Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–16

DOI: 10.1177/0956797617740646
www.psychologicalscience.org/PS
SAGE

Easier Seen Than Done: Merely Watching Others Perform Can Foster an Illusion of Skill Acquisition

Michael Kardas and Ed O'Brien
Booth School of Business, The University of Chicago

Abstract

Modern technologies such as YouTube afford unprecedented access to the skilled performances of other people. Six experiments (N = 2,225) reveal that repeatedly watching others can foster an illusion of skill acquisition. The more people merely watch others perform (without actually practicing themselves), the more they nonetheless believe they could perform the skill, too (Experiment 1). However, people's actual abilities—from throwing darts and doing the moonwalk to playing an online game—do not improve after merely watching others, despite predictions to the contrary (Experiments 2–4). What do viewers see that makes them think they are learning? We found that extensive viewing allows people to track what steps to take (Experiment 5) but not how those steps feel when taking them. Accordingly, experiencing a "taste" of performing attenuates the illusion: Watching others juggle but then holding the pins oneself tempers perceived change in one's own ability (Experiment 6). These findings highlight unforeseen problems for self-assessment when watching other people.

Keywords

self-assessment, empathy gap, repeated exposure, open data, open materials, preregistered

Received 3/30/17; Revision accepted 10/5/17

> One must learn by doing the thing; though you think you know it, you have no certainty until you try.
>
> —Sophocles (~500 BC/2015, p. 191)

Although people have extolled learning by *doing* for centuries, modern technologies have allowed learning by *seeing* to proliferate. YouTube houses millions of instructional videos depicting complex techniques from playing guitar to dancing ballet. Ratings for professional sports have reached record numbers by streaming onto phones and on-demand services (Statista, 2017a). SyberVision, a highly popular instructional video provider, promises "the more you see and hear pure movement . . . the more likely you are to perform it as a conditioned reflex" (cited in Druckman & Swets, 1988, p. 7).

Watching others is enjoyable and convenient, but people typically cannot master new skills from sight alone, even after watching from multiple angles and in slow motion (Austin & Miller, 1992). Instead, people acquire skills not merely by watching but by doing: practicing and performing themselves (Ericsson, Krampe, & Tesch-Römer, 1993; Kolb, 2014; Newell, 1991; Ullén, Hambrick, & Mosing, 2016; Willingham, 1998; Wulf, Shea, & Lewthwaite, 2010).

Alas, when people want to learn a skill, where do they begin? Many people likely begin by merely watching others, whether by choice (e.g., the ease of loading a video online) or necessity (e.g., lacking the equipment or confidence to jump right in). In a preregistered survey (see the Supplemental Material available online), we asked 500 participants to indicate which form of help for learning new skills they seek first and use most, and which they believe is most widely available, easiest to process, and most effective. For each, they chose one of five options: *watching others perform it, reading text-based instructions, hearing verbal instructions,*

Corresponding Author:
Michael Kardas, Booth School of Business, The University of Chicago, 5807 South Woodlawn Ave., Chicago, IL 60637
E-mail: mkardas@ChicagoBooth.edu

other, or *all options equal*. Watching others was reported to be the first-sought (62.80%) and most-used learning aid (69.20%) and was perceived as most available (48.20%), easiest to process (74.60%), and most effective (72.20%; all critical ps ≤ .002).

Although people may have good intentions when trying to learn by watching others, we explored unforeseen consequences of doing so: When people repeatedly watch others perform before ever attempting the skill themselves, they may overestimate the degree to which they can perform the skill, which is what we call an illusion of skill acquisition. This phenomenon is potentially important, because perceptions of learning likely guide choices about what skills to attempt and when. Although boosted confidence might encourage people to try activities they would otherwise avoid (Bandura, 1977), perceptions of learning that exceed actual changes in ability could cause viewers to budget too little time for practice or hastily attempt risky activities, naive to their low chances of success (especially on initial attempts). People today have ubiquitous outlets to learn by watching others, but merely watching others may problematically inflate self-assessments.

Why might people overestimate how much they have learned from merely watching? Watching gives people vivid, direct access to the performer's actions and hence provides insight about what, exactly, to do. Furthermore, watching a performance is dynamic: The more people watch, the more fluently these actions are processed (Song & Schwarz, 2008; Weaver, Garcia, Schwarz, & Miller, 2007), the less surprising they seem (Campbell, O'Brien, Van Boven, Schwarz, & Ubel, 2014), the greater the number of actions that are noticed (Scully & Newell, 1985), and so on. All of this added information may lead viewers to believe they have "got it" ("I bet I could do that!"). However, no matter how many times people watch a performance, they never gain one critical piece: the feeling of doing. Subtleties of performing are difficult to detect by sight alone (Adams, 1984), and the kinesthetic, sensory, and emotional states evoked within the moment of performing are difficult to mentally simulate (Van Boven, Loewenstein, Dunning, & Nordgren, 2013). If viewers do not fully adjust for this gap between seeing (tracking what the performance looks like) and doing (experiencing what the performance feels like), they may come away feeling they have learned sufficiently diagnostic information to perform the skill themselves—but learning what the steps are may be insufficient without incorporating how those steps actually feel on taking them.

In six experiments, we explored this hypothesis. First, we tested whether repeatedly watching others increases viewers' belief that they can perform the skill themselves (Experiment 1). Next, we tested whether these perceptions are mistaken: Mere watching may not translate into better actual performance (Experiments 2–4). Finally, we tested mechanisms. Watching may inflate perceived learning because viewers believe that they have gained sufficient insight from tracking the performer's actions alone (Experiment 5); conversely, experiencing a "taste" of the performance should attenuate the effect if it is indeed driven by the experiential gap between seeing and doing (Experiment 6).

Experiment 1: Repeated Watching and Perceived Ability

First, we documented the basic effect. We hypothesized that the more people merely watch others, the more they believe they can perform the skill themselves. Moreover, we compared the effect of extensively watching with extensively reading or thinking about the skill, highlighting its potentially unique role in inflating perceived abilities.

Method

In this and all of our experiments, we predetermined sample sizes of at least 50 participants per cell (Simmons, Nelson, & Simonsohn, 2018), and we doubled this number or more for online experiments. This matches many cell sizes in past research using similar designs (e.g., about 35 per cell in Carpenter, Wilford, Kornell, & Mullaney, 2013; about 30 per cell in Andrieux & Proteau, 2016; Baumeister, Alquist, & Vohs, 2015). In addition, we conducted power analyses using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). Cells of 50 provide 80% power to detect an effect size (d) of 0.57 (two-tailed, $\alpha = .05$). The average size of the critical effect across experiments was 0.60 ($N = 2{,}225$), achieving 84% power. Our experiments appear to have been amply powered.

We report all manipulations, measures, and exclusions. All data and materials are available on the Open Science Framework (OSF). The majority of our experiments were preregistered (Experiments 1, 3, 4, and 6); all of these files can be found at https://osf.io/h49y7/.

Participants. Participants ($N = 1{,}003$) were recruited from Amazon's Mechanical Turk (age: $M = 36.07$ years, $SD = 11.87$; 56.00% female; 74.18% Caucasian) to complete the study for $0.75.[1]

Procedure. Participants assessed their own abilities to perform the "tablecloth trick." In the trick, performers stand at the edge of a table with a tablecloth and several dishes and must pull the tablecloth off the table without upsetting the dishes. We chose this trick because it is

unlikely that participants could actually (a) practice during the study, thereby isolating the effects of our manipulations, or (b) learn to perform this complex trick merely from watching others many times, providing a more conservative test.

All participants were shown the same image of a set table with a spot marked where they were to hypothetically stand (see the materials at OSF). They were told to imagine encountering this exact table and being given one chance at the trick, attempted right there and then without any other practice or experience, as if we were assessing their natural ability to perform it. Participants were asked to rate from 1 (*I feel there's no chance at all I'd succeed on this attempt*) to 7 (*I feel I'd definitely succeed without a doubt on this attempt*) how well they would do via the following item: "You jump in and give the trick ONE SHOT yourself. What do you feel are the chances that you'd successfully pull it off?" This was our dependent variable.

Before making an estimate, however, each participant was randomly assigned to one of six training conditions in a 3 (type of exposure: watch, read, think) × 2 (amount of exposure: low, high) between-subjects design.

In the watch conditions, participants were told that they would be given more information about the trick before making their estimate. They clicked to the next page and saw a video of a man performing the trick himself (see the materials at OSF). The video was displayed for either 5 s (which showed the trick once in full; low-exposure version) or 60 s (which repeated this trick 20 times; high-exposure version). On completion, the page automatically continued, and participants made their estimate. These were the critical conditions. We hypothesized that merely watching many times would lead people to rate their own chance of success as significantly improved.

In the read conditions, participants were also told they would be given more information about the trick before making their estimate. They clicked to the next page and saw step-by-step written instructions for how to perform the trick (see the materials at OSF). The instructions displayed for either 5 s (low-exposure version) or 60 s (high-exposure version), matching the timing of the watch conditions. When the timer was up, the page automatically continued, and participants made their estimate. This procedure allowed us to assess whether (a) overexposure to other kinds of additional information about the trick would boost perceived abilities or (b) overexposure to watching is indeed critical. It also helped rule out general demand in the design (e.g., being given more time may lead participants to infer that they should give higher ratings or allow participants more time to simulate the trick with their hands). Reading text-based instructions was the second most highly rated learning aid in our survey, across all measures (see the introduction and the Supplemental Material).

In the think conditions, participants were given additional time in between reading the scenario and making their estimate but no actual additional information about the trick itself. When they clicked the continue button to load the page with the estimate item, they saw the following message: "Our survey is loading. The page will automatically continue when the timer expires." This loading screen was displayed for either 5 s (low-exposure condition) or 60 s (high-exposure condition), matching the timing of the other conditions. The page then continued automatically, and participants made their estimate. This procedure provided a baseline with no external learning aid, testing whether merely having ample time increases perceived abilities (e.g., inferred demand from the design, ample time to simulate and imagine).

After making their estimate, all participants responded to an attention check, "What kind of 'more information' did we give you?" (forced-choice between three items describing each type of exposure condition), and a manipulation check, "About how much of this 'more information' do you feel you were given?" (1 = *very little/went by quickly*, 7 = *a lot/displayed for a while*). Last, all participants reported whether they had any technical difficulties (yes/no) and whether they had ever previously attempted a tablecloth trick in their everyday lives (yes/no).

Results

Only 9.70% of participants failed the attention check, and 16.10% of participants reported they had previously attempted a tablecloth trick. We included all participants to maximize power. We conducted a multivariate general linear model (GLM) with type of exposure, amount of exposure, and the Type × Amount interaction as independent variables. The key ability measure and the manipulation check were dependent variables.

The basic effect: perceived ability. For our primary results, there was a main effect of type, $F(2, 997) = 33.29$, $p < .001$, $\eta^2 = .06$, and a main effect of amount; specifically, high exposure ($M = 3.17$, $SD = 1.66$) versus low exposure ($M = 2.89$, $SD = 1.61$) generally inflated participants' beliefs that they could perform the trick themselves, $F(1, 997) = 8.87$, $p = .003$, $\eta^2 = .01$. Critically, however, this depended on the type of information they were exposed to, as evidenced by a significant interaction, $F(2, 997) = 6.05$, $p = .002$, $\eta^2 = .01$ (see Fig. 1).

First and most important, pairwise comparisons revealed the basic effect of watching: High exposure to

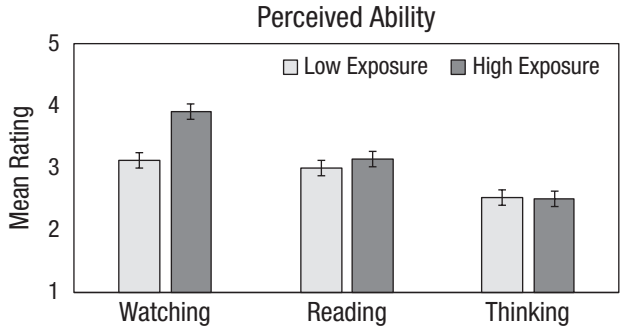

Fig. 1. Perceived ability to successfully perform the tablecloth trick in Experiment 1, separately for each type and amount of exposure. Error bars show ±1 *SE*.

watching someone else perform the tablecloth trick led participants to feel that they themselves would be significantly more likely to successfully perform the trick on their first attempt (M = 3.91, SD = 1.60)—knowing they would have no other practice or training in the interim—compared with watching the video just once (M = 3.12, SD = 1.59), $F(1, 997) = 20.24$, $p < .001$, $\eta_p^2 = .02$ ($d = 0.50$), 95% confidence interval (CI) for the mean difference = [0.44, 1.12]. As hypothesized, merely watching others perform many times increased perceptions of one's own ability to perform the same skill.

Second, this boost did not emerge from overexposure to other kinds of information (reading or thinking): High exposure to text instructions did not significantly increase one's own perceived abilities (M = 3.14, SD = 1.57) compared with low exposure to text instructions (M = 3.01, SD = 1.61), $F(1, 997) = 0.51$, $p > .250$, $\eta_p^2 = .001$ ($d = 0.08$), 95% CI for the mean difference = [−0.46, 0.22], and likewise more time to think and imagine the trick did not significantly increase one's own perceived abilities (M = 2.51, SD = 1.54) compared with low exposure (M = 2.53, SD = 1.58), $F(1, 997) = 0.01$, $p > .250$, $\eta_p^2 < .001$ ($d = 0.01$), 95% CI for the mean difference = [−0.32, 0.35]. Simply having additional time was not enough. Moreover, although access to text instructions boosted perceived abilities compared with just thinking with no other aid—as one might expect—extensive access to reading text instructions did not translate into correspondingly higher perceived abilities.

Manipulation check. The aforementioned results are bolstered by our results for the manipulation check. The manipulation indeed worked as intended for each type of information, as evidenced by a main effect of amount, $F(1, 997) = 251.41$, $p < .001$, $\eta^2 = .20$. There was also an incidental main effect of type, $F(2, 997) = 208.38$, $p < .001$, $\eta^2 = .30$, and an incidental interaction, $F(2, 997) = 36.94$, $p < .001$, $\eta^2 = .07$. Pairwise comparisons revealed that high-exposure participants felt they were more informed than low-exposure participants, whether it was having more time to watch (M = 5.19, SD = 1.78) versus less time to watch (M = 2.90, SD = 1.72), $F(1, 997) = 169.69$, $p < .001$, $\eta_p^2 = .15$ ($d = 1.34$), 95% CI for the mean difference = [1.95, 2.64]; having more time to read (M = 4.96, SD = 1.73) versus less time to read (M = 2.84, SD = 1.65), $F(1, 997) = 148.02$, $p < .001$, $\eta_p^2 = .13$ ($d = 1.25$), 95% CI for the mean difference = [1.79, 2.47]; or having more time to think (M = 1.99, SD = 1.53) versus less time to think (M = 1.61, SD = 1.11), $F(1, 997) = 4.83$, $p = .028$, $\eta_p^2 = .01$ ($d = 0.28$), 95% CI for the mean difference = [0.04, 0.72]. When rerunning the manipulation-check analyses to compare only the watch and read conditions, we found only the key main effect of amount, $F(1, 660) = 274.88$, $p < .001$, $\eta^2 = .29$, with no interaction, $F(1, 660) = 0.38$, $p > .250$, $\eta^2 = .001$, and no main effect of type, $F(1, 660) = 1.18$, $p = .278$, $\eta^2 = .002$. Together, these findings suggest that the basic effect applies most directly to watching, presumably because of the especially vivid, direct, and dynamic information about what to do that watching provides (see also our survey in the introduction).

Experiment 1 provides initial evidence for our hypothesis: The more that people merely watched others, the more they felt like they could perform the skill themselves. These findings also suggest that people do not feel more confident after high exposure to any form of declarative or externally generated information (e.g., Fisher, Goddu, & Keil, 2015); rather, only watching boosted perceived ability, likely because of highlighting the steps especially clearly and fluently.

These results warrant a closer look at the effects of watching others on performance, which we pursued in our remaining experiments. First, we moved beyond self-report and tested whether high exposure indeed fails to boost performance as much as viewers come to believe, across various kinds of skills (Experiments 2–4). If overexposure to watching others does not translate into better actual performance, these perceptions may indeed (at least sometimes) be illusory and therefore problematic. Next, we shed light on why viewers may mistakenly feel like they are learning and tested what they need to help debias their perceptions (Experiments 5 and 6).

To begin, Experiment 2 tested the accuracy of people's perceptions. Given the importance of physical practice for acquiring skills (Ericsson et al., 1993; Wulf et al., 2010), it seems unlikely that merely watching actually enhanced viewers' immediate abilities in Experiment 1, despite their perceptions otherwise. Experiment 2 directly tested this idea by comparing perceived ability to actual ability, in a domain with a clear criterion for success: earning points in darts.

Experiment 2: Throwing Darts

Participants watched a dart-throwing video 1 time or 20 times, and each was assigned to be either a predictor or a performer. Predictors estimated how many points they would earn throwing one dart. Performers actually threw one dart. We hypothesized that repeated watching would enhance predicted, but not actual, scores.

Method

Participants. Participants (N = 202) were recruited from our university subject pool (age: M = 21.69 years, SD = 7.56; 54.92% female; 81.27% Caucasian) to complete the study for $2.00.

Procedure. Participants were led to the study room, where they sat at a computer. Before watching the video, participants viewed a photo of the dartboard to orient them to the task. The dartboard contained seven rings labeled "10," "20," "30," "40," "50," "60," and "80" surrounding a bull's-eye at the center of the dartboard. These numbers corresponded to the point values of the rings, and participants were told that the bull's-eye was worth 100 points.

Next, each participant was randomly assigned to one cell in a 2 (exposure: low, high) × 2 (role: predictor, performer) between-subjects design. Predictors watched a video in which a person throws a dart and hits the bull's-eye in the center of the dartboard (see OSF for the video). One repetition of the video lasted approximately 3 s. The predictors watched the video either 1 time or 20 times in a row. Then they estimated how many points they would earn in a single dart throw: "Suppose we let you throw one dart yourself. How many points do you think you would earn?" (0 = *I'd miss the rings entirely,* 10, 20, 30, 40, 50, 60, 80, 100 = *I'd hit the bull's-eye*[2]). In contrast, performers watched the same video either 1 time or 20 times and then threw one dart themselves while we recorded the number of points that they actually scored. We sought to test whether (a) high-exposure predictors expected to earn more points than low-exposure predictors, replicating Experiment 1, and (b) these higher expectations translated into higher actual dart scores among high-exposure performers compared with low-exposure performers.

The dartboard was hung on a wall with the bull's-eye positioned 68 in. above the floor and the dart thrower positioned 93.25 in. from the base of the wall, as marked with a piece of tape on the floor. These dimensions matched the recommended standards set forth by the Professional Darts Corporation (2018). The dart-throwing video was filmed by the researchers inside the lab room where participants completed the experiment.

Among predictors, we also assessed other variables to replicate the basic effect of low versus high exposure as in Experiment 1. Predictors were asked the following questions: "Suppose we let you throw one dart yourself. How close do you think your dart would land to the bull's-eye?" (1 = *extremely far away/off the board*, 7 = *extremely close/hit the bull's-eye*); "Suppose we let you throw one dart yourself. What are the chances you would hit the bull's-eye?" (0% = *I'd definitely miss the bull's-eye*, 100% = *I'd definitely hit the bull's-eye*; increments of 10%); "To what extent did watching the video help you learn dart-throwing technique?" (1 = *not at all*, 7 = *very much*); and "To what extent did watching the video make you better at throwing darts?" (1 = *not at all*, 7 = *very much*). We expected each of these measures to replicate the basic effect: that high-exposure predictors would report greater abilities than low-exposure predictors.

After, all participants completed an attention check: "Think back to the original dart-throwing video. In the video, where did the person's dart throw land?" (*It missed the circular rings entirely* vs. *It landed in one of the circular rings, but missed the bull's-eye* vs. *It landed in the bull's-eye at the center of the dartboard*). High-exposure participants responded to an additional attention check: "Think back to the original dart-throwing video. Did we show you many different, unique dart throws or did we show you the same dart throw repeatedly?" (*You showed me many different, unique dart throws* vs. *You showed me the same dart throw repeatedly*).

Results

We needed to exclude 9 participants a priori: 5 who did not throw the dart, 1 who withdrew, 1 who was a repeat participant, and 2 because of experimenter error. Among the final N of 193, only 3.63% failed an attention check. We included all of these participants to maximize power.

Overestimating performance. For our primary analysis, we conducted a univariate GLM with exposure, role, and the Exposure × Role interaction as independent variables and dart score (predicted or actual score of the dart throw) as the dependent variable. There was no main effect of exposure $F(1, 189) = 0.27$, $p > .250$, $\eta^2 = .001$, and there was an incidental main effect of role; specifically, predictors generally overestimated their score (M = 38.85, SD = 22.51) relative to performers (M = 23.88, SD = 27.07), $F(1, 189) = 17.54$, $p < .001$, $\eta^2 = .09$ (d = 0.61), 95% CI for the mean difference = [7.87, 21.88].

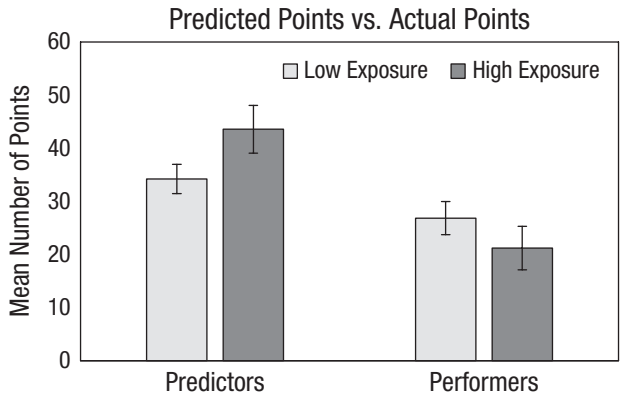

Fig. 2. Mean dart score in Experiment 2, separately for predictors and performers in each exposure condition. Error bars show ±1 *SE*.

More important, we observed the critical interaction, $F(1, 189) = 4.47$, $p = .036$, $\eta^2 = .02$ (see Fig. 2).

Pairwise comparisons revealed that high-exposure predictors expected to score more points ($M = 43.57$, $SD = 23.47$) than low-exposure predictors ($M = 34.21$, $SD = 20.70$), $F(1, 189) = 4.19$, $p = .042$, $\eta_p^2 = .02$ ($d = 0.39$), 95% CI for the mean difference = [0.35, 18.38]. This finding replicates the basic effect from Experiment 1: the more that people merely watch others, the better they think they could perform the skill themselves. But, critically, these boosted expectations did not translate into significant boosts in reality: High-exposure performers did no better ($M = 21.19$, $SD = 26.52$) than low-exposure performers ($M = 26.84$, $SD = 27.72$), $F(1, 189) = 1.08$, $p > .250$, $\eta_p^2 = .01$ ($d = 0.23$), 95% CI for the mean difference = [−16.38, 5.08]. Merely watching others many times did not actually help.

Analyzing the data within exposure conditions is also informative: Whereas low-exposure predictors more accurately imagined low-exposure performance, $F(1, 189) = 2.10$, $p = .149$, $\eta_p^2 = .01$ ($d = 0.30$), 95% CI for the mean difference = [−2.67, 17.40], high-exposure predictors significantly overestimated high-exposure performance, $F(1, 189) = 20.37$, $p < .001$, $\eta_p^2 = .10$ ($d = 0.92$), 95% CI for the mean difference = [12.60, 32.16]. Repeated observation inflated people's perceptions of learning.

Additional variables. Predictors also completed additional measures that served to further replicate the basic effect. Within predictor data, we conducted independent-samples *t* tests with exposure as the independent variable and these additional measures as dependent variables. Consistent with our hypothesis, results showed that high-exposure predictors expected their dart throws to land closer to the bull's-eye ($M = 4.18$, $SD = 1.45$) than did low-exposure predictors ($M = 3.42$, $SD = 1.51$), $t(111) = 2.71$, $p = .008$, $d = 0.51$, 95% CI for the mean difference = [0.20, 1.31]; predicted that they were more likely to hit the bull's-eye ($M = 31.61$, $SD = 24.99$) than did low-exposure predictors ($M = 20.35$, $SD = 18.12$), $t(111) = 2.74$, $p = .007$, $d = 0.52$, 95% CI for the mean difference = [3.13, 19.38]; reported learning more technique by watching ($M = 2.70$, $SD = 1.44$) than did low-exposure predictors ($M = 2.09$, $SD = 1.17$), $t(111) = 2.47$, $p = .015$, $d = 0.46$, 95% CI for the mean difference = [0.12, 1.10]; and reported improving more by watching ($M = 2.21$, $SD = 1.50$) than did low-exposure predictors ($M = 1.60$, $SD = 1.13$), $t(111) = 2.48$, $p = .015$, $d = 0.47$, 95% CI for the mean difference = [0.12, 1.11]. Our actual dart-score data suggest that these additional perceptions of learning do not necessarily reflect reality.

Experiment 2 provided further support for the hypothesis. Actual performance (the score of a dart throw) was not immediately boosted after watching others perform the skill many times (throwing a bull's-eye), but mere observers believed that it would be. Next, we sought to replicate this effect in a different performance domain—dancing—and using a within-subjects design: The same participants who made predictions then attempted the performance. This afforded a more conservative test (predictors might temper their confidence if they know they have to make the attempt) and further boosted generalizability (in everyday life, performers might consider how well they will do before actually performing; perhaps the act of setting a high prediction indeed improves performance and therefore erases the effect).

Experiment 3: Doing the Moonwalk

Participants watched a moonwalk video 1 time or 20 times. Participants predicted how well they could do the moonwalk, then actually attempted it. We hypothesized that repeated watching would enhance predicted, but not actual, moonwalking performances.

Method

Participants. First, participants ($N = 100$) were recruited from our university subject pool (age: $M = 26.26$ years, $SD = 11.29$; 54.00% female; 38.00% Caucasian) to complete the moonwalk phase for $1.00. They predicted how well their attempt at a moonwalk would be judged by a group of outside raters, and then they made their attempt in front of a video camera. Next, participants ($N = 100$) were recruited from Amazon's Mechanical Turk (age: $M = 33.06$ years, $SD = 8.98$; 30.00% female; 81.00% Caucasian) to complete the ratings phase for $5.00. They watched the performance videos and judged each one on the same rating scale that performers had used for their predictions.

We chose moonwalking as the performance domain because we assumed that many participants by default might feel unskilled or embarrassed by the thought of their attempt and even more so knowing their performance would be videotaped and judged. These forces might compel participants against inflating their predicted abilities, providing a more conservative test.

Moonwalk procedure. Participants were led to a private study room where they sat at a computer. They were informed that they would watch a training video of a moonwalk dance move. They would then get one shot at attempting this same move in the video without any additional practice or training, and we would video-record this attempt. Their video-recorded moonwalks would be shown to a separate group of raters in the second phase of the study. Participants were told that the raters would first watch the same training video and then rate each participant's attempt on a scale from 1 (*pretty bad attempt*) to 10 (*pretty good attempt*). Participants then watched the training video, in which a person performs the moonwalk (see OSF for the video). One repetition of the video lasted approximately 6 s. Following random assignment to condition, low-exposure participants watched the video 1 time, and high-exposure participants watched the video 20 times consecutively.

After watching but before actually performing, all participants were reminded that their attempt would be judged by a group of outside raters and were asked to predict "how an average rater would rate YOUR attempt." They made predictions on a sliding scale from 1 (*pretty bad attempt*) to 10 (*pretty good attempt*). The score showed on the side as participants slid along the scale, displaying to the hundredth decimal place. After making their prediction, all participants then actually attempted a single moonwalk in the lab room by moonwalking from one piece of tape to another marked on the floor. A stationary video camera recorded the attempt. Both the model's video and participants' performance videos showed the performer from the neck down.

After attempting their moonwalk, all participants completed two forced-choice questions: (a) "Now that you've actually made your attempt, how was it for you?" (*It turned out to be easier than I expected, as compared to how I felt right after my video training* vs. *It turned out to be as easy/hard as I expected, as compared to how I felt right after my video training* vs. *It turned out to be harder than I expected, as compared to how I felt right after my video training*) and (b) "Now that you've actually made your attempt, how well do you think you did?" (*I ended up doing better than I predicted, as compared to how I felt right after my video training* vs. *I think I ended up doing as good/bad as I predicted, as compared to how I felt right after my video training* vs. *I ended up doing worse than I predicted, as compared to how I felt right after my video training*). We did not make a priori predictions about these items (see the preregistered materials on OSF). However, if the basic effect were to be replicated, we were interested in getting a sense of whether participants realize that their predictions were indeed inflated after they actually experienced the move (we returned to this idea in Experiment 6).

Finally, all participants completed an attention check: "Earlier you watched a video in which another person performed the moonwalk. How many times did we show you this video? (*You showed me this video 1 time* vs. *You showed me this video 20 times in a row*)."

Ratings procedure. In the next phase of the study, we showed the moonwalk videos to a sample of 100 outside raters to test the accuracy of performers' predictions. First, all raters were told about the lab procedure and watched the original training video once. Raters knew that the lab participants had seen the same video prior to their attempts. Then, raters watched and rated each of the 100 moonwalks, one at a time in randomized order, from 1 (*pretty bad attempt*) to 10 (*pretty good attempt*). Each rating screen was prefaced with the phrase "Compared to the original training video" and had a link to rewatch the training video if desired. Thus, each rater evaluated all 100 videos (i.e., each moonwalk was evaluated by 100 different raters). As preregistered, we calculated the mean rating for each video and treated this mean as a single actual performance score for each performer, which could be directly compared with each performer's predicted score.

Results

Only 1.00% of lab participants failed an attention check, and 1.00% of raters reported technical difficulties. We include all moonwalkers and all raters in the following analyses.

Overestimating performance. For our primary analysis, we conducted a repeated measures GLM with exposure (low, high) as a between-subjects factor and role (predictor, performer) as a within-subjects factor, with the moonwalk scores as the dependent measure. There was a main effect of exposure, $F(1, 98) = 5.26$, $p = .024$, $\eta^2 = .051$, and there was no main effect of role, $F(1, 98) = 2.69$, $p = .104$, $\eta^2 = .027$. More important, we observed the critical interaction, $F(1, 98) = 10.93$, $p = .001$, $\eta^2 = .100$ (see Fig. 3).

Pairwise comparisons revealed that high-exposure participants expected to perform better moonwalks

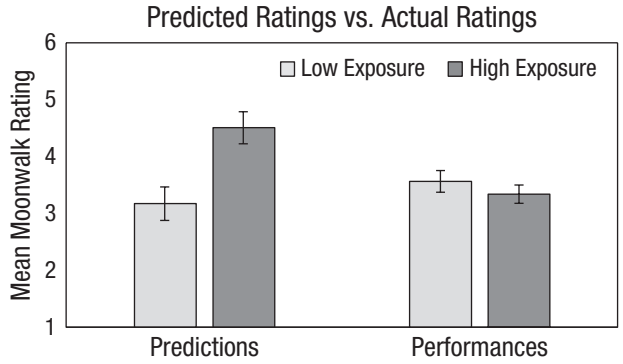

Fig. 3. Mean predicted and actual moonwalk ratings in Experiment 3, separately for each exposure condition. Error bars show ±1 *SE*.

(M = 4.51, SD = 1.99) than low-exposure participants (M = 3.17, SD = 2.09), $F(1, 98) = 10.73$, $p = .001$, $\eta_p^2 = .10$ ($d = 0.66$), 95% CI for the mean difference = [0.53, 2.14]. This replicates the basic effect from Experiments 1 and 2: The more that people merely watch others, the better they think they could perform the skill themselves. But critically, these boosted expectations did not translate into significant boosts in reality: High-exposure participants moonwalked no better (M = 3.34, SD = 1.14) than low-exposure participants (M = 3.57, SD = 1.35), $F(1, 98) = 0.81$, $p > .250$, $\eta_p^2 = .01$ ($d = 0.18$), 95% CI for the mean difference = [−0.27, 0.72]. As in Experiment 2, merely watching others many times did not actually enhance participants' immediate abilities, despite their predictions otherwise.

Analyzing the data within exposure conditions was also informative: Whereas low-exposure participants accurately imagined the quality of their low-exposure moonwalks, $F(1, 98) = 1.39$, $p = .242$, $\eta_p^2 = .01$ ($d = 0.16$), 95% CI for the mean difference = [−1.06, 0.27], high-exposure participants significantly overestimated the quality of their high-exposure moonwalks, $F(1, 98) = 12.22$, $p = .001$, $\eta_p^2 = .11$ ($d = 0.52$), 95% CI for the mean difference = [0.51, 1.83]. Repeated observation inflated people's perceived ability.

Additional variables. Lab participants also completed two exploratory measures so we could gauge their reactions after performing. For how good they thought their attempt turned out, most high-exposure participants felt their attempt was worse than expected (58.00% worse, 2.00% better, 40.00% as expected), whereas most low-exposure participants felt their attempt was as expected (32.00% worse, 4.00% better, 64.00% as expected). A logistic regression testing for differences in these choices (dummy codes: 1 = worse than expected, 2 = as expected) confirmed a significant effect of exposure, $b = 1.07$, $SE = 0.42$, Wald = 6.36, $df = 1$, $p = .012$, Exp(b) = 2.90. These results mirrored the basic effect: High-exposure participants were indeed overconfident in their moonwalking abilities, which they realized firsthand on actually attempting the move.

We observed similar patterns for the item regarding how difficult participants ended up finding the task: Fewer low-exposure participants found the task harder than expected (32.00% harder, 8.00% easier, 60.00% as expected) compared with high-exposure participants (42.00% harder, 14.00% easier, 44.00% as expected), although the logistic regression results were not statistically significant, $b = 0.58$, $SE = 0.44$, Wald = 1.79, $df = 1$, $p = .180$, Exp(b) = 1.79.

These results extend the basic effect to a within-subjects design. Participants thought their dancing abilities had improved after repeatedly watching someone else perform the dance. In reality, this boosted confidence was mistaken—merely watching did not actually help. So far, we found that viewers' actual abilities did not improve after merely watching others throw a dart (Experiment 2) and perform a dance (Experiment 3), despite their predictions otherwise. As an additional test to establish this basic effect, in Experiment 4 we used a within-subjects design with the same exact participants providing predicted scores and actual scores. Moreover, we sought to test an easier-to-scale performance domain: abilities to play a computer game.

Experiment 4: Playing a Game

Participants played a "mirror-tracing" game. They first watched a video of someone playing, predicted their own score, and then played the game themselves. We hypothesized that watching many times would enhance predicted, but not actual, scores.

Method

Participants. Participants (N = 270) were recruited from Amazon's Mechanical Turk (age: M = 32.61 years, SD = 9.28; 54.44% female; 65.93% Caucasian) to complete the study for $1.00.

Procedure. Participants assessed their abilities to play a mirror-tracing game and then actually played the game themselves. The game was modeled from a game used by Cusack, Vezenkova, Gottschalk, and Calin-Jageman (2015), who developed the game as a behavioral methods tool to study complex motor movements through online platforms. We hired a programmer to build a version of their game that we could implement within our Qualtrics survey software and use on Mechanical Turk (see OSF for the game).

In the game, players see an image of a curved maze at the top of the screen. In an empty box below, players must recreate the shape of this maze by tracing it with the computer cursor. The only points marked in this

tracing box are a dot for where to start and a dot for where to end. Players therefore must simulate the path in between as closely and as quickly as possible. As players move, they see an automated running tally of their score, which ranges from 0% to 100% corresponding to the percentage match to the correct path (i.e., a score of 100% means the player is simulating the shape of the maze perfectly, whereas a score of 0% means the player is completely deviating). Finally, adding further complexity to the experience of playing the game, players cannot use a mouse but instead have to trace the shape by carefully moving their fingers along their computer's track pad, and furthermore, their movements throughout the task are traced in reverse (e.g., when the maze goes up and players need to trace upward, they need to move their fingers down on the track pad).

For our experiment, participants were told that they would get one shot at playing the game without any practice or training beyond our instruction screens. All participants began by clicking through detailed step-by-step instruction screens explaining what the game is and how it works (including the full scoring procedure and all controls), culminating in watching a video of someone playing the game (which we recorded). The player in the video does well, earning a score of 94%. The video shows a split-screen performance of the player's hand movements on the track pad as well as what is happening in real time on the screen (see OSF for the video). One repetition of the video lasted about 8 s. Following random assignment to condition, low-exposure participants watched the video 1 time, and high-exposure participants watched the video 20 times consecutively, as in our other experiments. All participants were instructed to be passive viewers and watch the video without doing anything else, including not practicing or mimicking the person in the video.

After watching but before actually playing, all participants were reminded that their task was to trace the maze as quickly as they can while earning the highest percentage score that they could. They predicted their score on a sliding scale from 0% to 100%. After making their prediction, all participants then actually played the game, and we recorded their score (also between 0% and 100%). There was no opportunity to lie about one's performance: The game was programmed to automatically copy scores into the data file after participants completed the maze, and participants could play the game and attain a score just one time.

After playing, all participants completed an attention check: "Before you actually played the game yourself, we showed you a video of a person playing the game. What did you see in the video?" (*I saw the person play the level once and that was that* vs. *I saw the person play the level once but the video replayed 20 times in a row*). We also included three exploratory checks about the study experience overall: (a) "Which of the following best describes what you were doing while watching the video?" (*I was basically just watching like a passive viewer, without practicing the hand motions myself* vs. *I was more like an active viewer, practicing the hand motions myself while watching*); (b) "To what extent were you yourself practicing the hand motions before you actually played the game?" (*not at all, a little bit, moderate, quite a bit, a lot*); and (c) "While playing the game, did you end up going as fast as you can?" (*Yes, I went as fast as I could while trying to get a good score* vs. *No, I ended up slowing down/stopping/etc. in order to get a higher score*). Last and in a similar vein, the game tacitly recorded how long it took participants to finish the maze. We presumed an ideal test of our key hypothesis would find no differences in these items across watching conditions.

Results

Only 0.74% of participants failed the attention check. We included all participants in the following analyses.

Overestimating performance. For our primary analysis, we conducted a repeated measures GLM with exposure (low, high) as a between-subjects factor and score (predicted, actual) as a within-subjects factor. There was a main effect of exposure, $F(1, 268) = 9.20$, $p = .003$, $\eta^2 = .03$, as well as an incidental main effect of score: Participants generally overestimated how well they would do ($M = 62.41$, $SD = 20.63$) relative to how they ended up doing ($M = 48.65$, $SD = 25.13$), $F(1, 268) = 52.72$, $p < .001$, $\eta^2 = .16$ ($d = 0.45$), 95% CI for the mean difference = [9.81, 17.10]. More important, we observed the critical interaction, $F(1, 268) = 7.76$, $p = .006$, $\eta^2 = .03$ (see Fig. 4).

Pairwise comparisons revealed that high exposure to watching someone else play the game led participants to predict that they would earn a significantly higher score ($M = 67.76$, $SD = 17.67$) compared with getting low-exposure to the video ($M = 56.38$, $SD = 22.07$), $F(1, 268) = 22.10$, $p < .001$, $\eta_p^2 = .08$ ($d = 0.57$), 95% CI for the mean difference = [6.62, 16.15]. This replicates the basic effect from all previous experiments: The more that people merely watch others, the better they think they could perform the skill themselves. But critically—replicating the performances in Experiments 2 and 3—these boosted expectations did not translate into significant boosts in reality: High-exposure performers went on to score no higher ($M = 49.15$, $SD = 24.67$) than low-exposure performers ($M = 48.09$, $SD = 25.73$), $F(1, 268) = 0.12$, $p > .250$, $\eta_p^2 < .001$ ($d = 0.04$), 95% CI for the mean difference = [−7.10,

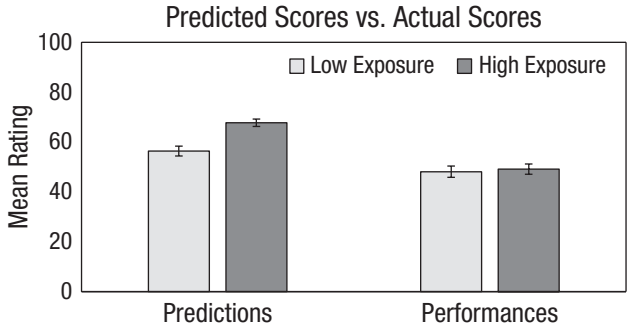

Fig. 4. Mean predicted and actual mirror-tracing scores in Experiment 4, separately for each exposure condition. Error bars show ±1 *SE*.

4.98]. Merely watching others many times did not actually help.

Analyzing the data within exposure conditions was also informative. As reported earlier, all participants were generally overconfident. However, high-exposure participants were far more overconfident, $F(1, 268) = 53.65$, $p < .001$, $\eta_p^2 = .17$ ($d = 0.66$), 95% CI for the mean difference = [13.61, 23.62]—their predictions were higher over the mark—compared with low-exposure participants, $F(1, 268) = 9.45$, $p = .002$, $\eta_p^2 = .03$ ($d = 0.25$), 95% CI for the mean difference = [2.98, 13.60]. Perceptions of learning were especially inflated following repeated observation.

Additional variables. Finally, the results of our exploratory checks further isolated the effect of watching and suggested that participants had otherwise similar study experiences (see the Supplemental Material for complete results): An equal majority of participants in both conditions reported following instructions to watch the video passively (low exposure: 90.55%; high exposure: 88.11%), not practicing before playing (low exposure: $M = 1.57$, $SD = 0.99$; high exposure: $M = 1.45$, $SD = 0.85$), and tracing the maze quickly (low exposure: 86.61%; high exposure: 86.71%); all ps > .250. Likewise, low-exposure participants took just as long to finish the maze ($M = 42.48$ s, $SD = 35.80$ s) as high-exposure participants ($M = 35.58$ s, $SD = 35.42$ s), $t(268) = 1.59$, $p = .113$, $d = 0.19$, 95% CI for the mean difference = [−1.65, 14.46].

Together with Experiments 2 and 3, these findings robustly highlight the same basic effect: Across a variety of skills, watching others perform many times leads people to overestimate how much their own abilities have improved. In our next set of experiments, we moved toward better understanding mechanisms: We more specifically tested what viewers attend to (Experiment 5) and what they fail to take into account (Experiment 6) that may be inflating perceptions of learning.

In Experiment 5, we sought to better discern why viewers believe they have improved after merely watching. What are high-exposure viewers actually reacting to? We have proposed that viewers are exposed to direct, vivid information about what the performer is actually doing, leading them to feel like they have learned enough (without having incorporated how those steps feel). Our survey in the introduction as well as Experiment 1 support this possibility. Note, however, that repeated watching also overexposes viewers to success, and reflecting on success could enhance viewers' confidence, whether or not they also attend to steps of the performance (Hall, Ariss, & Todorov, 2007; Ruvolo & Markus, 1992). Still another possibility is that simply having ample time to think or mentally prepare drives the effect (although Experiment 1 suggests otherwise).

Experiment 5 tested for more direct evidence that viewers were indeed being influenced by specifically tracking the performer's actions over and above these other possibilities. High-exposure viewers should not show the boost when it is difficult to track the performer's actions, despite seeing the same successful outcome so many times. Moreover, this design holds possible demand constant by comparing conditions of equally high exposure.

Experiment 5: Visual Insight

Participants watched the tablecloth video from Experiment 1. We manipulated whether participants could see both the tablecloth and performer or only the tablecloth. We hypothesized that seeing what to do many times (and not high exposure per se) may elicit the effect.

Method

Participants. Participants ($N = 400$) were recruited from Amazon's Mechanical Turk (age: $M = 33.57$ years, $SD = 9.69$; 40.80% female; 78.50% Caucasian) to complete the study for $0.25.

Procedure. Participants were assigned to one cell in a 2 (performer: present, absent) × 2 (exposure: low, high) between-subjects design. Participants in the performer-present condition watched the full video depicting the person performing a tablecloth trick. Participants in the performer-absent condition saw the same exact video, except it was cropped such that viewers could see the table set with dishes but could not see the performer's specific hand placements and movements (see OSF for

the videos). Otherwise the video was identical. Note that these participants nonetheless saw the same successful outcome (and everything else in the video) and watched just as many times as the other participants. Any differences between high-exposure conditions therefore cannot be attributed to these more general exposure effects.

After, all participants responded to three dependent variables, presented in randomized order: "To what extent did watching the video make you better at doing this?" (1 = *not at all*, 7 = *very much*), "To what extent did watching the video prepare you to do this yourself?" (1 = *not at all*, 7 = *very much*), and "How much technique did you learn from watching the video?" (1 = *none at all*, 7 = *quite a bit*). These questions were designed to capture a more general assessment of perceived learning from watching beyond the single-score estimates in our other experiments.

Finally, participants reported whether they had ever tried a tablecloth trick (yes/no) and responded to three attention checks: "How many times did we show you the same video?" (*1, 2, 5, 10, 15, 20*), "What did you see in the video?" (*A person dunked a basketball* vs. *A person threw a bowling ball* vs. *A person threw a dart* vs. *A person pulled a tablecloth* vs. *A person played with a yo-yo*), and "Did you watch the entire video? (no penalty for honesty!)" (*yes/no*).

Results

Only 4.50% of participants failed any attention check, and 16.00% of participants reported that they had previously attempted a tablecloth pull. We include all participants to maximize power. The dependent measures were collapsed to form a perceived-skill-acquisition scale ($\alpha = .90$), although the effects hold for each item individually as well (see the Supplemental Material). We conducted a univariate GLM with performer, exposure, and the Performer × Exposure interaction as independent variables and the perceived-skill-acquisition scale as the dependent variable. As hypothesized, there was a main effect of performer, $F(1, 396) = 19.69$, $p < .001$, $\eta^2 = .05$; a main effect of exposure, $F(1, 396) = 14.47$, $p < .001$, $\eta^2 = .04$; and the critical interaction, $F(1, 396) = 4.23$, $p = .040$, $\eta^2 = .01$ (see Fig. 5).

Marking the source of this interaction, pairwise comparisons revealed a replication of the basic effect among participants who could see the actual performer and his actions: High exposure to this video again led viewers to report significantly higher skill acquisition ($M = 2.95$, $SD = 1.55$) compared with low exposure ($M = 2.15$, $SD = 1.22$), $F(1, 396) = 16.76$, $p < .001$, $\eta_p^2 = .04$ ($d = 0.59$), 95% CI for the mean difference = [0.42, 1.18]. Merely watching many times inflated perceived learning.

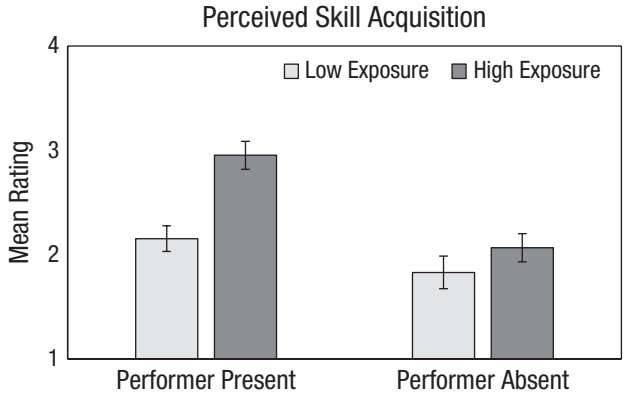

Fig. 5. Mean rating of perceived skill acquisition in Experiment 5, separately for each video type and exposure condition. Error bars show ±1 *SE*.

In contrast, the basic effect was attenuated among participants who could not see the performer's specific actions: Viewers did not feel like they had learned any more after high exposure ($M = 2.07$, $SD = 1.34$) than after low exposure ($M = 1.83$, $SD = 1.33$), $F(1, 396) = 1.56$, $p = .212$, $\eta_p^2 = .004$ ($d = 0.17$), 95% CI for the mean difference = [–0.14, 0.61]. Despite watching others many times, these participants did not come away feeling like they were better off themselves.

These results provided moderation-based evidence for our framework, helping rule out pure effects of high exposure (general fluency, extra time to think and reflect, effort justification, observing success, etc.) and highlighting what viewers might actually be noticing that leads them to exhibit the effect. Watching others many times does not inflate perceptions of skill acquisition if viewers cannot specifically see the performer's actions—that is, people feel that they are learning while merely watching only if they can track what the specific steps and actions look like (despite never experiencing what the performance feels like, which may prove critical).

Experiment 6: Getting Back in Touch

Finally, we tested three strategies for calibrating self-assessments. Participants watched a performance, then (a) reflected on the task, (b) read technical details about the task, or (c) personally interacted with the objects involved. If the illusion is driven by viewers neglecting the feeling of doing, then giving them a taste of doing should most attenuate it.

Method

Participants. Participants ($N = 150$) were recruited from the Museum of Science and Industry in Chicago,

Illinois (age: $M = 32.42$ years, $SD = 13.30$; 47.33% female; 71.33% Caucasian), to complete the study in exchange for a gift pen.

Procedure. Participants entered the study room and sat at a computer. They were told that they would watch a video in which a person juggles three bowling pins. Then they were shown one actual bowling pin and were told that they may be asked to juggle bowling pins later. Participants then watched the video 20 times in a row (see OSF for the video). Each repetition of the video lasted approximately 5 s. After watching, participants completed the same dependent measures from Experiment 5, plus an additional item explicitly about ability: "How well could you perform this yourself if you actually tried?" (1 = *extremely poorly*, 7 = *extremely well*).

Participants were then assigned to one of the three debiasing conditions, each of which was designed to provide additional information that might help inform people's judgments about how much they had learned while watching. The first two conditions below provide control comparisons: We gave participants different kinds of additional information about the juggling video, but this information did not provide direct access into the feeling of the task in action and therefore did not bridge the experiential gap between seeing and doing per se.

First, participants in the explanation condition were asked to spend additional time reflecting on the task. They responded to the following item: "Now please write a detailed, step-by-step explanation of how the person juggled the bowling pins. Please write out the sequence you saw in as much detail as possible." Other research has found that asking people to explain how something works often reminds them they do not know it as well as they thought at first glance (the "illusion of explanatory depth"; Rozenblit & Keil, 2002). We tested whether such a task could temper perceived skill acquisition. Participants were given 1 min to reflect and write (we will return to the illusion of explanatory depth in the General Discussion).

Second, other participants in the technical-information condition were given the following true information about each of the bowling pins shown in the video: "Weight = 3.5 pounds (1.6 kg); Length = 15 inches (38 cm); Minimum diameter = 1.8 inches (4.6 cm); Maximum diameter = 4.8 inches (12.2 cm); Surface material = plastic." Reading these details may help people more accurately imagine what the experience is like (although the read conditions in Experiment 1 provided additional evidence against this possibility). Participants were given 1 min to read and reflect on the information.

Of critical interest, still other participants were indeed given direct access to the feeling of the performance: Participants in the sensory-experience condition were asked to hold the bowling pins for 1 min. Equally critical, participants were instructed to hold the pins but not to juggle them: This provided a small taste of doing without prompting them to try the task and fail (and so unsurprisingly conclude that they had not learned in Phase 1). In other words, these participants simply received additional information about the task and did not get any actual feedback about their abilities (similar to participants in the other two conditions). The pins were identical to the ones seen in the video and that had been described to participants in the technical-information condition.

After the debiasing period, all participants then completed slightly modified perceived-skill-acquisition items, which piped in their earlier responses in place of the letter "X": "You originally said, in Phase 1, that the video made you X/7 better at doing this. Now, as you think back on the video, to what extent did watching the video in Phase 1 make you better at doing this?" and likewise for the other items. Changes in ratings on the perceived-skill-acquisition scale and the perceived-ability item from Time 1 (having watched many times) to Time 2 (having then received a form of additional information about the task) were our dependent variables. Again, any possible demand in this task or in these items was held constant; pure demand predicts significant drops in perceived learning for all conditions, whereas our framework predicts a significant drop only for one: the key sensory condition.

Finally, all participants answered an attention check: "Did we show you the same video footage one time or many times repeatedly?" (*one time* vs. *many times repeatedly*). They also indicated whether they had ever tried juggling bowling pins prior to the experiment (*yes/no*).

Results

We had to exclude 5 participants a priori: 4 because of experimenter error and 1 because the participant withdrew prior to finishing all procedures. Among the final $N = 145$, 1 failed the attention check, and 10 reported previous experience juggling bowling pins. We included all these participants to maximize power.

Perceived skill acquisition. The perceived-skill-acquisition measures were highly correlated in both Phase 1 ($\alpha = .85$) and Phase 2 ($\alpha = .85$), so we collapsed them into scales, although the effects held for each item individually as well (see the Supplemental Material). We conducted a repeated measures analysis of variance with condition as the between-subjects factor (three levels: one of three kinds of debiasing task) and time (two

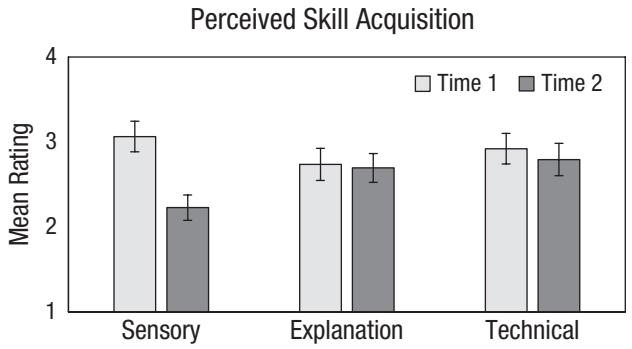

Fig. 6. Mean reduction in perceived learning from merely watching in Experiment 6, separately for participants in each of the three conditions at Time 1 and Time 2. Error bars show ±1 *SE*.

levels: perceived learning at Time 1, before the debiasing task, and perceived learning at Time 2, after the debiasing task) as the within-subjects factor.

There was no main effect of condition, $F(2, 142) = 0.41$, $p > .250$, $\eta^2 = .10$, but there was a main effect of time: Participants generally adjusted their perceptions of learning following their debiasing task, $F(1, 142) = 30.63$, $p < .001$, $\eta^2 = .18$. Critically, however, this depended on the type of additional information that participants were given, as demonstrated by a significant interaction, $F(2, 142) = 17.07$, $p < .001$, $\eta^2 = .19$ (see Fig. 6).

Pairwise comparisons revealed that participants who received a small taste of doing by simply holding the bowling pins themselves then reported that they had learned significantly less than what they had initially thought after merely watching (Time1: $M = 3.06$, $SD = 1.23$; Time 2: $M = 2.23$, $SD = 1.01$), $F(1, 142) = 61.98$, $p < .001$, $\eta_p^2 = .30$ ($d = 0.89$), 95% CI for the mean difference = [0.63, 1.05]. However, no such adjustments were made following the other, nonphysical, debiasing tasks: Perceived learning remained just as high after participants reflected on how the person was able to juggle and wrote an explanation of it (Time 1: $M = 2.74$, $SD = 1.31$; Time 2: $M = 2.69$, $SD = 1.18$), $F(1, 142) = 0.16$, $p > .250$, $\eta_p^2 = .001$ ($d = 0.06$), 95% CI for the mean difference = [–0.17, 0.25]. Likewise, perceived learning remained just as a high after participants read detailed technical information about the size and weight of the pins (Time 1: $M = 2.92$, $SD = 1.28$; Time 2: $M = 2.79$, $SD = 1.32$), $F(1, 142) = 1.51$, $p = .221$, $\eta_p^2 = .01$ ($d = 0.27$), 95% CI for the mean difference = [–0.08, 0.33].

Perceived ability. These results were observed for the perceived-ability item, replicating our preceding experiments. There was no main effect of condition, $F(2, 142) = 0.27$, $p > .250$, $\eta^2 = .004$; the same main effect of time, $F(1, 142) = 6.14$, $p = .014$, $\eta^2 = .04$; and the same critical interaction, $F(2, 142) = 12.62$, $p < .001$, $\eta^2 = .15$ (see Fig. 6). Marking the source of the interaction, pairwise comparisons revealed that participants who received a small taste of doing indeed lowered their perceived ability from what they had initially thought after merely watching (Time 1: $M = 1.85$, $SD = 0.98$; Time 2: $M = 1.40$, $SD = 0.74$), $F(1, 142) = 30.04$, $p < .001$, $\eta_p^2 = .18$ ($d = 0.60$), 95% CI for the mean difference = [0.29, 0.61]. But again, no such adjustments were made after writing and reflecting on an explanation of the task (Time 1: $M = 1.67$, $SD = 1.08$; Time 2: $M = 1.71$, $SD = 1.01$), $F(1, 142) = 0.27$, $p > .250$, $\eta_p^2 = .002$ ($d = 0.08$), 95% CI for the mean difference = [–0.12, 0.20], or after reading additional technical information about the task (Time 1: $M = 1.74$, $SD = 0.99$; Time 2: $M = 1.80$, $SD = 1.12$), $F(1, 142) = 0.58$, $p > .250$, $\eta_p^2 = .004$ ($d = 0.16$), 95% CI for the mean difference = [–0.10, 0.22].

Finally, the perceived-skill-acquisition scale and the perceived-ability item were highly correlated across conditions, both before ($r = .62$) and after ($r = .69$) the interventions. As might be expected, perceptions of learning were tightly linked to actual ability beliefs, and both of these evaluations may have become elevated merely from watching others (even in the absence of any actual doing).

Experiment 6 provided converging support for our framework. Our previous study revealed that viewers track the specific steps of others' performances while watching, leading them to feel like they could perform the skill themselves. Conversely, the current results suggest that viewers indeed take this information at face value and do not fully appreciate how those actions actually feel when doing them. That participants backtracked in their perceptions of learning after gaining direct information about the feeling of doing—but not after gaining additional details or trying to explain the performer's technique themselves—suggests that viewers do not incorporate this critical piece into their initial assessments.

General Discussion

Modern media afford unprecedented opportunities to watch and learn from others. Six experiments suggest that merely watching may have unforeseen costs for self-assessment. The more people watch others perform (without corresponding practice), the more they think they can perform the skill, too (Experiment 1). However, repeated watching does not necessarily improve immediate abilities, despite predictions otherwise (Experiments 2–4). These effects may reflect learning how performances look through repeated exposure (Experiment 5), without incorporating how those performances feel

within the moment of doing (Experiment 6). The experiential gap between seeing and doing may sometimes lead people to assume that they have learned more from merely watching than they have, fostering an illusion of skill acquisition.

Psychologists have long been interested in the link between observation and actual learning (Bandura, 1986; Sheffield, 1961). Our novel contribution highlights the role of prediction: Regardless of whether observation promotes actual skill acquisition, viewers may think they have learned more than warranted. While observation is commonly praised as beneficial for learning—and certainly better than doing nothing (Newell, 1991; Wulf et al., 2010)—our findings suggest that these benefits must be weighed against the possible costs of overestimating one's abilities (especially on the first try). Consider the X Games, an Olympics-style event featuring extreme sports attracting 30 million viewers annually (Statista, 2017b). Avid viewers may feel prematurely inspired to attempt similar actions themselves, with tragic consequences. In daily life, too, people may develop inflated confidence after watching others perform tasks from cooking to home repair (e.g., after a quick search for YouTube tips), causing people to rely too readily on themselves and forego better results from outsourcing to experts.

This insight echoes and extends classic research on overconfidence. People generally think they know more than they do and do not consider their ignorance until pressed (Dunning, 2005; Fisher et al., 2015; Marteau, Wynne, Kaye, & Evans, 1990; O'Brien, 2013; Rozenblit & Keil, 2002). Our findings suggest that one must press wisely: Showing a video over and over (vs. extensive reading or reflection) may increase perceived knowledge rather than emphasize a task's many complexities. Even when people initially recognize a task as difficult (Kruger, 1999), they may quickly turn overconfident after mere observation, swayed by their additional (but insufficient) preparation.

Our findings raise important directions for research. First, longer-term dynamics should be explored. Observation is necessary for understanding, so repeated watching may help in the long run; perhaps high-exposure viewers ultimately learn quicker despite overestimating their immediate abilities. Alternatively, because watching may not draw attention to critical features of the performance, high-exposure viewers could misunderstand the kind and amount of practice needed during subsequent training and therefore be no better prepared.

Second, interpersonal challenges may arise between parties with different experiential knowledge. For example, when swimming instructors model a backstroke, novices are unlikely to notice the head position, hip rotation, and kicking maneuver simultaneously while watching. Like a curse of knowledge (Camerer, Loewenstein, & Weber, 1989), instructors feel these techniques while demonstrating and may neglect novices' insensitivity to this subtle information. Instructors may overestimate the pedagogical value of behavioral modeling, causing frustration and reducing the time learners spend doing.

Third, identifying additional moderators and mediators would improve generalizability beyond our documented effects of specific videos, on specific performances, among specific populations. At the level of prediction, why does extensive watching (e.g., vs. reading) so influence perceived learning? Experiment 5 suggested that viewers lock onto the steps of the performance, which likely manifest most clearly and fluently via watching. Perhaps extremely vivid text-based tutorials operate similarly. Likewise, perhaps merely reading about feelings is sufficient to reduce the illusion; our experiments do not disentangle whether predictors fail to realize that such feelings are present from whether predictors are aware but misperceive their impact. Highlighting task complexity in still other ways (e.g., watching unskilled others or watching others work through a learning curve) may also inform predictions. More research like Experiments 5 and 6 is needed to discern what, exactly, viewers notice or infer versus miss or discount.

Relatedly, at the level of performance, why did high exposure not improve immediate abilities given that observation is known to elicit automatic simulations of real-time feelings of the experience (e.g., research on implicit procedural learning and mirror-neuron mimicry; Lyons, Young, & Keil, 2007; Mattar & Gribble, 2005; Stefan et al., 2005)? The activation of this system depends on having past personal experience with the observed action (Heyes, 2001) and is stronger when observing simple tasks (Heyes & Foster, 2002). We assessed novel, complex tasks. Perhaps this system was not so engaged, explaining why extensive watching did not help. Or perhaps this system was engaged but was fed incomplete information; if viewers do not even look at a moonwalker's hips, their simulations may not incorporate hips. Another possibility is the dynamic nature of repetition. Extensive actual consumption creates desensitization, at which point people struggle to recall the intensity of initial reactions (Campbell et al., 2014). Perhaps extensive simulation works similarly, undermining abilities to then resimulate the first live step.

Finally, Experiment 6 suggested that perceived learning is reduced by a taste of doing but not other potentially useful information. In daily life, this taste frequently comes too late (e.g., after an audience has gathered or one has precommitted to a task). Future

studies should test the effectiveness of other proxies for doing for calibrating self-assessments. Fruitful candidates include watching first-person performance videos, miming the performer's actions or handling related objects while watching, and playing virtual-reality games.

Until these possibilities are tested, the current experiments suggest that today's ubiquity of opportunities to watch and learn from others—via YouTube or elsewhere—warrant a closer look. While people may feel they are acquiring the skills that athletes, artists, and technicians perform in front of their eyes, often these skills may be easier seen than done.

Action Editor

Leaf Van Boven served as action editor for this article.

Author Contributions

M. Kardas developed the study concept. Both authors contributed to the experimental designs. M. Kardas performed the data collection, analysis, and interpretation under the supervision of E. O'Brien. M. Kardas wrote the first draft of the manuscript, and E. O'Brien provided critical revisions. Both authors approved the final version for submission.

Acknowledgments

We thank Nick Epley, Jane Risen, and Eugene Caruso (and their lab groups), as well as Linda Hagen and Anuj Shah, for especially helpful feedback. Parts of this research were presented at the annual conferences for the Society of Judgment and Decision Making and the Society for Personality and Social Psychology.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by the Willard Graham Faculty Research Award.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617740646

Open Practices

All data and materials have been made publicly available via the Open Science Framework (OSF) and can be accessed at https://osf.io/u3byh and https://osf.io/h49y7, respectively. The design and analysis plans for Experiments 1, 3, 4, and 6 were preregistered at OSF (https://osf.io/h49y7/). The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617740646. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. See the Supplemental Material for full procedural details of all experiments (e.g., sampling strategies and incidental measures). The main text reports all critical information.
2. The spelling "bullseye" was used in study materials but has been changed to "bull's-eye" throughout the present article for consistency.

References

Adams, J. A. (1984). Learning of movement sequences. *Psychological Bulletin*, *96*, 3–28.

Andrieux, M., & Proteau, L. (2016). Observational learning: Tell beginners what they are about to watch and they will learn better. *Frontiers in Psychology*, *7*, Article 51. doi:10.3389/fpsyg.2016.00051

Austin, S., & Miller, L. (1992). An empirical study of the SyberVision golf videotape. *Perceptual and Motor Skills*, *74*, 875–881.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191–215.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Baumeister, R. F., Alquist, J. L., & Vohs, K. D. (2015). Illusions of learning: Irrelevant emotions inflate judgments of learning. *Journal of Behavioral Decision Making*, *28*, 149–158.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *The Journal of Political Economy*, *97*, 1232–1254.

Campbell, T., O'Brien, E., Van Boven, L., Schwarz, N., & Ubel, P. (2014). Too much experience: A desensitization bias in emotional perspective taking. *Journal of Personality and Social Psychology*, *106*, 272–285.

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, *20*, 1350–1356.

Cusack, M., Vezenkova, N., Gottschalk, C., & Calin-Jageman, R. J. (2015). Direct and conceptual replications of Burgmer & Englich (2012): Power may have little to no effect on motor performance. *PLOS ONE*, *10*(11), Article e0140806. doi:10.1371/journal.pone.0140806

Druckman, D., & Swets, J. A. (1988). *Enhancing human performance: Issues, theories, and techniques*. Washington, DC: National Academies Press.

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself*. New York, NY: Psychology Press.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*, 363–406.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*, 674–687.

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, *103*, 277–290.

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences*, *5*, 253–261.

Heyes, C. M., & Foster, C. L. (2002). Motor learning by observation: Evidence from a serial reaction time task. *The Quarterly Journal of Experimental Psychology: Section A*, *55*, 593–607.

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice-Hall.

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

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences, USA*, *104*, 19751–19756.

Marteau, T. M., Wynne, G., Kaye, W., & Evans, T. R. (1990). Resuscitation: Experience without feedback increases confidence but not skill. *British Medical Journal*, *300*, 849–850.

Mattar, A. A., & Gribble, P. L. (2005). Motor learning by observing. *Neuron*, *46*, 153–160.

Newell, K. M. (1991). Motor skill acquisition. *Annual Review of Psychology*, *42*, 213–237.

O'Brien, E. (2013). Easy to retrieve but hard to believe: Metacognitive discounting of the unpleasantly possible. *Psychological Science*, *24*, 844–851.

Professional Darts Corporation. (2018). *Rules of darts*. Retrieved from https://www.pdc.tv/players/rules-darts

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521–562.

Ruvolo, A. P., & Markus, H. R. (1992). Possible selves and performance: The power of self-relevant imagery. *Social Cognition*, *10*, 95–124.

Scully, D. M., & Newell, K. M. (1985). Observational learning and the acquisition of motor skills: Toward a visual perception perspective. *Journal of Human Movement Studies*, *11*, 169–186.

Sheffield, F. D. (1961). Theoretical considerations in the learning of complex sequential tasks from demonstration and practice. In A. A. Lumsdaine (Ed.), *Student response in programmed instruction* (pp. 13–32). Washington, DC: National Academy of Sciences–National Research Council.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science*. Advance online publication. doi:10.1177/1745691617698146

Song, H., & Schwarz, N. (2008). If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological Science*, *19*, 986–988.

Sophocles. (2015). Trachiniae. In *The dramas of Sophocles, rendered in English verse, dramatic and lyric* (G. Young, Trans.). London, England: Sagwan Press. (Original work published ~500 BC)

Statista. (2017a). *Number of Olympic Games TV viewers worldwide from 2002 to 2014 (in billions)*. Retrieved from http://www.statista.com/statistics/287966/olympic-games-tv-viewership-worldwide/

Statista. (2017b). *Number of people who watched the X Games on TV within the last 12 months in the United States from spring 2008 to spring 2015 (in millions)*. Retrieved from http://www.statista.com/statistics/229093/people-who-watched-the-x-games-on-tv-within-the-last-12-months-usa/

Stefan, K., Cohen, L. G., Duque, J., Mazzocchio, R., Celnik, P., Sawaki, L., . . . Classen, J. (2005). Formation of a motor memory by action observation. *The Journal of Neuroscience*, *25*, 9339–9346.

Ullén, F., Hambrick, D. Z., & Mosing, M. A. (2016). Rethinking expertise: A multifactorial gene-environment interaction model of expert performance. *Psychological Bulletin*, *142*, 427–446.

Van Boven, L., Loewenstein, G., Dunning, D., & Nordgren, L. F. (2013). Changing places: A dual judgment model of empathy gaps in emotional perspective taking. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 48, pp. 117–171). Burlington, VT: Academic Press.

Weaver, K., Garcia, S. M., Schwarz, N., & Miller, D. T. (2007). Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology*, *92*, 821–833.

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychological Review*, *105*, 558–584.

Wulf, G., Shea, C., & Lewthwaite, R. (2010). Motor skill learning and performance: A review of influential factors. *Medical Education*, *44*, 75–84.

Pilot Survey

Method

Participants. N = 500 participants were recruited from Amazon's Mechanical Turk (M_{age} = 34.37; SD_{age} = 11.21; 51.00% female; 69.40% Caucasian) to complete the study for $0.15.

Procedure. We preregistered the survey at https://osf.io/xrhbv/. Participants were told to imagine that they wanted to learn how to perform a new skill. Then they were asked, "If you had to choose one and only one way to get a good sense of things, which of the following do you think you would try to access FIRST?" (*Getting to repeatedly listen to someone else tell you how to perform the skill, such as having someone describe it to you verbally (and you get no other cues/information/access except this* / *Getting to repeatedly watch someone else perform the skill, such as pulling up a video of someone doing it online (and you get no other cues/information/access except this* / *Getting to repeatedly read text-based instructions about how to perform the skill, such as pulling up written text online (and you get no other cues/information/access except this)* / *All of these choices equally; I'd be equally fine just picking one at random* / *Other (something that doesn't involve any of the cues/information/access provided by these other choices)*). The first three options were presented in randomized order and the final two were fixed so that the "All" and "Other" options always appeared last.

Participants completed five items, presented in counterbalanced order, with the same choices as listed above. Each item asked participants to report a different belief: specifically, the five items asked participants which option "do you think you would try to ACCESS FIRST?"; "do you think you would try to access MOST OFTEN?"; "do you think would be most WIDELY AVAILABLE for you to access?"; "do you think would be EASIEST for you to process?"; and "do you think you would find MOST EFFECTIVE?"

Results

First we tested whether people chose each option equally, and for each item, choices differed from chance, χ^2s(4, N = 500) > 432.54, *p*s < .001. Therefore, we conducted follow-up analyses for each item.

Per our preregistration, we sought to test whether "watching" was selected relatively more frequently than any other option. Therefore, for each of the five items, we re-ran chi-square analyses using only "watching" and the second most frequently selected option for that item. Watching others was reported to be the first-sought (62.80% watching, 23.60% reading; $\chi^2(1, N = 432) = 88.93$, $p < .001$) and most frequently used learning aide (69.20% watching, 17.00% reading; $\chi^2(1, N = 431) = 158.05$, $p < .001$), and was perceived to be the most widely available (48.20% watching; 35.60% reading; $\chi^2(1, N = 419) = 9.47$, $p = .002$), the easiest to process (74.60% watching; 11.60% reading; $\chi^2(1, N = 431) = 230.22$, $p < .001$), and the most effective learning aide (72.20% watching; 11.80% reading; $\chi^2(1, N = 420) = 217.15$, $p < .001$). These results suggest that people may often watch others when they want to learn new skills, and this warrants further investigation about the actual effects of watching on people's abilities.

Experiment 1

To begin the experiment, all participants were asked if they had heard of the tablecloth trick before and generally knew what we meant (*Yes* / *No*). All participants proceeded to complete the experiment. However, as preregistered, in the main text and analyses we include only those who indicated yes (the reported sample of $N = 1{,}003$). We anticipated this exclusion rate based on a pilot test and therefore oversampled by 3.00%: in reality we requested 1,030 participants from Amazon Turk (i.e., 3.00% above an even 1,000). As it turns out, all significant

results remain significant and all null effects remain null when retaining all participants (see below for these analyses).

Our intention for this exclusion criterion was to help make the results of the "think" conditions more interpretable. These participants essentially serve as a no-information control comparison: they provide baseline estimates of abilities to perform the tablecloth trick without any additional knowledge about it—without watching a video *or* reading instructions. We were concerned that people who had never heard of the tablecloth trick but were randomly assigned to this condition would not provide meaningful estimates.

The basic effect: Perceived ability. We re-analyzed the data with all $N = 1{,}035$ participants, whether or not they indicated that they had heard of the tablecloth trick before. For our primary results, there was a main effect of Type, $F(2, 1029) = 30.95$, $p < .001$, $\eta^2 = .06$, and a main effect of Amount such that high exposure ($M = 3.16$, $SD = 1.66$) versus low exposure ($M = 2.91$, $SD = 1.62$) generally inflated participants' beliefs that they could perform the trick themselves, $F(1, 1029) = 6.54$, $p = .011$, $\eta^2 = .01$. Critically, however, this depended on the type of information they were exposed to, as evidenced by a significant interaction, $F(2, 1029) = 4.87$, $p = .008$, $\eta^2 = .01$.

First and most important, pairwise comparisons reveal the basic effect of watching: high exposure to watching someone else perform the tablecloth trick led participants to feel that they *themselves* would be significantly more likely to successfully perform the trick on their first attempt ($M = 3.83$, $SD = 1.62$)—knowing they would have no other practice or training in the interim—as compared to watching the video just once ($M = 3.15$, $SD = 1.59$), $F(1, 1029) = 15.95$, $p < .001$, $\eta_p^2 = .02$ ($d = 0.43$), 95% $CI_{difference}$ [0.35, 1.02]. As hypothesized, merely

watching others perform many times increased perceptions of one's own ability to perform the same skill.

Second, this boost did *not* emerge from overexposure to other kinds of information (reading or thinking): high exposure to text instructions did not significantly increase one's own perceived abilities (M = 3.11, SD = 1.57) as compared to low exposure to text instructions (M = 3.01, SD = 1.62), $F(1, 1029) = 0.33$, $p = .565$, $\eta_p^2 < .001$ ($d = 0.06$), 95% $CI_{difference}$ [-0.44, 0.24], and likewise more time to think and imagine the trick did not significantly increase one's own perceived abilities (M = 2.53, SD = 1.54) as compared to low exposure (M = 2.55, SD = 1.59), $F(1, 1029) = 0.02$, $p = .892$, $\eta_p^2 < .001$ ($d = 0.01$), 95% $CI_{difference}$ [-0.31, 0.36]. Simply having additional time was not enough. Moreover, although access to text instructions boosted perceived abilities compared to just thinking with no other aide—as one might expect—*extensive* access to reading text instructions did not translate into correspondingly higher perceived abilities.

Manipulation check. The above results are bolstered by our results for the manipulation check. The manipulation indeed worked as intended for each type of information, as evidenced by a main effect of Amount, $F(1, 1029) = 249.60$, $p < .001$, $\eta^2 = .20$. (There was also an incidental main effect of Type, $F(2, 1029) = 206.21$, $p < .001$, $\eta^2 = .29$, and an incidental interaction, $F(2, 1029) = 34.19$, $p < .001$, $\eta^2 = .06$.) Pairwise comparisons reveal that high-exposure participants felt they were more informed than low-exposure participants, whether it was having more time to watch (M = 5.19, SD = 1.79) versus less time to watch (M = 2.95, SD = 1.73), $F(1, 1029) = 166.34$, $p < .001$, $\eta_p^2 = .14$ ($d = 1.27$), 95% $CI_{difference}$ [1.90, 2.58]; having more time to read (M = 4.95, SD = 1.73) versus less time to read (M = 2.85, SD = 1.67), $F(1, 1029) = 145.44$, $p < .001$, $\eta_p^2 = .12$ ($d = 1.24$), 95% $CI_{difference}$ [1.76, 2.45]; or having more time to think (M = 2.03, SD = 1.58) versus less time to think (M = 1.62, SD = 1.11), $F(1, 1029) =$

5.73, $p = .017$, $\eta_p^2 = .01$ ($d = 0.30$), 95% $CI_{difference}$ [0.08, 0.76]. When re-running the analyses to compare only the "watch" and "read" conditions, there is only the key main effect of Amount, $F(1, 685) = 272.10$, $p < .001$, $\eta^2 = .28$, with no interaction, $F(1, 685) = 0.27$, $p = .605$, $\eta^2 < .001$, nor was there a main effect of Type, $F(1, 685) = 1.62$, $p = .203$, $\eta^2 = .002$. Together, these findings suggest that the basic effect applies most directly to watching, presumably due to the especially vivid, direct, and dynamic information about *what* to do that watching provides (see also our survey in the introduction).

Experiment 2

After the other dependent measures, we included some supplementary measures. Predictors rated, "How talented at dart throwing was the person in the video?" (1 = *not talented at all*; 7 = *extremely talented*). Performers rated, "Was the dart throwing task easier or harder than you expected?" (1 = *much easier than I expected*; 4 = *about what I expected*; 7 = *much harder than I expected*); "How well do you feel you performed?" (1 = *not well at all*; 7 = *extremely well*); and "How satisfied do you feel with your dart throw?" (1 = *not at all satisfied*; 7 = *extremely satisfied*). All participants rated, "How easy is it to hit the bullseye?" (1 = *not easy at all*; 7 = *extremely easy*); "How hard is it to hit the bullseye?" (1 = *not hard at all*; 7 = *extremely hard*): predictors made these ratings after making other predictions, and performers made these ratings after actually throwing the dart and completing the other post-performance items. Finally, all participants also rated, "Prior to this study, how much dart throwing experience did you have?" (1 = *no dart throwing experience at all*; 7 = *lots of dart throwing experience*).

The dart thrower was rated as marginally more talented among low-exposure predictors ($M = 5.72$, $SD = 1.03$) compared to high-exposure predictors ($M = 5.30$, $SD = 1.40$), $t(111) = 1.80$, $p = .075$, $d = 0.34$, 95% $CI_{difference}$ [-0.87, 0.04]. Among performers, ratings of whether the

dart throw was less or more challenging than expected did not differ between low exposure (M = 4.05, SD = 1.39) and high exposure (M = 4.48, SD = 1.35), $t(78) = 1.38$, $p = .171$, $d = 0.31$, 95% $CI_{difference}$ [-0.19, 1.03], although pooling across levels of exposure, performers reported that the dart throw was marginally more difficult than they expected compared to the scale midpoint of 4, "about what I expected", $t(79) = 1.79$, $p = .078$, $d = 0.20$, 95% *CI* [3.97, 4.58]. Evaluations of one's own dart throw did not differ between low exposure (M = 3.11, SD = 1.90) and high exposure (M = 2.88, SD = 1.94), $t(78) = -0.52$, $p > .250$, $d = 0.12$, 95% $CI_{difference}$ [-1.08, 0.63]. Likewise, satisfaction with one's own dart throw did not differ between low exposure (M = 3.47, SD = 1.91) and high exposure (M = 2.81, SD = 1.94), $t(78) = -1.54$, $p = .128$, $d = 0.34$, 95% $CI_{difference}$ [-1.52, 0.20].

Among predictors, hitting the bullseye was rated as less easy, and more hard, among low-exposure predictors (M_{easy} = 2.07, SD_{easy} = 1.02, M_{hard} = 5.63, SD_{hard} = 1.03) than among high-exposure predictors (M_{easy} = 2.88, SD_{easy} = 1.48, M_{hard} = 5.14, SD_{hard} = 1.38), ts(111) = -3.38, 2.14, ps = .001, .035, ds = 0.64, 0.40. Among performers, after throwing the dart, ratings of how easy and how hard it was to hit the bullseye did not differ between low exposure (M_{easy} = 2.26, SD_{easy} = 1.00, M_{hard} = 5.61, SD_{hard} = 1.08) and high exposure (M_{easy} = 2.29, SD_{easy} = 1.47, M_{hard} = 5.48, SD_{hard} = 1.35), ts(78) = 0.08, -0.47, ps = .937, .640, ds = 0.02, 0.11.

Finally, prior dart throwing experience did not vary by condition, $F(3, 189) = 0.42$, $p > .250$, $\eta^2 = .007$.

Experiment 3

In the moonwalk phase of the study, participants were informed about both levels of exposure before they watched the moonwalk training video. Specifically, they were told that participants in "Condition 1x" would watch the video 1x, then attempt the move themselves,

whereas participants in "Condition 20x" would watch the video 20x in a row, then attempt the move themselves. We did this because the raters in the second phase of the study would view both low-exposure and high-exposure moonwalks within-subjects during the same study session, and we wanted to ensure that moonwalkers were fully aware of the ratings procedure.

Before the prediction measures, we reminded participants that raters would see both low-exposure and high-exposure moonwalks within-subjects. Specifically, participants responded to the item: "Based on your training, how good do you feel your attempt will be, relatively speaking? Keep in mind that YOU saw the video 1x [20x] for training while other participants will see the video 20x [1x]. We'll show all of these videos to outside raters. Given this, predict how an average rater would rate YOUR attempt."

Experiment 4

The majority of participants (89.26%) reported following instructions to watch the video passively, with no differences by condition (90.55% of low-exposure participants, 88.11% of high-exposure participants), $\chi^2(1, N = 270) = 0.42$, $p > .250$; the majority of participants reported not practicing before playing ($M = 1.51$, $SD = 0.92$), with no differences by condition ($M = 1.57$, $SD = 0.99$ among low-exposure participants, $M = 1.45$, $SD = 0.85$ among high-exposure participants), $t(268) = 1.14$, $p = .257$, $d = 0.14$, 95% $CI_{difference}$ [-0.09, 0.35]; and the majority of participants (86.67%) reported following instructions to trace the maze quickly, with no differences by condition (86.61% of low-exposure participants, 86.71% of high-exposure participants), $\chi^2(1, N = 270) = .001$, $p > .250$. These results help further isolate the effect of watching by suggesting all participants had otherwise similar study experiences.

Experiment 5

We conducted the same analyses except treating each item in the scale individually. For each item there was a main effect of Performer (improvement: $F(1, 396) = 8.85$, $p = .003$, $\eta^2 = .02$; preparation: $F(1, 396) = 10.83$, $p < .001$, $\eta^2 = .03$; learning: $F(1, 396) = 36.14$, $p < .001$, $\eta^2 = .08$) and a main effect of Exposure (improvement: $F(1, 396) = 11.55$, $p < .001$, $\eta^2 = .03$; preparation: $F(1, 396) = 7.68$, $p = .006$, $\eta^2 = .02$; learning: $F(1, 396) = 18.87$, $p < .001$, $\eta^2 = .05$). The interaction was significant for one item and marginally significant for the other two (improvement: $F(1, 396) = 3.05$, $p = .081$, $\eta^2 = .01$; preparation: $F(1, 396) = 2.81$, $p = .095$, $\eta^2 = .01$; learning: $F(1, 396) = 5.01$, $p = .026$, $\eta^2 = .01$).

Pairwise comparisons reveal the same basic effect for each item: viewers who could see the actual performer and his actions reported significantly greater skill acquisition after high versus low exposure (improvement: $M_{View1x} = 1.94$, $SD_{View1x} = 1.26$, $M_{View20x} = 2.68$, $SD_{View20x} = 1.65$, $F(1, 396) = 12.92$, $p < .001$, $\eta_p^2 = .03$ ($d = 0.51$), 95% $CI_{difference}$ [0.34, 1.15]; preparation: $M_{View1x} = 2.23$, $SD_{View1x} = 1.64$, $M_{View20x} = 2.96$, $SD_{View20x} = 1.77$, $F(1, 396) = 9.65$, $p = .002$, $\eta_p^2 = .02$ ($d = 0.44$), 95% $CI_{difference}$ [0.27, 1.18]; learning: $M_{View1x} = 2.29$, $SD_{View1x} = 1.31$, $M_{View20x} = 3.22$, $SD_{View20x} = 1.68$, $F(1, 396) = 21.15$, $p < .001$, $\eta_p^2 = .05$ ($d = 0.66$), 95% $CI_{difference}$ [0.53, 1.33]). In contrast, there were no systematic differences across exposure among viewers who watched the cropped video (improvement: $M_{View1x} = 1.76$, $SD_{View1x} = 1.44$, $M_{View20x} = 2.00$, $SD_{View20x} = 1.39$, $F(1, 396) = 1.40$, $p = .238$, $\eta_p^2 = .004$ ($d = 0.17$), 95% $CI_{difference}$ [-0.63, 0.16]; preparation: $M_{View1x} = 1.97$, $SD_{View1x} = 1.57$, $M_{View20x} = 2.15$, $SD_{View20x} = 1.53$, $F(1, 396) = 0.62$, $p > .250$, $\eta_p^2 = .002$ ($d = 0.11$), 95% $CI_{difference}$ [-0.63, 0.27]; learning: $M_{View1x} = 1.75$, $SD_{View1x} = 1.28$, $M_{View20x} = 2.05$, $SD_{View20x} = 1.36$, $F(1, 396) = 2.27$, $p = .133$, $\eta_p^2 = .006$ ($d = 0.21$), 95% $CI_{difference}$ [-0.69, 0.09]).

Experiment 6

We conducted the same analyses except treating each item in the scale individually. For each item there was no main effect of condition (improvement: $F(2, 142) = 0.27$, $p > .250$, $\eta^2 = .004$; preparation: $F(2, 142) = 0.32$, $p > .250$, $\eta^2 = .005$; learning: $F(2, 142) = 1.11$, $p > .250$, $\eta^2 = .02$); a main effect of time (improvement: $F(1, 142) = 8.46$, $p = .004$, $\eta^2 = .06$; preparation: $F(1, 142) = 30.52$, $p < .001$, $\eta^2 = .18$; learning: $F(1, 142) = 24.35$, $p < .001$, $\eta^2 = .15$); and a significant interaction (improvement: $F(2, 142) = 16.92$, $p < .001$, $\eta^2 = .19$; preparation: $F(2, 142) = 9.28$, $p < .001$, $\eta^2 = .12$; learning: $F(2, 142) = 9.22$, $p < .001$, $\eta^2 = .12$).

Pairwise comparisons reveal the same basic effect for each item: viewers given a "taste" of the experience reported significant drops on all items (improvement: $M_{Time1} = 2.70$, $SD_{Time1} = 1.33$, $M_{Time2} = 1.94$, $SD_{Time2} = 1.01$, $F(1, 142) = 40.21$, $p < .001$, $\eta_p^2 = .22$ ($d = 0.68$), 95% $CI_{difference}$ [0.53, 1.01]; preparation: $M_{Time1} = 3.04$, $SD_{Time1} = 1.33$, $M_{Time2} = 2.13$, $SD_{Time2} = 0.95$, $F(1, 142) = 44.02$, $p < .001$, $\eta_p^2 = .24$ ($d = 0.71$), 95% $CI_{difference}$ [0.64, 1.19]; learning: $M_{Time1} = 3.45$, $SD_{Time1} = 1.49$, $M_{Time2} = 2.62$, $SD_{Time2} = 1.41$, $F(1, 142) = 39.45$, $p < .001$, $\eta_p^2 = .22$ ($d = 0.88$), 95% $CI_{difference}$ [0.57, 1.09]). However, there were no significant drops among participants in the explanation condition (improvement: $M_{Time1} = 2.17$, $SD_{Time1} = 1.42$, $M_{Time2} = 2.31$, $SD_{Time2} = 1.52$, $F(1, 142) = 1.49$, $p = .224$, $\eta_p^2 = .01$ ($d = 0.19$), 95% $CI_{difference}$ [-0.38, 0.09]; preparation: $M_{Time1} = 2.58$, $SD_{Time1} = 1.41$, $M_{Time2} = 2.42$, $SD_{Time2} = 1.15$, $F(1, 142) = 1.49$, $p = .224$, $\eta_p^2 = .01$ ($d = 0.18$), 95% $CI_{difference}$ [-0.10, 0.44]; learning: $M_{Time1} = 3.46$, $SD_{Time1} = 1.73$, $M_{Time2} = 3.35$, $SD_{Time2} = 1.56$, $F(1, 142) = 0.64$, $p > .250$, $\eta_p^2 = .004$ ($d = 0.12$), 95% $CI_{difference}$ [-0.15, 0.36]). Nor were there any significant drops among participants in the technical information condition (improvement: $M_{Time1} = 2.42$, $SD_{Time1} = 1.37$, $M_{Time2} = 2.44$, $SD_{Time2} = 1.42$, $F(1, 142) = 0.03$, $p > .250$, $\eta_p^2 < .001$ ($d = 0.04$), 95% $CI_{difference}$ [-0.25, 0.21]; preparation: $M_{Time1} = 2.80$, $SD_{Time1} = 1.31$, $M_{Time2} = 2.58$, $SD_{Time2} = 1.36$, $F(1, 142) = 2.71$, $p = .102$, $\eta_p^2 = .02$ ($d = 0.47$), 95%

$CI_{difference}$ [-0.04, 0.48]; learning: M_{Time1} = 3.54, SD_{Time1} = 1.59, M_{Time2} = 3.36, SD_{Time2} = 1.63, $F(1, 142) = 1.98$, $p = .162$, $\eta_p^2 = .01$ ($d = 0.21$), 95% $CI_{difference}$ [-0.07, 0.43]).

Additionally, note that we preregistered analyses of T1-T2 difference scores, for each condition individually (https://osf.io/d8w63/). All preregistered analyses were significantly confirmed and we report these analyses below. In hindsight, the interaction within a Repeated Measures ANOVA is the optimal test, so we report this in the main text instead instead. For the preregistered difference-score analyses, we first conducted these analyses for the perceived skill acquisition scale and then we proceeded to analyze the individual scale items as well as the perceived ability item.

For each participant we computed difference scores using the perceived skill acquisition scale (T2 minus T1) and performed a one-way ANOVA with condition as the independent variable and the difference score as the dependent variable. The effect of condition was significant, $F(2, 142) = 17.07$, $p < .001$, $\eta^2 = 0.19$. Next, we conducted planned, pairwise comparisons. The decline in perceived skill acquisition was greater in the personal experience condition ($M_{difference}$ = -0.84, $SD_{difference}$ = 0.94) than in the explanation condition ($M_{difference}$ = -0.04, $SD_{difference}$ = 0.71), $t(142) = 5.32$, $p < .001$, $d = 1.09$, 95% $CI_{difference}$ [0.50, 1.09], and greater in the personal experience condition than in the technical information condition ($M_{difference}$ = -0.13, $SD_{difference}$ = 0.48), $t(142) = 4.80$, $p < .001$, $d = 0.97$, 95% $CI_{difference}$ [0.42, 1.00]. The decline in perceived skill acquisition did not differ between the explanation and technical information conditions, $t(142) = -0.58$, $p > .250$, 95% $CI_{difference}$ [-0.21, 0.38], $d = 0.12$.

Next we performed the same "difference score" analyses for each scale item individually. For each of the three "perceived skill acquisition" items (improvement, preparation, learning), we computed a difference score for each participant (T2 minus T1) and then performed a one-

way ANOVA with condition as the independent variable and the difference score as the dependent variable. For each perceived skill acquisition item the effect of condition was significant, Fs(2,142) = 9.22, ps < .001, η_p^2 > 0.11. Next, we conducted planned, pairwise comparisons. The declines for each "perceived skill acquisition" item were greater in the personal experience condition ($M_{\text{difference-improve}}$ = -0.77, $SD_{\text{difference-improve}}$ = 1.13; $M_{\text{difference-prepare}}$ = -0.91, $SD_{\text{difference-prepare}}$ = 1.28; $M_{\text{difference-learn}}$ = -0.83, $SD_{\text{difference-learn}}$ = 0.94) than in the explanation condition ($M_{\text{difference-improve}}$ = 0.15, $SD_{\text{difference-improve}}$ = 0.77; $M_{\text{difference-prepare}}$ = -0.17, $SD_{\text{difference-prepare}}$ = 0.93; $M_{\text{difference-learn}}$ = -0.10, $SD_{\text{difference-learn}}$ = 0.90), ts(142) > 3.86, ps < .001, ds > 0.79, and greater in the personal experience condition than in the technical information condition ($M_{\text{difference-improve}}$ = 0.02, $SD_{\text{difference-improve}}$ = 0.47; $M_{\text{difference-prepare}}$ = -0.22, $SD_{\text{difference-prepare}}$ = 0.46; $M_{\text{difference-learn}}$ = -0.18, $SD_{\text{difference-learn}}$ = 0.87), ts(142) > 3.53, ps < .001, ds > 0.72. Declines for each of these items did not differ between the explanation and technical information conditions, ts(142) < 0.75, ps > .250, ds < 0.15.

Finally, we conducted the same analyses for the "perceived ability" item and obtained similar results. For each participant we computed a difference score for perceived ability (T2 minus T1) and performed a one-way ANOVA with condition as the independent variable and difference score as the dependent variable. The effect of condition was significant, $F(2, 142)$ = 12.62, p < .001, η^2 = 0.15. Next, we conducted planned, pairwise comparisons. The decline in perceived ability was greater in the personal experience condition ($M_{\text{difference}}$ = -0.45, $SD_{\text{difference}}$ = 0.75) than in the explanation condition ($M_{\text{difference}}$ = 0.04, $SD_{\text{difference}}$ = 0.50), $t(142) = 4.26$, p < .001, d = 0.87, 95% $\text{CI}_{\text{difference}}$ [0.26, 0.72], and greater in the personal experience condition than in the technical information condition ($M_{\text{difference}}$ = 0.06, $SD_{\text{difference}}$ = 0.37), $t(142) = 4.46$, p < .001, d = 0.91, 95% $\text{CI}_{\text{difference}}$ [0.28, 0.73]. Declines in perceived ability did not differ between

the explanation and technical information conditions, $t(142) = 0.16$, $p > .250$, 95% $CI_{difference}$ [-0.20, 0.24], $d = 0.03$.

OPEN PRACTICES DISCLOSURE

PLEASE COMPLETE AND RETURN TO EDITORIALOFFICE@PSYCHOLOGICALSCIENCE.ORG

***Psychological Science* manuscript #:** **PSCI-17-0471.R1** **Corresponding author:** **Michael Kardas**

Articles accepted to *Psychological Science* after January 1, 2014, are eligible to earn badges that recognize open scientific practices: publicly available data, material, or preregistered research plans. Please read more about the badges on our Open Practice Badges page, and you can also find information in the Open Science Framework wiki and FAQ.

☐ **Please check this box if you are not interested in participating.**

If you choose to participate, this form will be posted with your article as supplemental online material.

To apply for one or more badges acknowledging open practices, please check the appropriate box(es) below and provide the information requested in the relevant sections. You will not qualify for a badge for a given item unless you can provide a URL, doi, or other **permanent path** for accessing the specified information in a **public, open-access repository. Qualifying public, open-access repositories are committed to preserving data, materials, and/or registered analysis plans and keeping them publicly accessible via the web into perpetuity. Files must be registered with a time stamp and must not be able to be changed at a later time.** Examples of qualifying repositories include the Open Science Framework (OSF) and the various Dataverse networks. Hundreds of other qualifying data/materials repositories are listed at http://re3data.org/ and http://databib.org/. Preregistration of an analysis plan must take place via a publicly accessible registry system (e.g., OSF, ClinicalTrials.gov or other trial registries in the WHO Registry Network, institutional registration systems). **Personal websites and most departmental websites do not qualify as repositories.**

Authors who wish to publicly post third-party material in their data, materials, or preregistration plan must have the proper authority or permission agreement in order to do so.

There are circumstances in which it is not possible or advisable to share any or all data, materials, or a research plan publicly. For example, there are cases in which sharing participants' data could violate confidentiality. If you would like your article to include an explanation of such circumstances and/or provide links to any data or materials you have made available—even if not under conditions eligible to earn a badge—you may write an alternative note that will be published in the Open Practices note in the article**.** Please check this box if you would like your article to include an alternative note and provide the text of the note below:

☐ **Alternative Note:**

☑ **Application for Open Data Badge**

1. Provide the URL, doi, or other **permanent path** for accessing the data in a **public, open-access repository**:
   https://osf.io/u3byh/

☑ Confirm that there is sufficient information for an independent researcher to reproduce **all of the reported results**, including codebook if relevant.

☑ Confirm that you have registered the uploaded files so that they are **time stamped** and cannot be changed.

☑ Application for Open Materials Badge

1. Provide the URL, doi, or other **permanent path** for accessing the materials in a **public, open-access repository**:

   https://osf.io/h49y7/

   ☑ Confirm that there is sufficient information for an independent researcher to reproduce **all of the reported methodology**.

   ☑ Confirm that you have registered the uploaded files so that they are **time stamped** and cannot be changed.

☑ Application for Preregistered Badge

1. Provide the URL, doi, or other **permanent path** to the **public** registration in a **public, open-access repository**.*
   https://osf.io/h49y7/

2. Was the analysis plan registered prior to examination of the data or observing the outcomes? If no, explain.**
   Yes

3. Were there additional registrations for the study other than the one reported? If yes, provide links and explain.*
   No

4. Were there any changes to the preregistered analysis plan for the primary confirmatory analysis? If yes, explain.**
   No

5. Are all of the analyses described in the registered plan reported in the article? If no, explain.*
   Yes

*No badge will be awarded if (1) is not provided, **or** if (3) is answered "yes" without strong justification, **or** if (5) is answered "no" without strong justification.
**If the answer to (2) is "no," the notation DE (Data Exist) will be added to the badge, indicating that registration postdates realization of the outcomes but predates analysis. If the answer to (4) is "yes" with strong justification for changes, the notation TC (Transparent Changes) will be added to the badge, indicating that the analysis plan was altered but the preregistered analyses and rationale for the change are provided.

By signing below, authors affirm that the above information is accurate and complete, that any third-party material has been reproduced or otherwise made available only with the permission of the original author or copyright holder, and that publicly posted data do not contain information that would allow individuals to be identified without consent.

Name: **Michael Kardas** **Date:** **10/5/17**