Stage 2 Registered Report

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
2026, Vol. 37(2) 125–135

DOI: 10.1177/09567976251412721
www.psychologicalscience.org/PS
Sage

Registered Report: A Replication Examining Occupational Experience and Performance on the Water-Level Task

Elizabeth R. Tenney[1], Kylie Rochford[1], Amelia Stillwell[1], Coco Xinyue Liu[1], David Tannenbaum[1], Marie Hennecke[2], Jeanine K. Stefanucci[3], B. Ariel Blair[4], Jesse Graham[1], and Bryan L. Bonner[1]

[1]Management Department, David Eccles School of Business, University of Utah; [2]Department of Psychology, Ruhr-Universität Bochum; [3]Department of Psychology, University of Utah; and [4]Department of Business Administration and Marketing, Goddard School of Business & Economics, Weber State University

Abstract

This is a registered report to directly replicate the primary finding in Hecht and Proffitt (1995). Hecht and Proffitt found that those with occupational experience handling liquid in containers performed worse at solving a water-level problem than those in occupations that did not require handling liquids. Shortly after, Vasta et al. (1997) found the opposite: Experience was associated with superior performance on the task. The conflicting findings and the small sample sizes in each study leave the relationship between experience and water-level-task performance uncertain. We addressed these concerns with a high-powered direct replication of Hecht and Proffitt with adults in Germany (N = 407). We failed to replicate Hecht and Proffitt's results, finding that their study had less than 33% power to detect the small, nonsignificant difference that we observed between groups.

Keywords

occupational experience, water-level task, registered report, replication, perception, expertise, intuitive physics, cognitive psychology

Received 3/13/24; revision accepted 11/20/25

In 1995, Hecht and Proffitt documented evidence that experience can contribute to failure on certain intuitive physics tasks. Contrary to conventional wisdom that experience does not impair (and often improves) performance, individuals with professional experience handling liquid in containers (i.e., waitresses and bartenders) were more prone to errors than those without experience (i.e., bus drivers, students, and housewives) on a water-level problem-solving task. For this task, participants indicate the angle or level of a liquid in a tilted glass (for an illustration, see Fig. 1a; Piaget & Inhelder, 1948/1956). The laws of gravity dictate that the surface of a liquid stays level to the ground regardless of a container's tilt, whereas some people incorrectly anticipate that liquid levels correspond to the angle of its container.

In Hecht and Proffitt (1995), "experts" were bartenders and servers with over 5 years of experience at Munich's Oktoberfest, carrying 1-liter mugs of beer to customers a considerable distance away. The authors posited that repeatedly carrying liquid-filled glasses (with a focus on not spilling) leads workers to adopt an object-relative frame of reference. In this frame, people focus on the orientation of the water relative to its container. In contrast, when people adopt an environment-relative reference frame, the focus is on the orientation of the water relative to the ground (McAfee

Corresponding Author:
Elizabeth R. Tenney, University of Utah, David Eccles School of Business, Management Department
Email: elizabeth.tenney@eccles.utah.edu

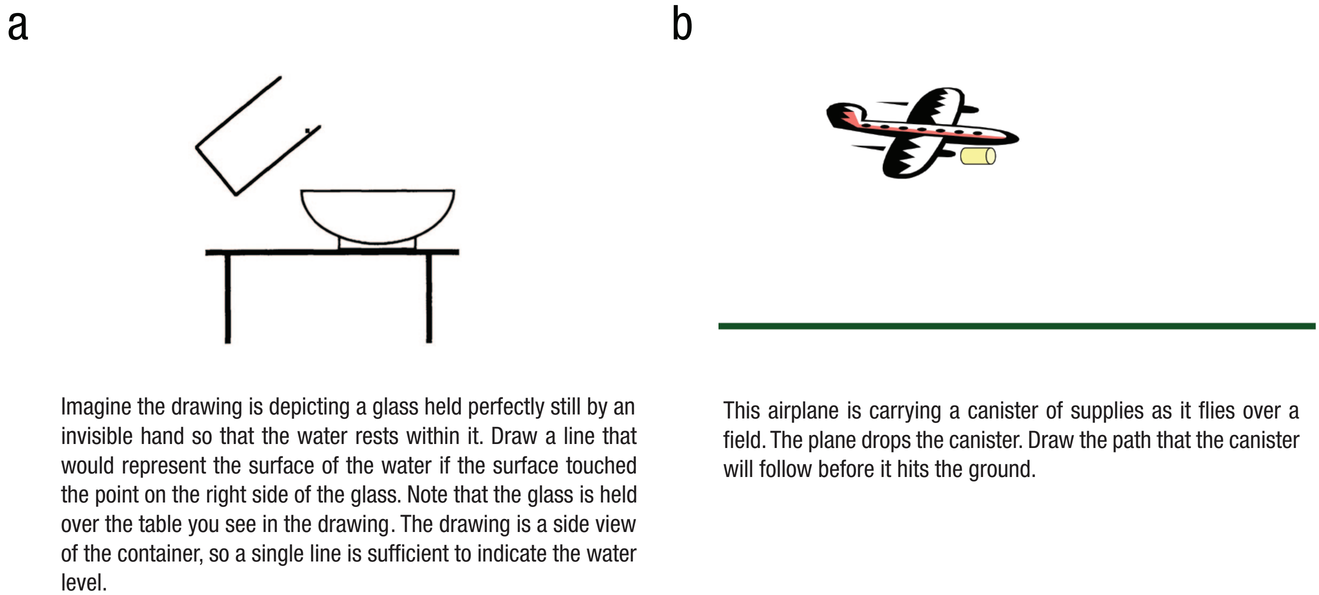

Fig. 1. Intuitive physics tasks: (a) water-level task and (b) falling-object task.

& Proffitt, 1991). Hecht and Proffitt (1995) argued that adopting an object-relative perspective creates a perceptual bias that is associated with greater errors on the water-level task. Their findings supported this premise and offered a rare example of relevant task experience undermining task performance.

Two years later, Vasta et al. (1997) attempted a replication and produced findings in the other direction. In their study, workers with occupational experience handling liquid demonstrated superior performance on the water-level task. Vasta et al. (1997) attribute their divergent findings to their study being better controlled than Hecht and Proffitt's (1995), with experimental and control groups better matched for age, education, and gender. Yet in both studies, sample sizes were small by modern empirical standards and likely underpowered (n = 20 per cell), raising the possibility that both findings may have capitalized on chance variation. Additionally, although the high-experience participants in Vasta et al. (1997) had several years of experience, they may not exhibit the same depth of expertise as Munich's Oktoberfest servers or the event city's bartenders. Vasta et al.'s control group also differed from that of Hecht and Proffitt; it consisted of salespeople and clerical employees, rather than bus drivers, students, and housewives, as in the original study. Finally, as noted by Vasta et al., the different results could also be influenced or explained by contextual discrepancies between the two studies (e.g., the studies were conducted in different countries). In sum, the role of occupational experience on performance in the water-level task is still unclear.

Findings from Hecht and Proffitt (1995) and Vasta et al. (1997) have been cited over 230 times according to Google Scholar (as of June 13, 2024) and have been used to inform subsequent research in a range of scientific disciplines, including cognitive psychology (Bilalić et al., 2008), marketing (Kilgour & Koslow, 2009), management (Dane, 2010) and education (Matthes et al., 2024). Often articles cite Hecht and Proffitt to evoke the general idea that experience can impair performance, without recognizing the conflicting findings from Vasta et al. (e.g., Dane, 2010; Dror et al., 2011; Kilgour & Koslow, 2009; Matthes et al., 2024). Given the ambiguity created by the inconsistent findings in these two articles, attempting a high-powered in-the-field replication is a worthwhile endeavor, in line with recent propositions that ability to remove uncertainty should be a major consideration for replication targets (Hardwicke et al., 2018; Isager et al., 2023). We particularly highlight the value of registered reports and direct replications for verifying empirical findings with modern methodological standards (e.g., preregistration; Nosek et al., 2022; Zwaan et al., 2018). Unlike many in the scientific community who may prioritize ease of data accessibility (Pittelkow et al., 2023), we embrace the challenges presented by data collection in field settings to rigorously test the focal hypotheses in real-world conditions.

The Current Study

Our goal is to assess whether the primary finding in Hecht and Proffitt (1995) replicates. The literature points to different possibilities. According to Hecht and Proffitt, because of occupational demands, servers and bartenders are attuned to not spilling liquid and focus on the level of a liquid relative to the rim of the glass (an object orientation), rather than relative to the ground, when solving the water-level problem. Thus, occupational experience should impair performance on the water-level task. On the other hand, Vasta et al.'s (1997) replication found that occupational experience improved performance on the water-level task for both men and women and supports the commonly held belief that experts often outperform novices in tasks relevant to their area of expertise (for a review, see Ericsson & Lehmann, 1996). It is also possible that occupational experience has no meaningful effect on whether a person adopts an object-relative or environment-relative reference system, and thus has no impact on performance on the water-level task. Holding a liquid in a glass is a common everyday experience, so additional occupational experience may not alter performance.

Extension: Occupational Experience and Falling-Object Task Performance

To further investigate the mechanism proposed in the original article, we built on Hecht and Proffitt (1995) by including an additional intuitive physics task unrelated to fluid dynamics. Specifically, we asked participants to anticipate the trajectory of a moving object as it fell (i.e., the "falling objects problem"; McCloskey, 1983; Riener et al., 2005). As shown in Figure 1b, some participants, when presented with this task, expected the object to fall straight down or diagonally instead of in a forward parabolic motion. If occupational experience related to handling liquid is the primary cause of poor performance on the water-level task, as suggested by Hecht and Proffitt, then we should observe impaired performance only for intuitive physics tasks related to liquid dynamics and not for other intuitive physics tasks, such as the falling-object task (i.e., classical mechanics).

Research Transparency Statement

General disclosures

Conflicts of interest: All authors declare no conflicts of interest. **Funding:** This research was financially supported by a grant from the Marriner S. Eccles Institute for Economics and Quantitative Analysis at the University of Utah, awarded to the first author. **Artificial intelligence:** ChatGPT and Copilot were used as intelligent thesaurus and copyediting tools during the creation of this article. No other artificial-intelligence-assisted technologies were used in this research or the creation of this article. **Ethics:** This research received approval from the University of Utah ethics board (ID No. IRB_00174435) and met the ethical guidelines and legal requirements of Ruhr-University Bochum, Germany. **Open Science Framework (OSF):** To ensure long-term preservation, we registered all OSF files at https://doi.org/10.17605/OSF.IO/YCV82.

Study disclosures

Preregistration: The hypotheses, methods, and analysis plans were preregistered as a Stage 1 registered report (https://doi.org/10.17605/OSF.IO/ETQ5A) on July 16, 2024, prior to data collection, which began on September 17, 2024. There were minor deviations from the preregistration (for details, see Table S1 in the Supplemental Material available online). **Materials:** All study materials are publicly available (https://osf.io/e6hps/). **Data:** All primary data are publicly available (https://osf.io/e6hps/). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/e6hps/). **Computational reproducibility:** The computational reproducibility of the results in the main article (but not the Supplemental Material) has been independently confirmed by the journal's STAR team.

Method

Design

Similar to Hecht and Proffitt (1995), we compared participants with occupational experience in handling liquids (servers and bartenders) with participants inexperienced in handling liquids professionally (bus drivers and students).[1] There was no randomization or counterbalancing. Researchers were not blind to groups, but to limit experimenter demand effects we followed a protocol for interacting with participants, and we handed participants paper materials without verbally explaining the task itself.[2]

Sample size

Hecht and Proffitt (1995) combined servers and bartenders into an experienced group ($n = 40$) to compare with those in the inexperienced group ($n = 80$). We planned to obtain a sample size 2.5 times that of the original article (see Simonsohn, 2015). We aimed to balance sample size across groups (whereas Hecht and Proffitt had fewer experienced than inexperienced participants), and we obtained a sample size of 207 experienced and 200 inexperienced participants (i.e.,

Table 1. Descriptive Statistics

Variable	Inexperienced			Experienced		
	Students	Bus drivers	Total	Servers	Bartenders	Total
n[a]	153	17	170	146	54	200
Gender[b]						
Men	50	14	64	74	30	104
Women	92	3	95	69	24	93
Age[c]	21 (2.7)	47 (10.0)	24 (8.7)	31 (10.2)	30 (9.9)	31 (10.1)
Physics[c]	4.4 (2.7)	3.8 (3.1)	4.3 (2.8)	5.0 (2.6)	4.4 (3.5)	4.9 (2.9)
Beverage service[c]	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	9.5 (8.1)	7.7 (8.7)	9.0 (8.3)
Education[d]	5.1 (.39)	3.6 (1.3)	5.0 (.72)	4.6 (1.3)	4.4 (1.5)	4.5 (1.4)
WLT error[e]	8.8 (11)	15 (16)	9.5 (11)	6.4 (8.7)	17 (18)	9.2 (13)
WLT correct[f]	54%	29%	51%	65%	43%	59%
FOT correct[g]	42%	41%	42%	34%	23%	31%

Note: Sample size per cell may vary because of missing values. WLT = water-level task; FOT = falling-object task.
[a]We excluded thirty participants who reported some degree of beverage-service experience in the inexperienced group and 8 participants who failed to draw a line in the glass (one person was excluded for both reasons, so 37 participants were excluded in total). [b]Other gender categories were excluded. [c]Mean years are shown with standard deviations in parentheses. [d]The scale ranged from 1 to 7, excluding 1 additional participant who selected "other." Standard deviations are in parentheses. [e]The mean absolute error in degrees on the water-level task with standard deviations in parentheses. [f]The percentage correct on the water-level task. [g]The percentage correct on the falling-object task.

2.5 times that of the largest group in Hecht and Proffitt). Our sample consisted of 186 men, 207 women, 3 gender diverse, 5 who chose not to disclose, and 6 nonresponses. The average age in our sample was 27.3 years, with an age range of 17 to 65 years (7 participants did not respond to the question about age). See Table 1 for demographics of each group.

We had planned for even sample sizes across all four subsamples of participants, but because of unforeseen contingencies, we made changes during data collection that deviated from our Stage 1 preregistration. We were unable to collect responses from 100 bartenders by the end of Oktoberfest, because it was harder to reach bartenders than servers, and instead increased the number of servers in our sample to reach our target of 200 experienced participants (i.e., 60 bartenders and 147 servers). We were also unable to collect responses from 100 bus drivers by the end of Oktoberfest because the rest stop we had planned for recruitment was closed for renovation, and many bus drivers we approached declined to participate, did not speak German (the language of our survey), or both. For this reason, we decided to continue recruiting bus drivers in another city in Germany after Oktoberfest ended, but this also proved difficult (i.e., we collected data from 6 bus drivers during Oktoberfest and 14 bus drivers after; see the preregistration-deviation information in Table S1 in the Supplemental Material available online). Consistent with our preregistration, after learning the total number of bus drivers reached, we collected additional responses from university students to reach our target of 200 inexperienced participants.[3] We stopped data collection when we reached our target sample size for each group (oversampling by seven because we noticed that several participants did not draw a line in the glass and would have to be omitted; see the preregistration-deviation table).

Exclusions

We excluded 30 participants in the inexperienced group who reported some degree of beverage-service experience (i.e., greater than 0 years of experience). We note that there were 117 missing responses to this question, primarily from university students (and that all missing responses were from participants in the inexperienced group), presumably because they thought the question did not apply to them and so did not answer it. We planned to exclude participants who drew a three-dimensional shape (rather than a line) on the water-level task, but no participants did so. Finally, we excluded 7 participants who failed to draw a line in the glass at all (one additional participant did this, but was already excluded by the other criterion). These exclusions apply to all reported analyses, and they left us with 170 inexperienced and 200 experienced participants.

Further, we also analyzed our data after performing a second round of exclusions similar to those used by Hecht and Proffitt (1995) and Vasta et al. (1997). We excluded participants who reported previous knowledge of the water-level task, as well as bartenders or servers who reported fewer than 5 years of experience.

This second round of exclusions left us with 162 inexperienced participants and 123 experienced participants. Because we did not observe meaningful differences in results[4] when we compared the first and second set of exclusions (i.e., there were similar effect sizes, and virtually all significant and nonsignificant results did not change when we used the more restrictive exclusion set), we report results below only using the larger, more inclusive sample; we report results using our second round of exclusions in the Supplemental Material.

Materials

One of the current authors (a bilingual German/English speaker) translated the original materials from Hecht and Proffitt (1995) into German. These materials were then revised in response to feedback from the original Hecht and Proffitt (1995) authorship team. Final materials were verified via back translation by a different researcher not on the authorship team (Brislin, 1976). As in the original study, participants responded using paper-and-pencil surveys.

Measures

Water-level task. In this task from Piaget and Inhelder (1948/1956), participants were presented with a cross-section drawing of a glass, tilted 50° clockwise from vertical. The glass is held above a bowl sitting on a table; the surface of the table is parallel to the ground. Participants were instructed to draw a single line representing the surface of the water that connects to a point marked on the right side of the glass. Figure 1a provides an illustration of the water-level task.

Performance on the water-level task was measured by how much the line angle drawn by participants deviated from horizontal. A protractor was placed parallel to the surface of the horizontal table to measure each participant's line angle. If participants did not draw a straight line, we extracted a line from the two endpoints of their drawing (as in Barhorst-Cates et al., 2020). Data coders measuring line angles were blind to participant group. Two or three coders coded each response. If there were three coders and they were within a degree of each other, then the average was taken of the three measurements. If two of three coders produced the same number, then that number was chosen as the final number. Otherwise, discrepancies greater than 1° were resolved by discussion. After coding absolute angle of error, a team coded the direction of the error while still blind to participant occupation. See the preregistration-deviation table (Table S1) for minor departures from our planned coding method.

Filler task. We used a similar filler task to that used by Hecht and Proffitt (1995). The task depicts two containers of different diameters, the first filled with water and the second empty. Participants were asked to draw the corresponding water level in the second container, after the contents of the first container had been poured into the second. As reported by Hecht and Proffitt, this filler task was employed so that participants would not spend too much time mulling over the water-level task or wondering whether the task was a trick question. Hecht and Proffitt also had a second filler task, which we replaced with the falling-object task.

Falling-object task. In this task from McCloskey (1983), participants were presented with an image of an airplane bearing a canister and moving in a horizontal direction across the page. We asked them to draw the trajectory of the canister when it was dropped from the airplane (see Fig. 1b for an illustration). We measured the falling-object task categorically as correct or incorrect (similar to Riener et al., 2005). Data coders who coded responses on this task were blind to participant occupation. We coded responses as correct if fall lines were drawn parabolically in the direction of the airplane's flight path; all other responses (e.g., a straight line in the direction of the plane's flight path, a straight line directly down from the plane, or any trajectory in the opposite direction of the plane's flight path) were coded as incorrect (see the Supplemental Material for examples). Two or three coders coded each response, and discrepancies were resolved by discussion.

Gender. Participants reported their gender as male, female, gender diverse ("divers" in German), or "prefer not to answer."

Age. Participants reported their age in years.

General education. Participants reported their highest level of education on a 7-point ordinal scale (1 = *no degree*, 7 = *doctorate*), plus an "other" option with space to write in a response.

Physics education. Participants reported the total number of years of physics education received.

Major. University students wrote in their major and selected the type of program to which their major belonged (*natural sciences, engineering*, *medicine*, *social science*, *arts and humanities*, *or other/don't know*).

Previous task experience. Participants reported whether they had prior familiarity with each intuitive physics task by circling *yes* or *no* for each task.

Occupational experience. When collecting data, researchers recorded whether participants were bartenders, servers, bus drivers, or students. Participants also reported their current primary occupation, checking a box for *bartender*, *server*, *bus driver*, *student*, or *other* (with space to write in a response). Participants then reported years of experience in their current occupation, as well as years worked in any beverage-service role.

See Table 1 for a summary of the descriptive statistics of the sample.

Procedure

All participants were recruited in Germany, and all data collection occurred in person (similar to Hecht and Proffitt, 1995). Like Hecht and Proffitt's study, participants were tested in their workplace or at their university, and they took as much time as needed to complete the task. Participants read the instructions in the questionnaire themselves and completed the paper-and-pencil water-level task, followed by a filler task and, finally, the falling-object task. After each key task (i.e., the water-level task, the falling-object task), participants reported whether they had prior experience with the task. Last, they answered questions about their gender, age, education, and occupation.

Researchers recruited servers during Oktoberfest in Munich, Germany. Servers were recruited before their shift began or during breaks, bartenders before their shift or when they were not busy, and bus drivers at the end of bus lines or at rest stops; students were recruited from a university in the western part of Germany. Nonstudent participants received financial compensation for completing the study (a gift card to Amazon.de worth €10), and students received a chocolate bar as compensation (commensurate with norms for completing psychology studies at this university).

Here we summarize the procedural differences between our replication and the Hecht and Proffitt study (1995). First, we did not recruit housewives as one of the subpopulations comprising the inexperienced group. Second, we had a team of researchers administering the survey who did not automatically provide verbal instructions, whereas Hecht and Proffitt gave instructions verbally (in addition to identical written instructions). Third, we provided payment to participants who completed the study, whereas Hecht and Proffitt did not. Fourth, for our sample of university students, we administered surveys to graduate and undergraduate students from various programs of study, whereas Hecht and Proffitt recruited only graduate students (with half enrolled in social-science programs and half in natural-science programs). Finally, we did not recruit or select on the basis of gender for any occupation. (We do not know whether Hecht and Proffitt actively selected on the basis of gender, but they reported data from only female servers, male bartenders, and male bus drivers.) This research was reviewed by, and received approval from, the Institutional Review Board at the University of Utah (ID No. IRB_00174435) and met the ethical guidelines and legal requirements of the Ruhr-University Bochum, Germany.

Results

Experience and performance. We first examined whether participants with occupational experience performed differently on the water-level task compared with participants with no occupational experience. To assess this, we used absolute error (in degrees) as our dependent variable and conducted a two-tailed, two-sample t test using robust standard errors (i.e., assuming unequal variances). With our final sample size of 370 participants (170 inexperienced and 200 experienced participants), we had 80% power to detect an effect of $d \geq 0.29$ using a two-tailed t test and an alpha level of .05 and 90% power to detect an effect of $d \geq 0.34$. As a point of comparison, our sample size gave us more than 99% power to detect the original effect size of $d = 0.67$ observed by Hecht and Proffitt (calculated from data provided by Hecht; personal communication).

We failed to observe a significant difference in absolute error between experienced participants ($M = 9.19$, $SD = 12.62$) and inexperienced participants ($M = 9.46$, $SD = 11.40$), $t(366.62) = -0.22$, $p = .83$, $d = -0.02$. When using the same binary cutoff for performance used in Hecht and Proffitt (0 = more than five degrees [of] error, 1 = five degrees or less of error), there was no significant difference in performance between experienced participants (59.0% correct) and inexperienced participants (51.2% correct), $z = -1.51$, $p = .13$. We also note that our results are directionally opposite to those of Hecht and Proffitt.[5] See Figure 2.

We also tested for performance differences between groups after statistically adjusting for gender, age, and education. Using ordinary-least-squares regression, we regressed absolute error scores onto experience (0 = *inexperienced*, 1 = *experienced*) as well as gender (dummy-coded), age (in years), and education (dummy-coded). For this regression, as well as all others, we implemented robust standard errors. We again failed to find a statistically significant difference in absolute error between experienced participants (predicted $M = 9.00$) and inexperienced participants (predicted $M = 10.0$), $t(344) = -0.64$, $p = .52$. When using the same binary cutoff for performance as before,[6] we again failed to find a significant difference in correct responses between experienced participants (predicted

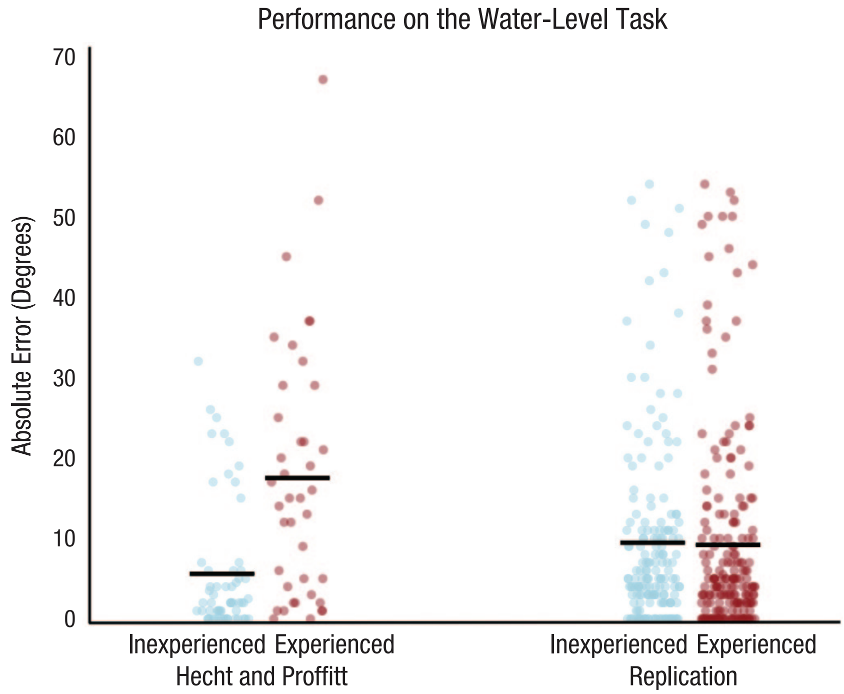

Fig. 2. Performance on the water-level task for those with occupational experience handling liquid (servers at Oktoberfest and bartenders) and those without (students and bus drivers) in Hecht and Proffitt (1995) and the current replication. Horizontal lines represent means. Higher values indicate worse performance.

probability = 59.5%) and inexperienced participants (predicted probability = 50.8%), $z = -1.42$, $p = .16$.

Replicability. We assessed the replicability of Hecht and Proffitt (1995) using the small-telescopes criterion, which asks whether our observed effect is large enough to have been detectable at 33% power based on the original sample size from Hecht and Proffitt (Simonsohn, 2015). Relying on this criterion,[7,8] an effect size reliably smaller than $d = 0.30$ would be inconsistent with a true effect large enough to have been detectable by Hecht and Proffitt; thus, we consider this a failed replication.[9] Using a one-sided t test, the difference we observed between experienced and inexperienced participants was reliably smaller than a detectable effect, $t(368) = 3.11$, $p = .001$. We observed a similar result after statistically adjusting for participant gender, age, and education, $t(344) = 2.90$, $p = .002$. We failed to replicate the results of Hecht and Proffitt.

Extensions. On the falling-object task, experienced participants were less likely to answer correctly (31.3%) than inexperienced participants (42.0%), $z = 2.13$, $p = .03$. This difference becomes statistically nonsignificant after adjusting for participant gender, age, and education (predicted probabilities were 31.7% vs. 40.8%, respectively; $z = 1.57$, $p = .12$).

We next examined group differences in performance across the two tasks (water level vs. falling object). First, we dichotomized performance on the water-level task similar to before (and similar to in Hecht & Proffitt, 1995) in order to compare performance across the two tasks. We then performed a logit regression in which we regressed task performance (0 = *incorrect answer*, 1 = *correct answer*) onto our predictors of experience (0 = *inexperienced*, 1 = *experienced*), task (0 = *water level*, 1 = *falling object*), and the interaction between experience and task. We implemented participant-clustered standard errors to account for potential nonindependence in performance across tasks. According to our Stage 1 preregistration, our coefficient of interest is the interaction based on the difference in average marginal effects (rather than the interaction term from the logit model; see McCabe et al., 2022).

On the basis of Hecht and Proffitt's (1995) original hypothesis we should expect a positive interaction effect, which would imply a larger detrimental effect of beverage experience on the water-level task than on the falling-object task. In fact, we observed a statistically significant negative interaction, $b = -0.19$, $SE = 0.06$, $z = -2.90$, $p = .004$. As discussed above, experienced participants (nonsignificantly) outperformed inexperienced participants on the water-level task, $z = -1.51$, $p = .131$, but performed worse than inexperienced participants on the falling-object task, $z = 2.13$, $p = .034$. We observed a similar negative interaction when adjusting for participant demographics, $b = -0.17$, $SE = 0.06$, $z = -2.62$, $p = .009$.

Exploratory analyses and data-quality checks. Hecht and Proffitt (1995) and Vasta et al. (1997) reported finding that men outperform women (also see Robert, 1990, and Tran & Formann, 2008; cf. Wu et al., 2017). Researchers have also found that participants who have more years of education, especially physics education, perform better on the water-level task (Riener et al., 2005). Hecht and Proffitt reported that younger participants performed best, but a well-powered study examining age found no decline in performance until around age 60 (Tran & Formann, 2008), which represents less than 1.5% of the participants in our sample. As an exploratory exercise and data-quality check, we examined whether younger participants, male participants, and more educated participants performed better on the water-level task than others (we also provide a correlation table between all variables in the Supplemental Material; see Table S2).

Absolute error on the water-level task was smaller for male participants ($M = 8.51$, $SD = 12.28$) than for female participants ($M = 10.30$, $SD = 12.12$), though the difference was not statistically significant, $t(348.5) = 1.38$, $p = .17$, $d = 0.15$. We observed a weak and nonsignificant positive correlation between age and absolute error (Pearson's $r = .096$, $p = .07$), and this relationship shrinks to virtually zero when we examined the rank-order correlation between age and absolute error (Spearman's $\rho = .004$, $p = .93$). We observed a negative and nonsignificant rank-order correlation between educational level[10] and absolute error on the water-level task (Spearman's $\rho = -.065$, $p = .22$). Finally, we observed a negative and significant correlation between years of physics education and absolute error on the water-level task (Pearson's $r = -.109$, $p = .04$; Spearman's $\rho = -.193$, $p < .001$). In sum, the only demographic characteristic reliably related to superior performance on the water-level task was years of physics education.

Last, Hecht and Proffitt (1995) reported that only 3% of participants drew a line that was less than −5 degrees from horizontal. We found that 8.1% of participants in our sample made this type of error.

Discussion

We conducted a registered report to replicate Hecht and Proffitt (1995), examining the relationship between occupational experience and performance on an occupationally relevant intuitive physics task. We extended their study by adding an intuitive physics task unrelated to occupational experience, as a further check of their theory. As in Hecht and Proffitt, we recruited participants with occupational experience handling liquids (servers and bartenders) and without this experience (students and bus drivers). We compared performance on the classic water-level task (Piaget & Inhelder, 1948/1956), and, as an extension, on an intuitive physics task unrelated to handling liquid (the falling-object task, McCloskey, 1983). Unlike Hecht and Proffitt, we did not find that participants with occupational experience performed worse on the water-level task. We observed a relatively precisely estimated null effect between groups.

Although we found no meaningful difference between the two groups on our target intuitive physics task (the water-level task), we did find that experienced participants performed relatively worse on an alternative intuitive physics task (the falling-object task). One possibility is that experienced participants in our sample show a baseline performance deficit on intuitive physics tasks relative to inexperienced participants and that their occupational experience affords a performance boost on the water-level task specifically. This interpretation would be consistent with Vasta et al.'s (1997) finding that participants with occupational experience outperformed inexperienced participants on the water-level task. However, this explanation raises the question of why occupational experience with liquids would be associated with a performance deficit on nonliquid intuitive physics tasks. Higher education attainment in the inexperienced group, which was composed primarily of undergraduate and graduate students, is unlikely to account for this discrepancy, as education did not reliably predict performance on intuitive physics tasks in our study, and the two groups had similar levels of physics education. We also note that this interaction became nonsignificant when we filtered on our secondary set of excluded participants. Therefore, given the total available evidence, we believe that the most parsimonious explanation of these results is that there is likely no meaningful difference between experienced and inexperienced groups on the water-level task.

Strengths of our replication include a larger sample, and thus more statistical power, than that used by Hecht and Proffitt (1995). Our sample size allowed us to detect small to medium effects with a high degree of power and gave us over 99% power to detect effect sizes reported in the original article by Hecht and Proffitt. Another strength of our design was the use of the registered-report method, which provides readers with evidence of decisions we made before data collection (e.g., about the sample and the analysis plans). We also collaborated with a member of the original authorship team to recreate their materials and closely follow their procedure, and a member of our authorship team had extensive prior experience coding the water-level task. Finally, we recruited participants in the same location and setting as the original study, unlike many replication studies (Hoffmann et al., 2025).

Despite these strengths, there are several reasons why we may have failed to replicate the results of Hecht and Proffitt (1995). One possibility is that the original study was a false positive; another possibility is that our replication is a false negative. Although our study had considerably higher power than that of Hecht and Proffitt, the true effect size may have been smaller than we had sufficient statistical power to detect. It is also possible that although we returned to the original location, participant characteristics could have meaningfully changed over time (e.g., perhaps our servers had fewer years of work experience and were younger), or perhaps situational characteristics were meaningfully different (e.g., perhaps our servers were less distracted when completing the survey). From our field observations, servers appeared more concerned with getting drinks to a table quickly—to put them down because they were heavy, and to get them to customers efficiently—than with spilling beverages. This tendency could have been different when Hecht and Proffitt studied servers, but it nevertheless calls into question the hypothesis that occupational experience with liquids elicits an object-relative reference system.

Transparency

Action Editor: Yoel Inbar
Editor: Simine Vazire
Author Contributions

Elizabeth R. Tenney: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Kylie Rochford: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing.

Amelia Stillwell: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Coco Xinyue Liu: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.

David Tannenbaum: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Marie Hennecke: Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing – review & editing.

Jeanine K. Stefanucci: Data curation, Methodology, Resources, Supervision, Writing – review & editing.

B. Ariel Blair: Funding acquisition, Investigation, Methodology, Resources, Writing – review & editing.

Jesse Graham: Funding acquisition, Investigation, Methodology, Writing – review & editing.

Bryan L. Bonner: Data curation, Funding acquisition, Investigation, Methodology, Writing – review & editing.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was financially supported by a grant from the Marriner S. Eccles Institute for Economics and Quantitative Analysis at the University of Utah, awarded to the first author.

Artificial Intelligence

ChatGPT and Copilot were used as intelligent thesaurus and copyediting tools during the creation of this article. No other artificial-intelligence-assisted technologies were used in this research or the creation of this article.

Ethics

This research received approval from the University of Utah Ethics Board (ID No. IRB_00174435) and met the ethical guidelines and legal requirements of Ruhr-University Bochum, Germany. We obtained informed consent from all participants.

Open Practices

Open practices for this article are described in the Research Transparency Statement section, which appears at the end of the Introduction section in the main text.

ORCID iDs

Elizabeth R. Tenney https://orcid.org/0009-0009-5327-7947
Kylie Rochford https://orcid.org/0000-0002-5707-2510
Amelia Stillwell https://orcid.org/0000-0001-9757-5326
Coco Xinyue Liu https://orcid.org/0009-0001-0890-6435
David Tannenbaum https://orcid.org/0000-0002-6603-7370
Marie Hennecke https://orcid.org/0000-0002-0263-4598
Jeanine K. Stefanucci https://orcid.org/0000-0003-4238-2951
B. Ariel Blair https://orcid.org/0000-0002-0414-9094
Jesse Graham https://orcid.org/0000-0001-8863-7978
Bryan L. Bonner https://orcid.org/0000-0002-7294-5789

Acknowledgments

We thank Matthew Aita, Zeynep Börek, Jacqueline M. Chen, Rebekka Daubert, Iara Delgado, Kyle Hudson Guttman, Alex Hall, Johannes Koopmann, Rebekah Obst, Johanna Röttsches, and Laura Schächtele for their insightful comments and research assistance, and we thank Heiko Hecht and Dennis Proffitt for sharing their wisdom and the data file from their original study.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976251412721

Notes

1. We chose not to recruit the housewives group used in Hecht and Proffitt (one of the subsamples that comprised the inexperienced group) for several reasons. First, the original group of housewives was a snowball sample from the mother of one of the authors (personal communication with Hecht, February 21, 2024). Consequently, it was unclear how to recruit a comparable sample of housewives to that of the original study. Second, we were concerned that the occupation of housewife or homemaker may not clearly represent low experience, as this role could include frequent tasks that involve carrying liquids, such as serving drinks within the home. Indeed, Hecht and Proffitt's data show a considerably smaller difference in performance between experienced participants and housewives (absolute error on the water-level task: d = 0.06) than between experienced participants and all other inexperienced participants (d = 1.01). Thus, focusing on the latter set of inexperienced participants (students and bus drivers) gave us the greatest chance of replicating Hecht and Proffitt's primary result.
2. To minimize collusion among participants, our research team sometimes asked participants not to speak to or discuss answers with others while completing the survey. Most participants were recruited directly from a researcher, and the researcher was almost always nearby while the participant completed the survey, ready to discourage collusion. However, given the noisy environment for data collection, we cannot be certain that no collusion occurred.
3. We estimated that we could recruit 30 bus drivers total given our ongoing efforts, and therefore we collected data from 170 students while we continued to recruit bus drivers. However, reaching this target proved difficult, so we stopped at 20 bus drivers and collected data from 10 additional students.
4. All results reported here above or below the significance threshold of $p < .05$ did not change when we used our secondary set of exclusions, except for two findings. First, we examined group differences in performance across the two tasks (water level vs. falling object) and found a negative interaction when we adjusted for participant demographics (b = –0.17, SE = 0.06, z = –2.65, p = .008); this result failed to reach statistical significance when we filtered on our secondary set of excluded participants (b = –0.16, SE = 0.08, z = –1.90, p = .058). Second, in the exploratory section, we reported a nonsignificant correlation between age and absolute error on the water-level task (Pearson's r = .096, p = .07); this result became statistically significant when we filtered out the additional set of excluded participants (Pearson's r = .142, p = .02).
5. For all analyses, test statistics and effect sizes were coded as negative when they were inconsistent with Hecht and Proffitt (1995).
6. When statistically adjusting for demographics for binary outcomes, we conducted the same set of analyses as before but used logit regression rather than ordinary-least-squares regression. We report estimates, test statistics, and p values based on the average marginal effects (i.e., the difference in predicted probabilities) rather than based on the log-odds coefficient from the logit model (for a discussion on this issue, see McCabe et al., 2022). We note that in our data both approaches tend to return similar test statistics and p values.
7. We used Hecht and Proffitt's (1995) total sample (including housewives) when performing our small-telescopes calculation, even though our sample did not include housewives. Doing so created a more stringent or conservative criteria for us to conclude a failed replication result.
8. In our Stage 1 preregistration, we had incorrectly reported this value as d = 0.28. Using either effect size does not change our results or conclusions.
9. Another method for assessing replicability is the use of prediction intervals (Spence & Stanley, 2024). Given the observed effect and sample size found in Hecht and Proffitt (1995) and our replication sample size, any standardized effect falling outside the prediction interval of [0.23, 1.11] would indicate a failed replication. Using this method, we again failed to replicate the results of Hecht and Proffitt: Our observed effect size of d = –0.02 fell outside the replication interval.
10. For correlations with education, we excluded 1 additional respondent who reported "other" as the degree of educational attainment. This participant was not dropped from the prior regression analyses, because educational attainment was included as a fixed effect (i.e., dummy-coded), which does not assume a linear relationship across education levels and the outcome variable.

References

Barhorst-Cates, E. M., Creem-Regehr, S. H., Stefanucci, J. K., Gardner, J., Saccomano, T., & Wright, C. (2020). Spatial reference frame but neither age nor gender predict performance on a water-level task in 8- to 11-year-old children. *Perception*, *49*(11), 1200–1212. https://doi.org/10.1177/0301006620964414

Bilalić, M., McLeod, P., & Gobet, F. (2008). Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psychology*, *56*(2), 73–102. https://doi.org/10.1016/j.cogpsych.2007.02.001

Brislin, R. W. (1976). Comparative research methodology: Cross-cultural studies. *International Journal of Psychology*, *11*(3), 215–229. https://doi.org/10.1080/00207597608247359

Dane, E. (2010). Reconsidering the trade-off between expertise and flexibility: A cognitive entrenchment perspective. *Academy of Management Review*, *35*(4), 579–603. https://doi.org/10.5465/amr.35.4.zok579

Dror, I. E., Pascual-Leone, A., & Ramachandran, V. (2011). The paradox of human expertise: Why experts get it wrong. In N. Kapur (Ed.), *The paradoxical brain* (pp. 177–188). Cambridge University Press.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*, 273–305. https://doi.org/10.1146/annurev.psych.47.1.273

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, *5*(8), Article 180448. https://doi.org/10.1098/rsos.180448

Hecht, H., & Proffitt, D. R. (1995). The price of expertise: Effects of experience on the water-level task. *Psychological Science*, *6*(2), 90–95. https://doi.org/10.1111/j.1467-9280.1995.tb00312.x

Hoffmann, J., Twardawski, M., Höhs, J. M., Gast, A., Pohl, S., & Sengewald, M. (2025). The design of current replication studies: A systematic literature review on the variation of study characteristics. *Advances in Methods and Practices in Psychological Science*, *8*(2), 1–22. https://doi.org/10.1177/25152459251328273

Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., Krueger, J. I., Perugini, M., Ropovik, I., van't Veer, A. E., Vranka, M., & Lakens, D. (2023). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods*, *28*(2), 438–451. https://doi.org/10.1037/met0000438

Kilgour, M., & Koslow, S. (2009). Why and how do creative thinking techniques work?: Trading off originality and appropriateness to make more creative advertising. *Journal of the Academy of Marketing Science*, *37*, 298–309. https://doi.org/10.1007/s11747-009-0133-5

Matthes, J., Schneider, M., & Preckel, F. (2024). The relation between prior knowledge and learning in regular and gifted classes: A multigroup latent growth curve analysis. *Journal of Educational Psychology*, *116*(2), 278–296. https://doi.org/10.1037/edu0000848

McAfee, E. A., & Proffitt, D. R. (1991). Understanding the surface orientation of liquids. *Cognitive Psychology*, *23*(3), 483–514. https://doi.org/10.1016/0010-0285(91)90017-I

McCabe, C. J., Halvorson, M. A., King, K. M., Cao, X., & Kim, D. S. (2022). Interpreting interaction effects in generalized linear models of nonlinear probabilities and counts. *Multivariate Behavioral Research*, *57*(2–3), 243–263. https://doi.org/10.1080/00273171.2020.1868966

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324). Erlbaum.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Piaget, J., & Inhelder, B. (1956). *The child's conception of space* (F. J. Langdon & J. L. Lunzer, Trans.). Routledge & Kegan-Paul. (Original work published 1948)

Pittelkow, M. M., Field, S. M., Isager, P. M., van't Veer, A. E., Anderson, T., Cole, S. N., Dominik, T., Giner-Sorolla, R., Gok, S., Heyman, T., Jekel, M., Luke, T. J., Mitchell, D. B., Peels, R., Pendrous, R., Sarrazin, S., Schauer, J. M., Specker, E., Tran, U. S., Vranka, M. A., Wicherts, J. M., Yoshimura, N., Zwaan, R. A., & Van Ravenzwaaij, D. (2023). The process of replication target selection in psychology: What to consider? *Royal Society Open Science*, *10*(2), Article 210586. https://doi.org/10.1098/rsos.210586

Riener, C., Proffitt, D. R., & Salthouse, T. (2005). A psychometric approach to intuitive physics. *Psychonomic Bulletin & Review*, *12*(4), 740–745. https://doi.org/10.3758/BF03196766

Robert, M. (1990). Sex-typing of the water-level task: There is more than meets the eye. *International Journal of Psychology*, *25*(2), 475–490. https://doi.org/10.1080/00207599008247878

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Spence, J. R., & Stanley, D. J. (2024). Tempered expectations: A tutorial for calculating and interpreting prediction intervals in the context of replications. *Advances in Methods and Practices in Psychological Science*, *7*(1), 1–13. https://doi.org/10.1177/25152459231217932

Tran, U. S., & Formann, A. K. (2008). Piaget's water-level tasks: Performance across the lifespan with emphasis on the elderly. *Personality and Individual Differences*, *45*, 232–237. https://doi.org/10.1016/j.paid.2008.04.004

Vasta, R., Rosenberg, D., Knott, J. A., & Gaze, C. E. (1997). Experience and the water-level task revisited: Does expertise exact a price? *Psychological Science*, *8*(4), 336–339. https://doi.org/10.1111/j.1467-9280.1997.tb00449.x

Wu, S., Li, Y., & Kong, M. (2017). Sex and ability differences in neural strategy for Piaget's water level test: An EEG study. *Perceptual and Motor Skills*, *124*(2), 351–365. https://doi.org/10.1177/0031512516687902

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, Article e120. https://doi.org/10.1017/S0140525X17001972