Debiasing Training Transfers to Improve Decision Making in the Field

Anne-Laure Sellier, HEC Paris

Irene Scopelliti, City University of London

Carey K. Morewedge, Boston University

Abstract

The primary objection to debiasing training interventions is a lack of evidence that they transfer to improve decision making in field settings, where reminders of bias are absent. We gave graduate students in three professional programs (*N* = 290) a one-shot training intervention that reduces confirmation bias in laboratory experiments. Natural variance in the training schedule assigned participants to receive training before or after solving an unannounced business case modeled on the decision to launch the Space Shuttle Challenger. We used case solutions to surreptitiously measure their susceptibility to confirmation bias. Trained participants were 29% less likely to choose the inferior hypothesis-confirming solution than untrained participants. Analysis of case write-ups suggests that a reduction in confirmatory hypothesis testing accounts for their improved decision making in the case. The results provide promising evidence that debiasing training effects transfer to field settings and can improve decision making.

Keywords: Debiasing, Training, Confirmation Bias, Confirmatory Hypothesis Testing, Judgment and Decision Making

Biases in judgment and decision making affect experts and novices alike, yet there is considerable variation in individual decision making ability (e.g., Cokely, Feltz, Ghazal, Allan, Petrova, & Garcia-Retamero, 2018; Frederick, 2005; Mellers et al., 2015; Scopelliti, Min, McCormick, Kassam, & Morewedge, 2018; Scopelliti et al., 2015). To the extent that this variance reflects malleable differences, training interventions could be an effective and scalable way to debias and improve human reasoning. Successful training interventions are particularly well suited to generalize and improve reasoning in new and old contexts where other interventions such as nudges and incentives have not or cannot be implemented.

Early tests of training interventions found they reliably improved reasoning in specific domains, but often failed to generalize to novel problems and contexts unless training was extensive (e.g., statistics courses) or trainees knew they were being tested (Fischoff, 1982; Fong, Krantz, & Nisbett, 1986; Fong & Nisbett, 1991; Milkman, Chugh, & Bazerman, 2009). Postmortems of this research program have argued that training may teach people to recognize bias and to correct biased inferences when prompted, but its effects will not transfer to the field where reminders of bias are absent (Kahneman & Egan, 2011). This view suggests that, at best, debiasing training effects are domain specific (Milkman et al., 2009). At worst, training may be a Hawthorne effect or could impair decision making by interfering with generally useful heuristics (e.g., Arkes, 1991).

We report a field experiment examining whether the debiasing effects of one-shot serious game-based training interventions, which exhibited large and long-lasting debiasing effects in laboratory contexts (Morewedge et al., 2015), transfer to improve decisions in the field. The games incorporate four debiasing strategies proposed by Fischhoff (1982): warnings about bias, teaching its directionality, providing feedback, and extensive coaching and training. The large effects of the games appear to be due to the personalized feedback and practice they deliver to players across multiple bias-eliciting paradigms and domains (Morewedge et al., 2015). We administered one game-based training intervention targeting confirmation bias to business students before or after they completed, in one of their courses, an unannounced business case that measured their susceptibility to confirmation bias. No explicit connection was made between the intervention and the case. We analyzed case solutions to measure if the debiasing effects of the training intervention transferred to reduce confirmation bias in this different field decision, which required generalization of training to a new paradigm and domain.

Method

Open Science Practices. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The case, all bias measures, and data are available at: osf.io/mnz8j. We do not provide readers with access to the proprietary intervention, but a general summary is publically available (Symborski et al. 2017).

Participants. Three hundred and eighteen graduate business students at HEC Paris were enrolled in a course in which we administered a modified version of the case, *Carter Racing*. All students were offered free debiasing training through a special program run by the school. All but two students volunteered to receive it (N = 316; 101 women; M_{age} = 28.24 years, SD = 3.69). Participants included students enrolled in three different graduate programs: students completing a Master in Business Administration (n = 217), an MSc in Entrepreneurship (n = 64), or an MSc in Strategic Management (n = 35).

Training Intervention. The one-shot debiasing intervention consisted of playing a serious game, *Missing: The Pursuit of Terry Hughes*. Playing this video game once has been shown to significantly reduce the propensity of players to exhibit confirmation bias, bias blind spot, and correspondence bias on individual difference scales measuring each construct both immediately, from pretest to posttest in laboratory contexts, and as long as three months after game play in online follow-up surveys (Morewedge et al., 2015).

Game players act as amateur detectives and search for a missing neighbor, who is embroiled in a fraud committed by her employer, a pharmaceutical company. There are three episodes (i.e., levels), each with a play-teach loop structure. Players make bias-eliciting judgments and decisions during game “play.” Eight decisions during game play elicit confirmation bias (i.e., three in “Episode 1,” three in “Episode 2,” and two in “Episode 3”). At the end of each episode participants receive training in the “teach” portion of the game through an after action review. In the review, experts define the three biases targeted by the game and provide strategies to mitigate each bias. Narrative examples of cases in which professionals exhibited the bias are then provided (e.g., the conclusion of intelligence analysts that Iraq possessed WMD’s before the Iraq War). Next, participants receive personalized feedback on the degree of bias they exhibit in each scenario in that episode of the game, and how it might have been avoided. At the end of this portion of training, participants complete practice problems for confirmation bias (and the other two biases) and receive immediate feedback on their performance on those problems before the next level begins or the game ends.

The game uses three paradigms to elicit and teach game players about confirmation bias. The first is the Wason Four Card Selection task (Wason, 1968). Bias mitigation is taught by explaining the greater value of searching for hypothesis-disconfirming evidence rather than confirming evidence. In more colloquial language, players are taught when testing a rule with the structure, "If *P*, then *Q*," that testing for instances of "*P* and ~*Q*" allows one to make a more valid inference than does testing for instances of "*P* and *Q*". The second paradigm is based on Tschirgi's (1980) multivariate cause identification paradigm. Participants are informed of an outcome (e.g., a cake turned out well) that could have been caused by any of three variables (e.g., instead of typical ingredients, using margarine, honey, or brown wheat flour as substitutes). They are then asked how they would test whether a focal variable (e.g., using honey) caused the outcome. Participants are taught to test whether the outcome will replicate when they remove the focal variable and hold the other factors constant (e.g., make a cake using margarine, sugar, and brown wheat flour). The third paradigm is based on Snyder and Swann's (1978) trait hypothesis testing paradigm. Participants are taught, when searching for evidence that might confirm or disconfirm a focal hypothesis (e.g., testing if a person is an extravert), the value of searching for hypothesis-disconfirming evidence (e.g., asking questions that test if she is an introvert).

Course Case. We administered a modified version of *Carter Racing* to all students in the three programs within one of their courses (Brittain & Sitkin, 1988). The case elicits confirmation bias in decision making under uncertainty: a tendency to preferentially test, search for, and interpret evidence supporting existing beliefs, hypotheses, and opinions (e.g., Nickerson, 1998). In the case, modeled on the decision to launch the Space Shuttle Challenger, each student acts as the lead of an automotive racing team making a high-stakes binary decision: remain in a race despite a risk of an expensive engine failure (the hypothesis-confirming choice) or withdraw from the race, which would incur a significant sure cost (the hypothesis-disconfirming choice).

The case narrative and payoff structure, if engine failure is deemed unlikely, favor remaining in the race. By contrast, the data provided in the case reveals that withdrawing from the race is an objectively superior option. Engine failure is near certain at the low temperature recorded at the start of the race. That conclusion, however, requires students to compare two graphs: one depicting engine failures at different temperatures and one depicting races with no failure at different temperatures. These are plotted on y-axes with different scales (Exhibits 1 and 2, respectively; see supplemental methods). If students first examine Exhibit 1, the relationship between temperature and engine failure would appear inconclusive. Confirmatory hypothesis testing might then lead them to ignore temperature concerns and base their decision on the favorable payoffs for racing. Only if students continued to compare the two exhibits do the dangers of racing become fully clear.

We renamed and modified the case slightly to make the solution impossible to find online and increase comprehension for our diverse international sample (e.g., temperatures were presented in Celsius, not Fahrenheit). We note that the case structure was considerably different from the structure of the paradigms used to test and teach confirmation bias in the debiasing training intervention.

Procedure. University administration offered a free, serious game-based training intervention to all students in three different degree programs that they were told could improve their "managerial decision making ability." Volunteers signed up online for a single training session from a set of sessions offered over a twenty-day period. Students could sign up for any session available when the school announced the free training opportunity. The intervention was administered in a university computer laboratory, where groups of up to 20 students played the game at a time, in private, on separate computers. All students completed at least two levels of the game (i.e., were exposed to training for all three biases), and played for 80-100 minutes.

Between 6 and 49 days following the start of the gaming sessions, participants individually solved a modified version of the Carter Racing business case in of one of their regularly scheduled classes. We exploited natural variation in the time when participants completed gaming sessions to test whether the intervention improved decision-making in the complex business case, which was administered within one course in each participant's program. The case was not announced on the syllabi of the courses in which it was administered, the faculty administering the training and case were different, and no other connection was made between the case and the intervention. Thus, participants could not have known the game and case were related and could not plan to play the game to improve their case performance. The timing of the session in which participants received training determined their assignment to *Trained* and *Untrained* conditions, respectively. Average lag in the trained condition between training and case completion was 17.96 days (SD = 19.86).

Participants first submitted their case solution (i.e., race or withdraw) and a written justification of their solution. They then reported their decision confidence on a 7-point scale (1 = 50% confidence, 7 = 100% confidence). After the participants finished the case, they completed two pencil and paper scale-based measures assessing their susceptibility to the two other cognitive biases treated in the game: a 14-item measure of bias blind spot (Scopelliti et al., 2015) and a 10-item measure of correspondence bias (the Neglect of External Demands, or NED; Scopelliti et al., 2018). These measures served as manipulation checks for the efficacy of the debiasing training; the game has been shown to reduce bias on both scales in previous research. We also included a 3-item Cognitive Reflection Task (Frederick, 2005), a measure of the propensity to reflect upon seemingly intuitive answers. Comparing its effect size relative to that of the intervention on decision to race could thus serve as an informative benchmark. Participants then reported their age, gender, years of work experience, and the degree they were pursuing. Finally, since participants were not all native English speakers, they reported the extent to which they experienced difficulty comprehending the language used in the case on a 7-point scale, (1 = not at all; 7 = very much). We later were able to collect the cumulative GPA for all but one participant from the university registrar, and GMAT scores for participants with an exam score in their official record (n = 208).

Only after all participants solved the case and all gaming sessions concluded were participants fully debriefed in their classes. The case and training were thus administered in different contexts (i.e., classroom versus laboratory), domains (i.e., automotive racing versus corporate fraud), with different problem structures (i.e., a binary case decision versus multiple choice problems and scale ratings). The design then conservatively tested, when bias is surreptitiously measured, whether debiasing training effects transfer to field settings and improve decision-making in a novel context and paradigm.

Results

Exclusions and control measures. We only retained participants who were certain they were not familiar with the case. This filter excluded 26 participants from subsequent analyses. We report analyses on the remaining 290 participants.

Participants in the trained (n = 182) and untrained (n = 108) conditions did not differ in age, years of work experience, or English proficiency (Fs < 1). The proportions of male participants in the trained (73.1%) and in the untrained condition (63.0%) were not significantly different (95% $CI_{difference}$ = [-.8%, 21.2%], $\chi^2(1, N = 290) = 3.26$, $p = .071$). Twenty-two participants in the untrained control condition (20.4%) solved the case but did not complete a gaming session; they signed up for a session but did not show up for that session. Excluding them from the analyses does not substantively change the results.[1]

Scale measures. We first examined if the effects of the debiasing intervention replicated on the two scale-based measures administered immediately after the case was solved. Of the full sample, 225 participants completed both the bias blind spot scale (BBS; Cronbach's $\alpha = .81$) and the correspondence bias scale (NED; Cronbach's $\alpha = .90$). One group of 65 participants solved the case in a class where the instructor did not administer these scales. Replicating previous research, trained participants exhibited significantly lower levels of bias on both scale measures than did untrained controls [BBS: $M_{Trained} = .85$, 95% CI [.75, 1.02] vs. $M_{Untrained} = 1.26$, 95% CI [1.12, 1.42], mean difference = .40, 95% CI [.20, .61], $F(1, 223) = 15.48$, $p < .001$, $d = .53$; NED: $M_{Trained} = 2.38$, 95% CI [2.17, 2.58] vs. $M_{Untrained} = 3.09$, 95% CI [2.88, 3.30], mean difference = .72, 95% CI [.42, 1.01], $F(1, 223) = 21.99$, $p < .001$; $d = .63$].

We also estimated a linear regression model of the effect of the training intervention on decision confidence controlling for the case solution chosen. Although participants who decided to race were not more confident than participants who decided not to race, $\beta = .35$, 95% CI_β = [.00, .71], $t = 1.97$, $p = .050$, the intervention reduced confidence in the solution chosen, $\beta = -.43$, 95% CI_β = [-.78, -.08], $t = -2.41$, $p = .017$. This effect was robust to the inclusion of all covariates including gender, years of work experience, English proficiency, CRT scores, and GPA.

Case solutions. Most important, we examined whether the training intervention significantly reduced the choice share of the hypothesis-confirming case solution. It did. Logistic regression revealed that trained participants were significantly less likely to choose the hypothesis-confirming decision to race (58.8%) than untrained controls (72.2%), $\beta = -.60$, Wald $\chi^2(1) = 5.23$, $p = .022$, $\exp(\beta) = .549$, 95% $CI_{\exp(\beta)} = [.33, .92]$ (Figure 1, left panel). As a test of the longevity of this training effect, we compared the 125 participants (68.7% of the intervention group) exposed to the intervention ≤ 11 days before solving the case (short lag group) to the 57 participants (31.3% of the intervention group) exposed to the intervention between 43 and 52 days before solving the case (long lag group). This split of the sample was based on a natural discontinuity in the data; the next observed lag value after 11 days was 43 days. The debiasing effects of the game were no weaker in the short (56.8%) or long lag group (63.2%), 95% $CI_{difference} = [-2.07\%, 9.10\%]$, $\chi^2 (1, N = 182) = .65$, $p = .419$.

Robustness checks. As robustness tests against selection effects, we first examined whether the training effect persisted when we included the covariates of gender, years of work experience, English proficiency, cognitive reflection (CRT) scores, and GPA (Table 1, Model 2). We also estimated a model (Table 1, Model 3) including GMAT scores as an additional covariate on the subsample for which these scores were available. In both models, the effect of the training was significant. By contrast, cognitive reflection, GMAT scores, and GPA did not predict the decision to race. These findings suggest that the effect of training on decision making in the task is not attributable to a selection effect (e.g., better decision makers completing the training intervention earlier than worse decision makers). It is interesting to note that CRT (Frederick, 2005) scores were significantly higher for trained than untrained participants, $M_{Trained} = 2.44$, 95% CI [2.31, 2.57], $M_{Untrained} = 2.18$, 95% CI [2.00, 2.36], mean difference = .26, 95% CI [.04, .48], $F(1, 288) = 5.46$, $p = .020$, $d = .28$. Because of this difference, which could be diagnostic of natural differences between the trained and untrained groups, we controlled for CRT scores in our analyses. However, it is possible that the debiasing training intervention increased the propensity to engage in cognitive reflection.

We tested for selection effects in a second way, by estimating the effect of the intervention on participants who signed up for the game within short time intervals surrounding the case date. If there was a selection effect, these participants should be more similar across such influential individual differences than the full sample of participants, and the effect of training should become weaker with the narrowing of the time interval. We selected three short time intervals surrounding the case date, and only examined participants who played the game in those intervals: between three days prior to and three days after completing the case (6-day window subsample, n = 94), between two days prior to and two days after completing the case (4-day window subsample, n = 75), and between on day prior to and one day after completing the case (2-day window subsample, n = 50). In all three time intervals, participants in the training condition were significantly less likely to choose the hypothesis-confirming decision to enter the race than were untrained controls. In the 6-day window, trained participants were significantly less likely to decide to race (48.0%) than were untrained controls (72.7%), 95% CI for the mean difference [4.90%, 41.90%], β = -1.06, Wald $\chi 2(1)$ = 5.78, p = .016, exp(β) = .35, 95% $CI_{exp(\beta)}$ = [0.15, .82]. In the 4-day window, trained participants were significantly less likely to decide to race (48.0%) than were untrained controls (76.0%), 95% CI for the mean difference [4.30%, 46.20%], β = -1.23, Wald $\chi 2(1)$ = 5.06, p = .024, exp(β) = .29, 95% $CI_{exp(\beta)}$ = [0.10, .85]. In the 2-day window, trained participants were also significantly less likely to decide to race (50.0%) than were untrained controls (85.7%), 95% CI for the mean difference [5.70%, 54.30%], β = -1.79, Wald $\chi 2(1)$ = 4.62, p = .032, exp(β) = .17, 95% $CI_{exp(\beta)}$ = [0.03, .85].

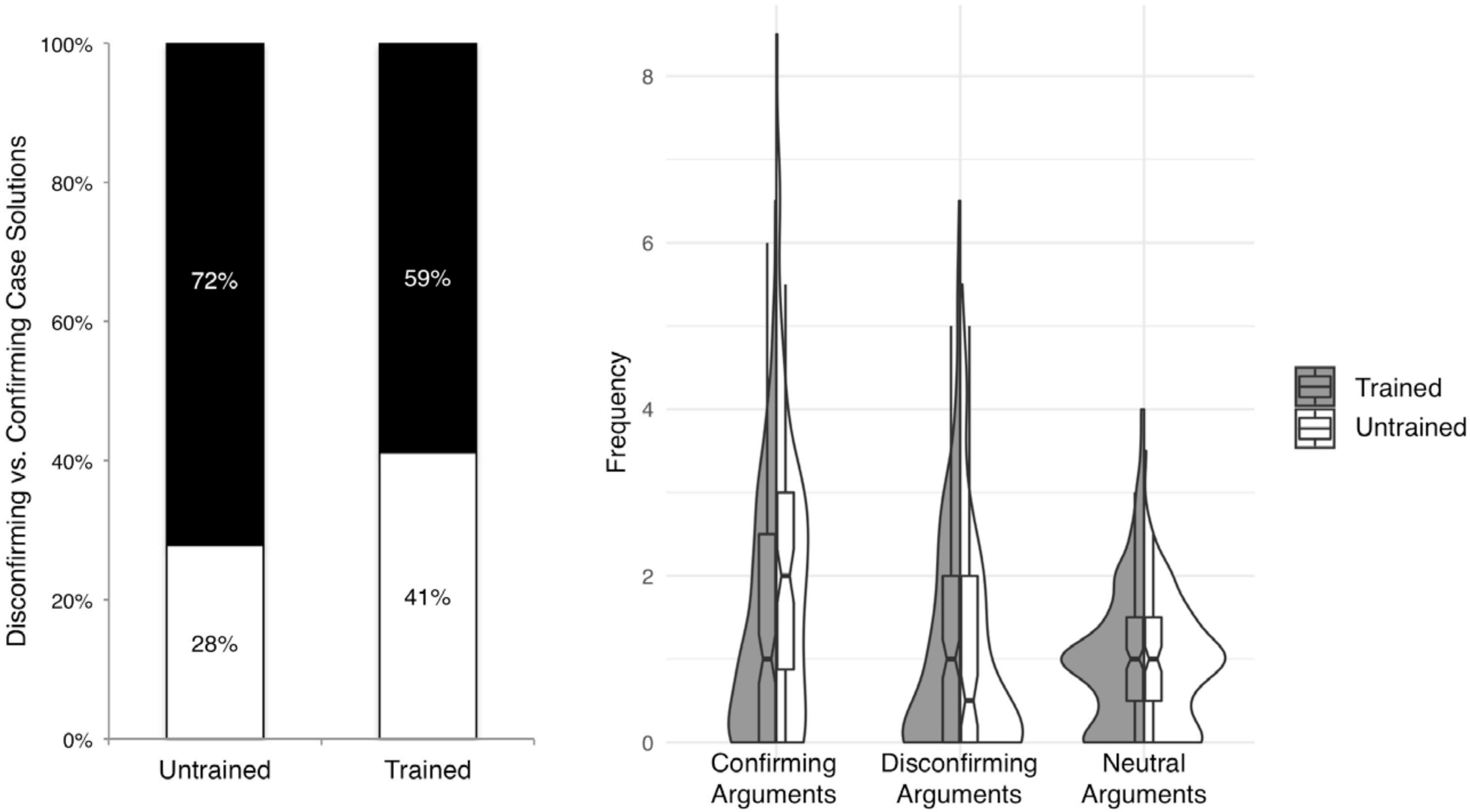

Figure 1. Left panel depicts choice share of the suboptimal hypothesis-confirming (black) and optimal hypothesis-disconfirming (white) case solutions by training condition. Right panel depicts frequency of confirming, disconfirming, and neutral arguments generated as reasons for choice of case solution by training condition. Plot width indicates the frequency of each observed value (i.e., probability density). Boxplots are centered at the median. Lower and upper hinges correspond to the first and third quartiles, respectively. The upper whisker extends from the third quartile to the largest observed value, no further than 1.5 times the interquartile range from the hinge. The lower whisker extends from the hinge to the smallest value, at most 1.5 times the interquartile range from the hinge.

Table 1. Logistic regression results and model comparisons for the decision to race

	Model 1		*Model 2*		*Model 3*		*Model 4*		*Model 5*		*Model 6*	
Variable	**B (SE)**	**exp(B)**	**B (SE)**	**exp(B)**	**B (SE)**	**exp(B)**	**B (SE)**	**exp(B)**	**B (SE)**	**exp(B)**	**B (SE)**	**exp(B)**
Intercept	.956*** (.215)	2.600	4.923* (2.378)	137.382	9.359* 4.579	11600.111	.161 (.459)	1.175	1.071 (.589)	2.919	7.515 (7.178)	1834.485
Training	-.600* (.262)	.549	-.559* (.271)	.572	-.880* (.347)	.415	-.528* (.266)	.590	-.432 (.655)	.649	-.523 (.702)	.593
Gender			.487 (.286)	1.627	.282 (.366)	1.326					.918 (.736)	2.504
Experience			-.008 (.044)	.992	.032 (.065)	1.032					-.050 (.117)	.951
Proficiency			-.113 (.086)	.893	-.179 (.106)	.836					-.062 (.203)	.940
CRT			-.094 (.145)	.911	.094 (.187)	1.098					-.248 (.390)	.780
GPA			-1.051 (.675)	.350	-.790 (1.034)	.454					-1.614 (1.988)	.199
GMAT					-.008 (.005)	.992						
Decision Confidence							.162 (.084)	1.176				
Confirming Arguments									3.020*** (.476)	20.496	2.988*** (.489)	19.854
Disconfirming Arguments									-2.776*** (.422)	.062	-2.816*** (.438)	.060
-2 Log-Likelihood	374.272		366.551		238.986		370.499		75.389		72.888	
df	1		6		7		2		3		8	
Nagelkerke R^2	.025		.057		.090		.043		.890		.894	
N	290		289		191		290		290		289	

Note: *p < .05, ***p < .001

Process tests. We next examined whether a reduction in confirmatory hypothesis testing among trained participants, relative to untrained controls, might account for their reduced propensity to choose the inferior hypothesis-consistent case solution. Two coders, blind to condition and hypotheses, coded all statements in participants' written justifications into three categories, confirming statements (i.e., for racing; ICC(2, 2) = .93; M = 1.68, 95% CI [1.50, 1.86]), disconfirming statements (i.e., against racing; ICC(2, 2) = .90; M = 1.17, 95% CI [1.01, 1.33]), and neutral statements (ICC (2, 2) = .70; M = 1.01, 95% CI [.91, 1.10]). The overall number of statements participants wrote was not significantly different across conditions, mean difference = .38, 95% CI [-.14, .90], $F(1, 288) = 2.41$, $p = .122$.[2]

A reduction in confirmatory hypothesis testing can be the outcome of two different processes, a reduction in the number of hypothesis-confirming arguments or an increase in the number of hypothesis-disconfirming arguments, $r(290) = -.21$, 95% CI [-.31, -.09], $p < .001$. We thus examined the effect of the intervention on counts of both confirming and disconfirming arguments generated by participants (counts illustrated in Figure 1, right panel). Trained participants generated significantly fewer confirming arguments than did untrained controls ($M_{Trained} = 1.45$, 95% CI [1.23, 1.66], $M_{Untrained} = 2.07$, 95% CI [1.73, 2.41], mean difference = -.63, 95% CI [-1.03, -.23], $F(1, 288) = 10.82$, $p = .001$, $d = .39$). They also generated more disconfirming arguments than did untrained controls, but the difference between the conditions was not statistically significant ($M_{Trained} = 1.23$, 95% CI [1.02, 1.43], $M_{Untrained} = 1.08$, 95% CI [.82, 1.34], mean difference = .15, 95% CI [-.18, .48], $F(1, 288) = .78$, $p = .377$, $d = .11$). This suggests that training reduced confirmatory hypothesis testing through a reduction in the number of confirming arguments generated by participants. Of course, this interpretation needs to be adopted with caution. It is possible that participants' written responses reflect post-hoc justifications of their case decisions rather than the arguments they considered before making their decisions (Nisbett & Wilson, 1977).

We next tested if a reduction in confirmatory hypothesis testing among trained participants could explain their improved decision making in the task. A logistic regression model including confirming arguments, disconfirming arguments, and the intervention as predictors of the decision to race (Table 1, Model 5), revealed that each set of arguments significantly affected, in opposing directions, the likelihood of deciding to race, $\beta_{Confirming} = 3.02$, SE = .48, Wald $\chi^2(1) = 40.31$, $p < .001$, $\exp(\beta) = 20.50$, 95% $CI_{\exp(\beta)} = [8.07, 52.07]$, $\beta_{Disconfirming} = -2.78$, SE = .42, Wald $\chi^2(1) = 43.18$, $p < .001$, $\exp(\beta) = .06$, 95% $CI_{\exp(\beta)} = [.03, .14]$, whereas in this analysis, the effect of the intervention was no longer significant, $\beta = -.43$, SE = .66, Wald $\chi^2(1) = .44$, $p = .509$, $\exp(\beta) = .65$, 95% $CI_{\exp(\beta)} = [.18, 2.34]$.

Estimating the indirect effects of the intervention (with 10,000 bootstrap resamples) through each set of arguments revealed that a reduction in the number of confirming arguments generated significantly mediated the effect of the intervention ($\beta = -1.90$, LLCI = -4.01, ULCI = -.72). The increased number of disconfirming arguments generated by trained participants, although a significant predictor of the decision to race, did not significantly mediate the effect of the intervention ($\beta = -.41$, LLCI = -1.48, ULCI = .58). Including demographic covariates (i.e., gender, years of work experience, English proficiency, cognitive reflection, and GPA) in the conditional process analysis did not alter the pattern of results. In short, the reduction in confirmatory hypothesis testing exhibited by participants who were trained beforehand appears to explain their lower likelihood of deciding to race in the case.

We also tested an alternative account of the effect of the intervention, whether debiasing training simply induced more risk aversion or conservative decision-making. Trained participants were indeed less confident in their decisions than were untrained participants, but decision confidence did not explain the effect of the intervention. When including decision confidence as a predictor in a logistic regression model examining the effect of training on the decision to race (Table 1, Model 4), the effect of confidence was not significant, $\beta = .16$, Wald $\chi 2(1) = 3.75$, $p = .053$, $\exp(\beta) = 1.18$, 95% $CI_{\exp(\beta)} = [1.00, 1.39]$, whereas the effect of training was still significant, $\beta = -.53$, Wald $\chi 2(1) = 3.93$, $p = .047$, $\exp(\beta) = .59$, 95% $CI_{\exp(\beta)} = [.35, .99]$.

Discussion

Debiasing effects of a one-shot training intervention transferred to a novel problem and context in a field setting. Trained students were 29% less likely to choose an inferior hypothesis-confirming case solution than untrained students. A reduction in confirmatory hypothesis testing appeared to explain their improved decision making in the case. The method of condition assignment obviously raises selection concerns, but they are allayed by two analyses. First, controlling for participants' GPA, GMAT, and CRT scores did not mitigate the training effect. Second, the training effect was stable even within short observation windows of 2, 4, and 6 days around the intervention, where samples should be least susceptible to selection bias.

Our results address two major critiques of training interventions. As heuristics and biases are often adaptive (Arkes, 1991), training could impair judgment and decision-making. We found debiasing training improved a decision in the field—it increased preferences for the optimal hypothesis-disconfirming solution to a risky managerial decision. Second, we found that debiasing training appears to have transferred without reminders or the influence of a Hawthorne effect (Kahneman & Egan, 2011). Training influenced the case decision in the absence of an explicit connection between the training intervention and case.

More research is needed to explain why this game-based training intervention transferred more effectively than has specialized expert training (Milkman et al., 2009). Games may be uniquely engaging training interventions. Providing intensive practice and feedback is another possibility. It has been present in other successful training interventions (Fischhoff, 1982), and differentiated this intervention from a similar but less effective instructional video-based training intervention in our previous work (Morewedge et al., 2015). A third possibility is the breadth of the training that the intervention delivered. Transfer may be facilitated when training describes biases and mitigating strategies at an abstract level, and includes practice mapping those strategies to different paradigms and domains.

References

Arkes, H.R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110*(3), 486-498.

Brittain, J., & Sitkin S. (1988). Carter Racing Case and Teaching Notes. *Stanford Case System* (#SOB-24), Graduate School of Business, Stanford University.

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). Skilled Decision Theory: From Intelligence to Numeracy and Expertise. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (2nd Eds.), *Cambridge Handbook of Expertise and Expert Performance*. New York, NY: Cambridge University Press.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds). *Judgment under Uncertainty: Heuristics and Biases* (pp. 422-444), Cambridge, UK: Cambridge University Press.

Fong, D. R., Krantz, D. H., & Nisbett, R. E. (1986), Teaching reasoning. *Cognitive Psychology, 18*(3)**,** 253-292.

Fong, G.T., & Nisbett, R.E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120*(1), 34-45.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42.

Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., & Ungar, L. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*(3), 267-281.

Milkman, K.L., Chugh, D., Bazerman, M.H. (2009). How can decision making be improved?. *Perspectives on Psychological Science*, *4*(4), 379-383.

Morewedge, C., Yoon, H, Scopelliti, I., Symborski, C.W., Horris, J.H., Hassam, K. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129-140.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84(3)*, 231-259.

Scopelliti, I., Morewedge, C.K., McCormick, E., Min, L., Lebrecht, S., & Kassam, K.S. (2015). Bias blind spot. Structure, measurement, and consequences. *Management Science,* *61*(10), 2468-2486.

Scopelliti, I., Min, L., McCormick, E., Kassam, K.S., & Morewedge, C.K. (2018). Individual differences in correspondence bias: measurement, consequences, and correction of biased interpersonal attributions. *Management Science, 64*(4), 1879-1910.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*(11), 1202.

Symborski, C.W., Barton, M., Quinn, M.M., Korris, J.H., Kassam, K.S., & Morewedge, C.K. (2017). The Design and Development of Serious Games Using Iterative Evaluation. *Games and Culture, 12*(3), 252-268.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51(1), 1-10.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273-281.

Acknowledgments

We thank Stephen Baum, Alain Bloch, Alejandra Cervio, Marie-Josée Durand, James Korris, Andrea Masini, Laurence Lehmann-Ortega, Mathis Schulte, Carl Symborski, and the HEC IT department for their invaluable assistance, and HEC Foundation for its financial support.

Author Contributions

A. L. Sellier: conceptualization, experiment supervision, methodology, writing

I. Scopelliti: statistical analyses, figures, writing

C. K. Morewedge: coding supervision, conceptualization, figures, methodology, writing

Competing interests: Authors declare no competing interests.

Footnotes

1. Participants who did not play the game were slightly older than those who did, M_{NoGame} = 29.91 years, 95% CI [28.07, 31.75], vs. M_{Game} = 27.67 years, 95% CI [26.91, 28.43], mean difference = 2.23 years, 95% CI [.26, 4.21], $F(1, 106) = 6.48$, $p = .012$, but did not differ in years of work experience M_{NoGame} = 5.93 years, 95% CI [4.57, 7.29], vs. M_{Game} = 4.74 years, 95% CI [4.12, 5.36], mean difference = 1.19 years, 95% CI [-.29, 2.67], $F(1, 106) = 2.88$, $p = .092$, or gender, $\chi^2(1, N = 108) = .01$, $p = .942$. Most important, they did not differ with respect to the main dependent variable, i.e., the case decision, $\chi^2(1, N = 108) = .35$, $p = .553$.
2. For an exploratory analysis, coders also rated mention of temperature on a 3-point scale, not at all (1), mentioned temperature (2), and incorporated temperature in an argument to

race or not race (3). Note that consideration of temperature could be accurately be used to justify withdrawing, or inaccurately be used to support remaining in the race. Coders exhibited high agreement; ICC (2, 2) = .86; M = 1.96, 95% CI [1.88, 2.03]. Attention to temperature was not different across conditions, mean difference = -.08, 95% CI [-.23, .07], F(1, 288) = 1.24, p = .267, suggesting that participants read the case at similar levels of depth in both the trained and untrained condition.