis-Blake & Pfeffer, 1989; Jones & Harris, 1967; Nisbett & Borgida, 1975; Ross, 1977).

Ross, Amabile, and Steinmetz (1977) examined this phenomenon in a study that paired participants and randomly assigned one of them to the role of quiz master, who made up

[1]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Corresponding Author:
Don A. Moore, Tepper School, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
Email: don.moore@alumni.carleton.edu

the questions, and one to the role of quiz taker, who answered them. Naturally, the quiz master knew some answers that the quiz taker did not. But rather than attributing this to the quiz master's role-conferred advantage, observers reported that the quiz master was the more knowledgeable of the pair. Attributions of knowledgeability were biased by an excessive belief in the correspondence between behaviors and dispositions. This is what Gilbert and Malone (1995) called the *correspondence bias* and what Ross (1977) called the *fundamental attribution error.*

Is the Correspondence Bias Really a Bias?

In Ross et al.'s (1977) experiment, as in many studies demonstrating the correspondence bias, it was difficult for individual participants to precisely determine the strength of the situation. On average, the quiz master was unlikely to be more knowledgeable than the quiz taker (given random assignment to roles), but that was little help for the individual who had to decide whether a specific quiz master is more or less knowledgeable than a particular quiz taker. To accurately judge the strength of the situation, participants in Ross et al.'s experiment needed to know what proportion of questions, on average, quiz takers failed to answer correctly. If participants had this information, they would have been better able to specify the strength of the situational differences between the quiz master and the quiz taker. But they did not get the information. If it is impossible to determine the strength of the situation, it becomes impossible to adjust for it when making attributions.

This fact raises the possibility that the correspondence bias, as it has been demonstrated previously, might simply be a problem of incomplete information. We address this possibility by examining whether the correspondence bias persists when people have all the information they need to adjust their attributions of individual abilities based on the influence of the situation. Participants in our experiments are given quantified information about both the behavior (i.e., grades) and the situation (i.e., grading norms). Previous research on the correspondence bias has not tested the bias in situations where participants have clear, quantified information about both the situation and the outcome. Furthermore, prior studies failed to specify the strength of the situation. This leaves open some important questions about the causes of the correspondence bias. Our research paradigm can help answer these questions. If the correspondence bias persists even in the presence of full information, that would strengthen its standing as a bias and not simply an incomplete information problem.

Hypothesis

The primary hypothesis we investigate is that absolute performance will be insufficiently discounted relative to the ease
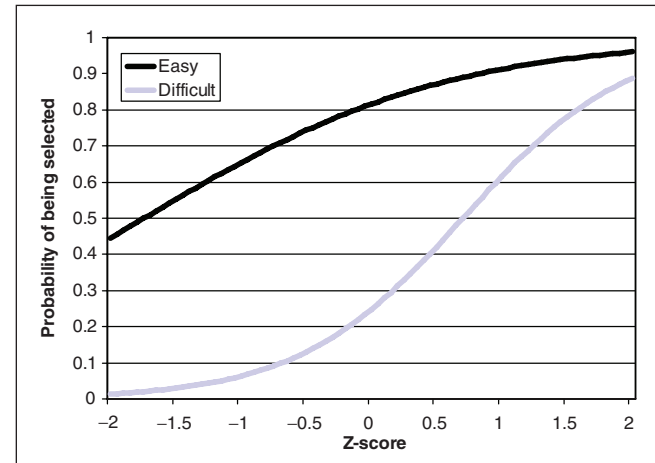

Figure 1. Probability of being selected, conditional on quiz difficulty and score relative to others on that quiz (Experiment 3)

of the task. Specifically, raw GPAs will be taken as evidence of academic performance and not sufficiently adjusted to account for the ease with which those grades were earned. In other words, an applicant's absolute GPA will have a stronger influence on admission decisions than will the grading leniency of their institutions. Rationally, evaluations of an applicant's prior academic performance should rely primarily on two things: (a) the rigor or quality of the institution or program of study and (b) the individual's performance relative to others in that same program (see Berry & Sackett, 2009). Our studies control for the first consideration and vary two aspects of the second: absolute performance (indicated by the individual's GPA) and the ease of obtaining a high score (indicated by the average GPA at the undergraduate institution). The easier the task, the less impressive high performance should become. Evaluations should give the leniency of grading (as measured by the average grade at the undergraduate institution) a decision weight equal in size and opposite in sign from that given to each candidate's GPA.

To draw an analogy, assume your goal is to pick the tallest players for your basketball team. In a desperate attempt to improve their chances of making the team, some of the players trying out have worn elevator shoes. If you know only the player's total height (with shoes) and the height of the shoes, then those two measurements should be weighed equally and oppositely in determining the player's shoeless height: Each inch contributed by the shoe will reduce the player's size by 1 in. when the shoes are removed. Instead, we hypothesize that the positive effect of individual performance (e.g., height) on evaluations will not be matched by the discounting effect of situational factors (e.g., elevator shoes). In the context of grades and admission, our hypothesis predicts that people will favor those from institutions with lenient grading because absolute GPA will be weighed heavily in evaluations of

applicants, but average grades at the institution will not be sufficiently discounted. In other words, those who show up in elevator shoes will be more likely to make the team.

Experiments 1 and 2 consider graduate school admission decisions. Experiment 3 replicates the same type of decision problem outside the context of GPAs and admissions decisions. All three experiments are consistent in showing that nominal performance is too readily taken at face value without discounting for obvious situational influences, even when the effect of those situational influences is obvious and easily quantified.

Experiment 1

This experiment put participants in the role of admissions decision makers and presented them with information about specific candidates' performance (GPA) as well as an indication of the distribution from which the GPA came (college average GPA). We manipulated these two factors in a 3 (GPA relative to average: above vs. equal vs. below) ×3 (average GPA at undergraduate institution: high vs. medium vs. low) within-subjects design. Candidates had GPAs that were .3 above their school's average, at their school's average, or .3 below their school's average. This manipulation was crossed with a manipulation of the average grade at the candidate's alma mater: Applicants came from colleges with average grades that were either high (average GPA of 3.6), medium (3.0), or low (2.4). Note that to reduce the obviousness of our manipulation, both the GPAs of the individual applicants and the average GPAs of their institutions varied slightly around these precise points (within .02).[2]

Naturally, we expected that being above average would have a positive effect on the probability of being admitted. Our more interesting hypothesis is that the school's average GPA would have a significant positive effect on the probability of admission: Candidates from colleges with high average grades would be more likely to be admitted. In other words, we expected that people will not sufficiently discount high grades that are due to lenient institutional grading practices.

Method

Participants. Fifty-five undergraduates at a research university in the Northeastern United States participated in the study in exchange for course credit in their introductory business courses.

Procedure. Participants were given the following instructions:

In this exercise, you will be playing the role of a member of the admissions committee at a selective MBA program. You are selecting students who would like to obtain masters degrees in business administration.

Your most important goal is to select the best candidates from among the applicants. In general, you usually have space to admit about half the applicants. You will see the applications of nine hypothetical students. The set of applicants that you will review all graduated from colleges of similar quality and selectivity. Please review each applicant carefully in order to assess the quality of their prior academic performance in college. Please review one candidate at a time. Answer the questions about each candidate before turning the page to read about the next candidate.

Information about the candidates included their GPA, the average GPA at the institution from which they obtained their undergraduate degree, and their grades in the last 10 classes they took. These classes were listed for each candidate. Both the candidate's grade and the class average for each course were shown. The candidate's grades in the 10 classes had the same mean as the candidate's overall GPA, and the average grades in each of the courses had the same mean as the undergraduate institution overall. To highlight each candidate's relative standing, the difference between his or her GPA and the average for the college was also specifically shown. This list of classes was counterbalanced across all conditions so as not to confound it with experimental condition.

For each candidate, participants were asked to (a) evaluate how successful the candidate had been in college on a 7-point scale, anchored at 1 (*very unsuccessful*) and 7 (*very successful*), and (b) report how likely they would be to admit them (as a numerical probability between 0% and 100%). After evaluating all nine candidates, they were asked to look back through the set and admit only four of the nine. In sum, for each candidate, each participant provided three ratings: (a) a rating of prior success, (b) an estimated probability of admission, and (c) an actual admission decision.

Participants were randomly assigned to one of nine randomly determined order conditions in a Latin squares design such that each candidate's position in the order was balanced. Names of the nine fictional colleges and course lists were counterbalanced across manipulations.

Results and Discussion

The descriptive statistics for the measures used as dependent variables are reported in Table 1. The three ratings of each candidate correlated strongly with each other (all *r*s above .6) and were therefore standardized by converting them to *z* scores and averaged to form a single measure of candidate admissibility (Cronbach's alpha = .86).

This admissibility assessment was then subject to a 3 (GPA relative to average) × 3 (average GPA at undergraduate institution) within-subjects ANOVA. Naturally, the results show a main effect of relative GPA, $F(2, 108) = 333.84$, $p < .001$,

Table 1. Ratings of Undergraduate Success, Estimated Probability of Being Offered Admission, and Observed Probability of Being Admitted to Graduate School Based on Undergraduate Grade Point Average (GPA) and Average Grades at Undergraduate Institution (Experiment 1)

Institution average GPA:	Low ≈ 2.4			Medium ≈ 3.0			High ≈ 3.6		
Individual GPA:	.3 below average	About average	.3 above average	.3 below average	About average	.3 above average	.3 below average	About average	.3 above average
Rated prior success (1-7)	2.33 (1.00)	3.42 (0.92)	4.16 (0.96)	3.05 (0.85)	4.25 (0.91)	5.05 (0.78)	3.64 (1.01)	4.49 (1.00)	6.25 (0.75)
Rated probability of acceptance	20% (14%)	36% (17%)	45% (18%)	32% (18%)	47% (19%)	61% (17%)	44% (19%)	54% (20%)	83% (15%)
Actual acceptance rate	2%	7%	56%	7%	50%	96%	30%	61%	94%

Standard deviations appear in parentheses.

$\eta^2 = .86$. Those with above-average GPAs received higher admissibility ratings ($M = 0.71$, $SD = 0.63$) than did those with below-average GPAs ($M = -0.67$, $SD = 0.41$). As hypothesized, the results also show a significant main effect for average GPA at undergraduate institution, $F(2, 108) = 96.35$, $p < .001$, $\eta^2 = .64$. Consistent with our expectations, candidates from institutions with lenient grading were more likely to receive higher admissibility ratings ($M = 0.46$, $SD = 0.80$) than were candidates from schools with strict grading ($M = -0.52$, $SD = 0.56$). The results also reveal a GPA × Leniency interaction effect, $F(4, 216) = 6.44$, $p < .001$, $\eta^2 = .11$. This interaction describes the fact that the effect of grading leniency differs as a function of relative GPA. Specifically, the benefit of having performed better than one's peers is stronger for those graded most leniently (where it increases admissibility by 1.26) than for those graded strictly (where it only increases admissibility by .81). However, as Table 1 shows, the effect of having been graded leniently persists across all conditions.

To further examine the decision-making process participants used to judge candidates and to identify where they attempted to correct their judgments (although insufficiently), we considered each of our three dependent variables separately. A 3 (GPA relative to average) × 3 (average GPA at undergraduate institution) within-subjects ANOVA using prior rating of success revealed the same results as above: a significant main effect of relative GPA, $F(2, 108) = 284.63$, $p < .001$, $\eta^2 = .84$; a significant main effect for average grade at undergraduate institution, $F(2, 108) = 94.45$, $p < .001$, $\eta^2 = .64$; and a significant interaction, $F(4, 216) = 10.19$, $p < .001$, $\eta^2 = .16$. Next, we considered the rated probability of admission as the dependent variable. The results did not vary in nature or significance: The main effect of relative GPA was significant, $F(2, 108) = 230.59$, $p < .001$, $\eta^2 = .81$; the main effect for average grade at undergraduate institution was significant, $F(2, 108) = 118.41$, $p < .001$, $\eta^2 = .69$; and the interaction was significant, $F(4, 216) = 11.47$, $p < .001$, $\eta^2 = .18$.

To analyze admission decisions, we employed a binary logistic regression, given the dichotomous nature of the dependent variable. This analysis also allowed us to fit the data to the version of the Lewinian equation with which we opened the article. This analysis is different from those reported so far, which used the experimental conditions as categorical independent variables. Instead, this new analysis uses two continuous variables: candidate GPA and average institution GPA. This approach has the advantage that, because the two independent variables are both quantified on the same scale, we can compare their effects more directly. However, it leads us to expect a negative effect of grading leniency. To understand why, it is useful to go back to the example of the tryouts for the basketball team, some of whom are wearing elevator shoes. Our prior analyses examined the independent effects of the person's height and the size of the elevator shoes. For the present analysis, however, we want to include the person's total height (with shoes) and the size of the shoes. We would expect the height of the shoes to have a discounting effect (a negative effect) on judgments of candidate height. The correspondence bias would predict that this effect would be smaller than it should be.

The results of this new analysis reveal that raw GPA and grading leniency are both significant predictors of admissions outcomes. Unsurprisingly, absolute GPA had a significant positive relationship with the rated probability of admission ($B = 7.22$, $SE = .66$, $p < .001$). We expected the effect of lenient grading to be negative in this analysis for the same reason the effect of the situation ought to be subtracted from behavior when making inferences about individual traits using the Lewinian equation: The situation (lenient grading) ought to be subtracted from the behavior (GPA) to make sensible attributions about the individual's dispositions. The result of the analysis is that coming from an institution with lenient grading did indeed have a negative effect on the chances of being offered admission ($B = -4.85$, $SE = .57$, $p < .001$), but the size of this effect was significantly smaller in magnitude than the effect of raw GPA, $\chi^2(1) = 65.30$, $p < .001$. In other

words, participants did discount a high GPA somewhat if it came from an institution with lenient grading, but this effect was too small to undo the strong positive effect of having a higher GPA. The size of the negative effect of grading leniency was only 67% the size of the positive effect of having a higher GPA.

The results of Experiment 1 are consistent with the hypothesis that absolute GPAs are taken as direct evidence of prior academic performance and are not appropriately discounted by the ease with which those grades were earned. However, it is obvious that participants did not completely ignore information about grading leniency; they just did not weight the discounting information as heavily as they did the nominal performance numbers.

This result is notable because the information we gave our participants on the strength of the situation is so much clearer than it has been in prior studies of the correspondence bias. The fact that we provided our participants with unambiguous quantifications of both people's behavior (their GPAs) and the situation that gave rise to that behavior (the average GPA at that institution) means that we can make stronger claims about bias than can prior research. In our experimental paradigm it is clear that GPA and grading leniency should have been equally and oppositely weighted. The fact that they are not allows us to pinpoint exactly how it is that our participants' decisions deviate from the optimal decision and how much this matters. In Experiment 1 grading leniency received a decision weight that was 67% the size of the GPA decision weight. We obtain two more estimates of this discrepancy from Experiments 2 and 3.

Experiment 2

The design of Experiment 1 paralleled the reality that individual grades will always show greater variance than variance in institution averages. Although this is realistic, it also raises the concern that it might increase the chance of finding a stronger influence for individual grades. After all, it is easier to detect the influence of independent variables with greater variability on the dependent measure. Consequently, we designed Experiment 2 to equalize the variability between individual grades and institutional averages.

Applicants came from colleges that varied with respect to grading leniency: Average grades were either high (average GPA of 3.6), medium (3.0), or low (2.4). This manipulation was crossed with a manipulation of the candidates' GPAs relative to their classmates: Candidates had GPAs that were high (3.6), medium (3.0), or low (2.4). We expected that, as in Experiment 1, applicants' nominal GPAs would exert a strong influence on admissions decisions but that the effect of the grading norms at their alma maters would not be as strong. However, the design of Experiment 2 leads us to predict a negative effect of more lenient

grading. By holding nominal GPAs constant, Experiment 2 makes a low GPA particularly bad, not just in absolute terms but also relative to the norm at the institution. The condition in which the candidate has a low GPA (2.4) and comes from an institution with more lenient grading (average GPA of 3.6) puts the applicant a full 1.2 grade points below the mean, a bigger difference than in any condition from Experiment 1. We expect to find, as in Experiment 1, that participants will (a) favor those with high GPAs and (b) insufficiently discount those GPAs based on grading leniency.

Method

Participants. Forty-five undergraduates from a research university in the Northeastern United States participated in the study in exchange for course credit.

Procedure. The study employed the same procedure and measures as in Experiment 1 except for the new individual grades and institutional averages.

Results and Discussion

As in Experiment 1, the three ratings of each candidate were standardized and averaged to form a single measure of admissibility (Cronbach's alpha = .81). This admissibility measure was then subject to a 3 (GPA: high vs. medium vs. low) × 3 (average GPA at undergraduate institution: high vs. medium vs. low) within-subjects ANOVA. The results demonstrate the expected main effect of GPA, $F(2, 88) = 492.8$, $p < .001$, $\eta^2 = .92$. Those with high GPAs received substantially higher admissibility scores ($M = 0.89$, $SD = 0.22$) than did those with low GPAs ($M = -0.91$, $SD = 0.19$). The results also show a more modest but significant main effect for grading leniency, $F(2, 88) = 29.73$, $p < .001$, $\eta^2 = .40$. Candidates from institutions with lenient grading received lower ($M = -0.27$, $SD = 0.87$) admissibility scores than did candidates from schools with strict grading ($M = 0.20$, $SD = 0.90$). The results also reveal a GPA × Leniency interaction effect, $F(4, 176) = 2.72$, $p = .03$, $\eta^2 = .06$. Here, the benefit of having performed better than one's peers appears to be stronger for those from institutions of moderate grading leniency (where it increases overall assessment, 1.88) than for those graded strictly (1.74) or leniently (1.79). But effect of having been graded leniency persists across all conditions, as Table 2 shows.

As we did in Experiment 1, we next considered each of our three dependent variables separately. A 3 (GPA relative to average) × 3 (average grade at undergraduate institution) within-subjects ANOVA using prior rating of success revealed the following results: a significant main effect of relative GPA, $F(2, 88) = 374.35$, $p < .001$, $\eta^2 = .90$; a significant main effect for average grade at undergraduate institution, $F(2, 88) = 33.30$, $p < .001$, $\eta^2 = .43$; and an insignificant interaction, $F(4, 176) = 1.15$, $p = .34$, $\eta^2 = .03$.

Table 2. Ratings of Undergraduate Success, Estimated Probability of Being Offered Admission, and Observed Probability of Being Admitted to Graduate School Based on Undergraduate Grade Point Average (GPA) and Average Grades at Undergraduate Institution (Experiment 2)

Institution average GPA:	Low ≈ 2.4			Medium ≈ 3.0			High ≈ 3.6		
Individual GPA:	About average	.6 above average	1.2 above average	.6 below average	About average	.6 above average	1.2 below average	.6 below average	About average
Rated prior success (1-7)	3.46	4.78	6.01	3.02	4.64	5.76	2.58	3.99	5.36
	(0.95)	(0.93)	(0.69)	(0.97)	(0.68)	(0.65)	(0.87)	(0.82)	(0.93)
Rated probability of acceptance	37%	58%	78%	32%	57%	75%	27%	50%	69%
	(18%)	(19%)	(13%)	(17%)	(14%)	(14%)	(17%)	(16%)	(19%)
Actual acceptance rate	4%	60%	98%	0%	56%	96%	0%	13%	76%

Standard deviations appear in parentheses.

Next, we considered the estimated probability of admission as the dependent variable. The results did not vary in nature or significance: a significant main effect of relative GPA, $F(2, 88) = 254.54$, $p < .001$, $\eta^2 = .85$; a significant main effect for average grade at undergraduate institution, $F(2, 88) = 21.30$, $p < .001$, $\eta^2 = .33$; and an insignificant interaction, $F(4, 176) < 1$, $p = .72$, $\eta^2 = .01$.

To compare the results of Experiment 2 with those of Experiment 1, we conducted another binary logistic regression predicting admission decisions using each candidate's GPA and average GPA at the undergraduate institution as the independent variables. Once again, both are significant predictors. Unsurprisingly, raw GPA had a significant positive relationship with the probability of admission ($B = 5.48$, $SE = .52$, $p < .001$). More interestingly, the negative effect of coming from an institution with lenient grading was significant ($B = -1.90$, $SE = .35$, $p < .001$) but substantially smaller in magnitude than the effect of raw GPA, $\chi^2(1) = 53.68$, $p < .001$. The discounting effect was only 34% of the size of the effect of a higher GPA.

Again, people somewhat discounted high GPAs when they came from institutions with lenient grading, but not to the degree that it sufficiently counteracted the strong positive effect of having a higher GPA. To be specific, an increase of 1 point in an applicant's GPA increased his or her rated probability of admission by 34.9%. But if that 1 point increase came solely because everyone at that institution got higher grades, it still increased the rated probability of admission by 27.3%. An increase in grading leniency of an entire grade point at the undergraduate institution only decreased the rated probability of admission by 7.59% for the same nominal GPA.

The results of Experiment 2 replicate the key finding of our first experiment. In Experiment 1, we showed that applicants from lenient-grading institutions were *more* likely to be admitted, holding constant their performance relative to their schools' average. However, in Experiment 2, we found that applicants from lenient-grading institutions were *less*

likely to be admitted, holding constant their nominal GPAs. As we explained earlier, these results are in fact perfectly consistent. Both experiments show participants (a) favoring those with high GPAs and (b) insufficiently discounting those GPAs based on grading leniency. Holding relative performance constant, candidates were evaluated more positively when they came from institutions with higher grades. Candidates with average grades from a college with an average GPA of 3.6 were admitted by 76% of our participants. Candidates with average grades from a college with an average GPA of 2.4 were admitted by 4% of our participants.

Experiment 3

One potential concern regarding Experiments 1 and 2 is that if people believe that high average grades are correlated with desirable features of a college or its graduates (despite our assurance that the institutions did not differ with regard to quality and selectivity) the tendency to favor graduates of institutions with high average grades makes sense. To rule out this explanation for our findings, we conducted a third experiment outside the domain of university admissions decisions. Instead, participants in Experiment 3 were asked to imagine that they had to select members for a "quiz bowl" trivia team. They reviewed the prior test performances of 10 applicants, 5 of whom had taken an easy test and 5 of whom had taken a difficult test. Both tests were on the subject of U.S. geography. Our hypothesis was that those who had high scores because they had taken the easy test would, like those who come from institutions with lenient grading, be more likely to be selected.

Method

Participants. Participants were 71 undergraduates at a research university in the northeastern United States participating for money.

Design and procedure. The experiment employed a 2 (experience: experience with task vs. no experience with task) × 2 (task difficulty: easy vs. hard) design. The first factor was manipulated between subjects, while the second factor was manipulated within subjects.

The task for Experiment 3 was similar to the first two experiments: to evaluate 10 candidates and eventually decide which 5 to select. However, participants in this experiment were asked to select the candidates they thought would perform above average on a third quiz, which was shown to participants at the time of selection:

In this study, we are interested in your ability to predict the performance of others. You will now see the scores of ten people who took one of two quizzes. For five of the people, you will see their scores on the first quiz. For the other five, you will see their scores on the second quiz. For each of the ten contestants, their correct answers are marked with a check and their incorrect answers are marked with an X. For each of the ten contestants, we will ask you to estimate their knowledgeability about US geography. All ten of these people also took a third quiz on the same topic of US geography. After examining ten contestants, we will ask you to identify the five people you think are most likely to perform well on the third quiz. This third quiz was the same for all ten contestants. You will earn $2 today for each contestant you pick whose score is in the top half of the performers on the third quiz. Therefore, if you correctly pick the five top scorers, we will reward you with $10 in cash for your performance. If the five contestants you pick are the five worst performers on the third quiz, you will not earn any additional money for this study.

The first factor we manipulated was experience with the task to test for the possibility that prior experience would reduce the bias observed in our first two studies. Based on previous findings (Epley, Savitsky, & Gilovich, 2002; Van Boven, Kamada, & Gilovich, 1999) we hypothesized that participants' experience with the task before making their judgments would attenuate the size of the correspondence bias. After all, experience with the task helps make salient the situational pressures through their effects on one's own behavior. This is in part why people are far more sensitive to situational effects on their own behavior than on the behavior of others (Jones & Nisbett, 1971). In this way, we hoped that personal experience with the situation could help people appreciate how situational constraints would affect the behavior and thus reduce the correspondence bias.

Participants in the experience condition were given an additional page with instructions at the beginning of the experiment:

Your first task in this study is to take two different trivia quizzes. Your goal is to answer as many questions correctly as you can, using your memory alone. You may not consult other people or information sources other than your own memory. Good luck!

Participants in the no-experience condition did not receive this additional page with instructions.

After the experience manipulation, participants evaluated each of the candidates for the quiz bowl trivia team. For each of the 10 candidates, participants saw actual completed quizzes from participants in a previous pilot study that included candidates' answers marked as correct or incorrect. Participants saw quizzes from 5 candidates who had taken a difficult quiz with questions like "How many U.S. states border Canada?" (mean score: 1 out of 10) and they saw quizzes from 5 who had taken a simple quiz with questions like "The Bronx is part of what U.S. city?" (mean score: 8.9 out of 10).

These 10 quizzes were selected such that the mean score and standard deviation for each type of quiz roughly matched the mean and standard deviation among all quiz takers in the pilot study from which the quizzes were selected. We divided these 10 quizzes into two sets such that the easy and difficult quiz scores of the candidates in each set were similar to each other. Set 1 included the easy quizzes of Candidates 2, 3, 4, 5, and 8, and the difficult quizzes of Candidates 1, 6, 7, 9, and 10. Set 2 included the easy quizzes of Candidates 1, 6, 7, 9, and 10, and the difficult quizzes of Candidates 2, 3, 4, 5, and 8. We also varied order as follows. We first randomized the order of the 10 candidates, and then we reversed this order to make a second order condition. Participants were randomly assigned to one of these four conditions created by our 2 (set) × 2 (order) between-subjects design. Note that the easy and difficult test scores were from the same, real individuals who previously participated in a pilot study.

After seeing a candidate's completed quiz, participants were then reminded of the candidate's score (out of 10) and were told the average score and standard deviation among all 10 test takers on that quiz. For the first set, the 5 easy scores participants saw were 9, 9, 9, 7, and 10 ($M = 8.8$, $SD = 1.1$). For the second set, the 5 easy scores were 10, 8, 8, 9, and 10 ($M = 9.0$, $SD = 1.0$). The 5 difficult scores participants saw from the first set were 1, 2, 0, 2, and 0 ($M = 1.0$, $SD = 1.0$). For the second set, the 5 difficult scores were 2, 1, 2, 0, and 0 ($M = 1.0$, $SD = 1.0$). Participants were then asked to rate how knowledgeable about U.S. geography they thought each contestant was using a 7-point scale that ranged from *not knowledgeable at all* to *very knowledgeable*.

Before making their selections, participants were reminded that each candidate they had seen had either taken a simple or a difficult quiz. Three questions then asked participants to

compare the two quizzes on 7-point scales: (a) "Do you think the two tests were equally good at testing candidates' trivia skills?" (from _simple is better_ to _difficult is better_), (b) "Do you think the two tests were equally fair measures of ability?" (from _simple is more fair_ to _difficult is more fair_), (c) "Do you think the two tests will be equally good predictors of performance if chosen for the team?" (from _simple is better_ to _difficult is better_).

After they had compared the two quizzes, participants read the following:

> Please select which candidates you think will do best on a quiz that was given to all quiz-takers. A copy of this quiz is below. Remember that for each person you select who performs better than average on the quiz at the bottom of this page you will earn $2.

The third test was also a geography test, of intermediate difficulty.

After participants made their selections, their choices were scored and payoffs were computed. After being paid, participants were thanked, debriefed, and dismissed.

Results and Discussion

We computed two averages for ratings of knowledgeability: one for the five contestants whose easy quizzes participants saw, and another for the five contestants whose difficult quizzes participants saw. These averages were then submitted to a 2 (experience) × 2 (test difficulty) mixed ANOVA with repeated measures on the second factor. The results reveal a significant within-subjects effect of test difficulty, $F(1, 69) = 136$, $p < .001$, $\eta^2 = .66$. When participants saw a contestant's easy quiz, that contestant was rated as significantly more knowledgeable ($M = 5.13$, $SD = 1.13$) than was the same contestant rated by participants who had seen his or her difficult quiz ($M = 2.65$, $SD = 1.01$). The main effect of experience was not significant, $F(1, 69) < 1$, $p = .59$. If experience taking the two quizzes helped participants avoid the correspondence bias, it would have shown up as an Experience × Difficulty interaction, wherein experience reduced the effect of difficulty on rated knowledgeability. This interaction does not quite attain significance, $F(1, 69) = 3.12$, $p = .08$, $\eta^2 = .04$. However, this marginally significant effect is not due to a debiasing influence provided by the experience manipulation. Although the difference between ratings of the easy ($M = 4.98$) and hard ($M = 2.88$) tests are significant among those without experience, $t(33) = 6.42$, $p < .001$, this difference is marginally larger among those _with_ experience, reflecting a stronger difference between the easy ($M = 5.28$) and hard ($M = 2.43$) tests, $t(36) = 10.43$, $p < .001$.

Participants were also more likely to pick contestants whose easy quiz scores they had seen when predicting which contestants would score better on the third quiz. Although those who had taken the easy quiz represented 50% of the contestants participants saw, they represented 68% of contestants selected. This 68% is significantly above the 50% we would have expected, had participants perfectly predicted contestants' scores on the third quiz and only selected those, $t(70) = 7.30$, $p < .001$. It is also significantly above the 60% we might have expected if participants had been following a justifiable strategy of picking the top two scorers on the easy and difficult quizzes, and then always selecting the next best easy quiz scorer for their fifth pick, $t(70) = 3.35$, $p = .001$.

To compare the results of Experiment 3 with those of Experiments 1 and 2, we conducted a binary logistic regression in which selection was the dependent variable. The independent variables in this regression were (a) the quiz score from each contestant the participant saw and (b) the difficulty of that quiz, as measured by the mean score. Consistent with our hypothesis and with the findings of the other experiments, the results reveal that the contestant's actual score was weighted more heavily ($B = 1.61$, $SE = .13$, $p < .001$) than was the difficulty of their quiz ($B = -1.37$, $SE = .12$, $p < .001$), $\chi^2(1) = 81.10$, $p < .001$. In this case, the discounting effect due to quiz ease was 85% of the size of the effect of quiz performance.

Specifically, what this means is that going from an average score on the difficult quiz (1.11 out of 10) to an average score on the easy quiz (8.78 out of 10) increases a contestant's probability of being selected from 27% to 70%. This effect is illustrated in Figure 1. To construct this graph, we conducted two binary logistic regressions using quiz score performance to predict the probability of being selected. One regression used easy quiz scores and another used difficult quiz scores. The results show a large effect for quiz difficulty, where easy quiz takers were substantially more likely to be chosen regardless of their relative performance on the quiz.

When participants were then asked to explicitly compare the virtues of the easy and the difficult quizzes, participants rated the difficult quiz as a better test of ability than the simple quiz, as indicated by the fact that each rating is above the rating scale's midpoint of 4 ($M = 4.54$, $SD = 1.95$), $t(70) = 2.32$, $p = .024$. They also rated the difficult test as more fair than the simple quiz ($M = 4.44$, $SD = 1.87$), $t(70) = 1.97$, $p = .053$, and as a better predictor of future performance than the easy quiz ($M = 4.75$, $SD = 1.90$), $t(70) = 3.31$, $p = .001$. It would appear that the only way to reconcile these ratings with participants' systematic preference for takers of the easy quiz is that they believed that the difficult test was better at revealing just how inept the takers of the difficult quiz were.

General Discussion

The results of the three experiments we present here are consistent in showing that information about the strength of the situation—in this case, task difficulty—tends not to be used sufficiently to discount information about an individual's performance even when performance and the situation's influence on it are obvious and quantified. As a result, students from institutions with lenient grading benefit from their high grades.

Contributions to Theory and Research

Our results suggest that neither underestimating the impact of the situation nor overestimating the impact of behavior are necessary conditions for producing the correspondence bias. Our participants did not need to estimate either in the research paradigms we employed. In addition, the results suggest that the correspondence bias can persist even when information about both behavior and situation are known with equal clarity and are presented in the same format and modality. This is testament to the bias's robustness. Perhaps more importantly, the present results afford a useful quantification of the size of the correspondence bias. Its hallmark is that the judgmental weight attached to the situation is lower than the weight attached to behavior.

In our results, we find that the situation is weighted between 34% and 85% of what it should be. Clearly, there are factors that varied between our experiments that influenced the size of the correspondence bias. Identifying these moderators of the effect size will be a useful task for future research. Another potential avenue for future research is to investigate the moderators of the effect of performance relative to peers. In Experiment 1, our results suggested that outperforming peers had the strongest effect on those from lenient-grading institutions. The results from Experiment 2 suggested that outperforming peers was most important at institutions of moderate grading standards. In Experiment 3 we found that outperforming peers was most valuable on hard tests, where the grading standards were toughest. We would only note that none of these interactions eliminated the benefits of lenient grading and task ease. Our goal in this research was to document the effect of situational influences on perceptions of individual performance using experimental designs that allow us to estimate exactly how much the correspondence bias affects judgments of performance. This represents a step beyond prior work on the topic.

This research also contributes to the prior literature on the correspondence bias by precisely showing why such an effect matters for real decisions by experienced professionals. The same effects documented here appear in actual admissions decisions (Swift, Moore, Sharek, & Gino, 2010). Moreover, when professional admissions staffers are asked to make the same judgments as those made by participants in our laboratory experiments, the results are indistinguishable from those of the student participants presented in this article: Both display the correspondence bias to a similar degree (Swift et al., 2010). Consequently, graduate programs are collectively choosing to select students who come from undergraduate programs with lenient grading rather than selecting the best students. The consequences could be substantial for both the quality of students selected and the quality of those graduate programs (Berry & Sackett, 2009).

Practical Implications

Three experiments supported the hypothesis that people rely heavily on nominal performance (such as GPA) as an indicator of success while failing to sufficiently take into account information about the distributions of performances from which it came. The question of whether people—especially decision makers such as admissions officers—can correct for the correspondence bias in judgments of others is fundamental to problems of social inequality and class mobility. A meritocracy depends on being able to identify merit that, in reality, is often clouded by variations in circumstance. Given persistent disparities in the difficulty of the conditions into which Americans are born (Neckerman & Torche, 2007; Wilson, 1990), it is essential for colleges and employers to be able to adjust their estimations of ability appropriately based on the ease with which individual promise can result in nominal performance. The results of the present study suggest pessimism—people will too often be judged based on their nominal performances, with insufficient regard to the difficulty of achieving those results.

Can we offer constructive advice to those in admissions offices, personnel offices, and hiring committees responsible for making such selection decisions? We believe we can. The advice is consistent with a great deal of other evidence that demonstrates the superiority of statistical over intuitive judgment (Dawes, 1972, 1979; Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996). The advice is that decision makers should not rely exclusively on their unaided intuitive judgments and they should instead obtain the help of a computational decision tool. In this case, what that means is simply that GPA ought to be considered exclusively as a percentile rank or *z*-score deviation from the mean at that person's school. Given the power and persistence of the effect we document, the implication seems to be that decision makers should not be allowed to see raw scores or absolute GPA and should only see the standardized score that shows relative performance.

Acknowledgments

DeVito, Bill Mangan, and Jessica Wisdom for help with data collection.

Declaration of Conflicting Interests

Funding

Notes

1. This is assuming similarity across institutions in both (a) institution quality and (b) within-institution variability.
2. This is also the case for Experiment 2.

References

Attewell, P. (2001). The winner-take-all high school: Organizational adaptations to educational stratificaiton. *Sociology of Education, 74*, 267-295.

Bagues, M. F., Sylos Labini, M., & Zinovyeva, N. (2008). Differential grading standards and university funding: Evidence from Italy. *CESifo Economic Studies, 54*, 149-176.

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. *Psychological Science*, *20*, 822-830.

Davis-Blake, A., & Pfeffer, J. (1989). Just a mirage: The search for dispositional effects in organizational research. *Academy of Management Review, 14*, 385-400.

Dawes, R. M. (1972). In defense of "bootstrapping." *American Psychologist, 27*, 773-774.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

Epley, N., Savitsky, K., & Gilovich, T. (2002). Empathy neglect: Reconciling the spotlight effect and the correspondence bias. *Journal of Personality and Social Psychology, 83*, 300-312.

Gilbert, D. T. (1994). Attribution and interpersonal perception. In A. Tesser (Ed.), *Advanced social psychology* (pp. 99-147). New York, NY: McGraw-Hill.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21-38.

Goldman, R. D., & Widawski, M. H. (1976). A within-subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement, 36*, 381-390.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293-323.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*, 1-24.

Jones, E. E., & Nisbett, R. E. (1971). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79-94). Morristown, NJ: General Learning Press.

Lewin, K. (1951). *Field theory in social science*. New York, NY: Harper & Row.

Neckerman, K. M., & Torche, F. (2007). Inequality: Causes and consequences. *Annual Review of Sociology, 33*, 335-357.

Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology, 32*, 932-943.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173-220). New York, NY: Academic Press.

Ross, L., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology, 35*, 485-494.

Staw, B. M., Bell, N. E., & Clausen, J. A. (1986). The dispositional approach to job attitudes: A lifetime longitudinal test. *Administrative Science Quarterly, 31*, 56-68.

Staw, B. M., Sutton, R. I., & Pelled, L. H. (1994). Employee positive emotion and favorable outcomes at the workplace. *Organization Science, 5*, 51-71.

Swift, S. A., Moore, D. A., Sharek, Z., & Gino, F. (2010). *Seeing through performance: Attribution errors in performance evaluation by experts.* Unpublished manuscript.

Van Boven, L., Kamada, A., & Gilovich, T. (1999). The perceiver as perceived: Everyday intuitions about the correspondence bias. *Journal of Personality and Social Psychology, 77*, 1188-1199.

Wilson, W. J. (1990). *The truly disadvantaged: The inner city, the underclass, and public policy*. Chicago, IL: University of Chicago Press.