Supplementary Analysis and Tables

Results with Secondary Exclusions

Experience and Performance. We first examined whether participants with occupational experience performed differently on the water level task compared to participants with no occupational experience. To examine 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). Based on our final sample size of 285 participants (162 inexperienced and 123 experienced participants), we have 80% power to detect an effect of $d \geq 0.34$ using a two-tailed t-test and an alpha level of 0.05, and 90% power to detect an effect of $d \geq 0.39$. As a point of comparison, our sample size gives 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 ($M = 9.41$, $SD = 12.82$) and inexperienced participants ($M = 9.31$, $SD = 11.16$), $t(242.0) = –0.07$, $p = .95$, $d = –0.01$. When using the same binary cutoff for performance used in Hecht and Proffitt (0 = more than five degrees error, 1 = five degrees or less of error), experienced participants were not significantly less likely to answer correctly (56.1%) than inexperienced participants (51.9%), z = –0.71, $p = .48$. We also note that our results are directionally opposite to that found by Hecht and Proffitt.[1]

We also tested for performance differences between groups after statistically adjusting for gender, age, and education. Using ordinary least squares, 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 fail to find a statistically significant

[1] For all analyses, test statistics and effect sizes are coded as negative when they are inconsistent with Hecht and Proffitt (1995).

difference in absolute error between experienced participants (predicted marginal *M* = 8.66) and inexperienced participants (predicted marginal *M* = 10.19), *t*(261) = –0.77, *p* = .44. When using the same binary cutoff for performance as before,[2] we again failed to find a significant difference in correct responses between experienced participants (predicted marginal probability = 56.9%) and inexperienced participants (predicted marginal probability = 50.6%), *z* = –0.86, *p* = .39.

Replicability. We assess 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). Based on this criterion,[3,4] 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 (and thus we consider a "failed" replication).[5] Using a one-sided *t*-test, the difference we observe between experienced and inexperienced participants was reliably smaller than a detectable effect, *t*(283) = 2.39, *p* = .009. We observed a similar result after statistically adjusting for participant gender, age, and education, *t*(261) = 2.56, *p* = .006. We fail to replicate the results of Hecht and Proffitt.

[2] When statistically adjusting for demographics for binary outcomes, we conducted the same set of analyses as before but using logit regression rather than OLS regression. We report test statistics and *p*-values based on the average marginal effects (i.e., 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 both approaches tend to return very similar test statistics and *p*-values.

[3] We use Hecht and Profitt's (1995) total sample (including housewives) when performing our small telescopes calculation, even though our sample did not include housewives. Doing so creates a more stringent or conservative criteria for us to conclude a failed replication result.

[4] 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.

[5] Another method for assessing replicability is the use of prediction intervals (Spence & Stanley, 2024). Given the observed effect size and sample size found in Hecht and Proffitt (1995) and our replication sample size, any standardized effect falling outside the prediction interval [0.21, 1.13] would indicate a "failed" replication. Using this method, we again fail to replicate the results of Hecht and Proffitt — our observed effect size of –0.01 falls outside the replication interval.

Extensions. On the falling object task,[6] experienced participants were less likely to answer correctly (28.1%) than inexperienced participants (39.4%), but this difference was not statistically significant, $z = 1.95$, $p = .051$. This difference is also statistically nonsignificant after adjusting for participant gender, age, and education (predicted marginal probabilities were 30.2% versus 37.5%, respectively; $z = 1.06$, $p = .290$).

We next examined group differences in performance across the two tasks (water level versus 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. Per our stage 1 preregistration, our coefficient of interest is the interaction based on the difference in predicted probabilities (rather than the interaction term from the logit model; see McCabe et al., 2022).

Based on 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 observe a statistically significant *negative* interaction, $b = –0.16$, SE = 0.08, $z = –2.09$, $p = .036$. As discussed above, experienced participants (nonsignificantly) outperformed inexperienced participants on the water level task, $z = –0.77$, $p = .442$, but performed worse than inexperienced participants on the falling object task, $z = 1.98$, $p = .047$. When adjusting for participant

[6] For analyses examining performance on the falling object task, we also exclude 7 participants who reported prior knowledge of the falling object task.

demographics, the negative interaction is nonsignificant, *b* = –0.14, SE = 0.08, z = –1.77, *p* = .076.

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; 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 relatively higher on the water level task.

Absolute error on the water level task was smaller for male participants (*M* = 8.57, *SD* = 12.56) than for female participants (*M* = 10.28, *SD* = 11.61), though the difference was not statistically significant, *t*(247.7) = 1.16, *p* = .25, *d* = 0.14. We observed a small positive correlation between age and absolute error (Pearson's *r* = 0.142, *p* = .02), but a small and nonsignificant rank-order correlation between age and absolute error (Spearman's ρ = 0.042, *p* = .49) suggesting that the first correlation was likely driven by outliers on the water level task. We observe a negative and nonsignificant rank-order correlation between educational level[7] and absolute error on the water level task (Spearman's ρ = –0.059, *p* = .33). Finally, we observe a negative and significant correlation between years of physics education and absolute error on the water level task (Pearson's *r* = –0.124, *p* = .04; Spearman's ρ = –0.166,

[7] For correlations with education, we exclude 1 additional respondent who reported "other" as their degree of educational attainment. This participant is not dropped from the prior regression analyses, because it was included as a fixed effect (i.e., dummy-coded), which does not assume an ordinal relationship between education levels and the outcome variable.

p = .006). In sum, the only demographic characteristic reliably related to superior performance on the water level task was years of physics education.

Lastly, Hecht and Proffitt (1995) reported that only 3% of participants drew a line that was less than –5 degrees from horizontal. We found that 7.4% of participants in our sample made this type of error.

Table S1 Preregistration deviations.

#	Details		Original wording	Deviation description	Extent of deviation	Judgement of impact
1	Study	#1 of 1	We said that if by the end of Oktoberfest we do not have 100 servers, we will collect more data from bartenders to reach 200 experienced participants.	During data collection we realized that we would be unable to collect responses from 100 bartenders by the end of Oktoberfest, as it turned out to be harder to reach bartenders than servers. We instead increased the number of servers in our sample to reach our target of 200 experienced participants (i.e., 60 bartenders and 147 servers).	*Minor*	We do not believe this change affected the results or changed the risk of bias.
	Type	Methods				
	Reason	Plan not possible				
	Timing	During data collection				
2	Study	#1 of 1	We said that if by the end of Oktoberfest we do not have 100 bus drivers we will collect more data from students to reach 200 inexperienced participants.	We were unable to collect responses from 100 bus drivers by the end of Oktoberfest, only getting 6 responses, as the rest stop we had planned for recruitment was closed for renovation. Furthermore, many bus drivers we approached declined to participate and/or did not speak German (the language of our survey). For this reason, in consultation with the action editor, we decided to continue recruiting bus drivers in another city in Germany after Oktoberfest ended. We estimated that we could recruit 30 bus drivers total given our ongoing efforts, and therefore we	*Minor*	In Hecht & Proffitt (1995) and in our study, students and bus drivers performed similarly on the water level task. They also performed similarly on the falling object task. As a result, even though there were a large number of students, we do not believe this change affected the results or changed the risk of bias.
	Type	Methods				
	Reason	Plan not possible				
	Timing	During data collection				

				collected data from 170 students while continuing to recruit bus drivers. However, reaching even this target proved extremely difficult, so we stopped at 20 and collected data from 10 additional students. Thus, to reach our target of 200 inexperienced participants, we recruited 20 bus drivers (6 during Oktoberfest and 14 after) and 180 students total.		
	Study	#1 of 1	“We will stop on the respective day(s) that the sample size(s) for experience and inexperienced is reached.”	We stopped data collection when we hit 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). We did not keep going until the end of the day because we had good records of exactly when we reached our targets and had good communication between researchers.	*Minor*	We do not believe this change affected the results or changed the risk of bias.
	Type	Methods				
	Reason	New knowledge				
	Timing	During data collection				
4	Study	#1 of 1	“Two experimenters will code each response”	One research assistant doing the coding was having health issues and so we recruited a third data coder.	*Minor*	We do not believe this change affected the results or changed the risk of bias. It may have increased the precision of the scoring.
	Type	Methods				
	Reason	New knowledge				
	Timing	During data collection				
5	Study	#1 of 1	“Two experimenters will code each response”	Only after coding all responses for absolute error on the water level task did we realize that we needed to document the direction of the error for one analysis. We decided to single-code the	*Minor*	We do not believe this change affected the results or changed the risk of bias.
	Type	Methods				

	Reason	Miscommunication		direction as positive/negative/no slope, with spot checks for each data coder. No mistakes were found during spot checks.		
	Timing	During data collection				
6	Study	#1 of 1	“discrepancies greater than 1 degree will be resolved by discussion.”	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 degree were resolved by discussion. We made these changes to accommodate discrepancies less than 1 degree and to adjust for having a third coder.	*Minor*	We do not believe this change to our coding method affected the results or changed the risk of bias.
	Type	Methods				
	Reason	Plan not possible				
	Timing	During data collection				
7	Study	#1 of 1	We would test participants “individually” in their place of work.	The first 5 out of 8 participants we ran in the field in one location drew a line outside of the glass, which is an unusual response. We speculated that these participants talked to each other about the survey. After this, we monitored for collusion more closely, and sometimes we asked participants not to speak to others while taking the survey and not to discuss answers. Most participants were recruited directly from a researcher, and the researcher was almost always nearby while a participant completed the survey, ready to discourage collusion.	*Minor*	We do not believe that the few participants we suspected of colluding affected the results or changed the risk of bias in a meaningful way (drawing a line outside of the glass means we could not score their data, so they were excluded). If many participants colluded, this would increase bias in the sample, but we do not believe that this happened often enough to affect the results given the changes we made to our protocol early on. That said, it was a noisy environment to collect data, and we cannot be certain no collusion occurred.
	Type	Methods				
	Reason	Plan not possible				
	Timing	During data collection				

8	Study	#1 of 1	"an effect size reliably smaller than d = 0.28 would be inconsistent with a true effect large enough to have been detectable by Hecht and Proffitt (and thus we consider a "failed" replication)"	When we checked this effect size later using a sensitivity test in gpower, it turns out that the effect size should be d = 0.30. We are not sure why we first thought it was 0.28.	*Minor*	We do not believe this change affected the results or changed the risk of bias. Using either effect size does not change our results or conclusions.
	Type	Analyses				
	Reason	New knowledge				
	Timing	After data access				
9	Study	#1 of 1	In trying to explain how we would interpret various findings, in the paragraph called, "Current Study," we used language such as saying: "We can assess whether an effect that we obtain is most consistent with Hecht and Proffitt, Vasta et al. (1997), or a null hypothesis of no difference."	Upon reflection, we were concerned that this language implied that we would formally test whether our results were consistent with Vasta et al. (1997), when we did not intend to do that per our analysis plan and what we told reviewers in our response letter. Therefore, we cut three sentences from this paragraph to avoid confusion. From the response letter: "Since we intentionally stick closely to the design of the original study by Hecht and Proffitt, we don't believe it would be appropriate to conduct a third analysis that performs a small-telescopes test based on the findings of Vasta et al. (as our study is not meant to be a direct replication of their work)."	*Minor*	We do not believe this change affected the results or changed the risk of bias.
	Type	Interpretation				
	Reason	Miscommunication				
	Timing	After results known				
10	Study	#1 of 1	We sometimes made decisions that were not especially concise or clear (e.g. wording).	We made changes that we thought improved clarity and/or conciseness that we felt did not alter the spirit of any stage 1 decisions. For example, since we did not randomly assign participants to a profession, we now refer to "groups" rather than "conditions," and we use "researchers" or "data coders" instead of "experimenters." Another	*Minor*	We do not believe this change affected the results or changed the risk of bias.
	Type	Interpretation				
	Reason	Miscommunication				

	Timing	After results known		example of a wording change is we “limit” experimenter demand instead of “control” it. We also changed the wording “predicted probabilities derived from the model” to “difference in average marginal effects” which mean the same thing, but reflected our updated preference for how to describe it. We also clarified that occupational experience was noted by experimenters as they collected the data. As another example of a change, we combined the two figures illustrating intuitive physics tasks into one figure with two panels. We also rearranged the demographic questions to be after the descriptions of the intuitive physics tasks to align with the order that participants saw them in the survey, and we moved the section on exclusions to be near the section about sample size.		

Note. Adapted from Willroth & Atherton, 2023

Table S2

Pearson correlations (r)

Variable	1	2	3	4	5	6	7	8
1. Occupational experience[a]								
2. Gender[b]	-.12*							
3. Age[c]	.36**	-.24**						
4. Physics[c]	.10	-.10	.01					
5. Bev service[c]	.45**	-.06	.71**	.01				
6. Education[d]	-.20**	.13*	-.33**	.17**	-.17**			
7. WLT error[e]	-.01	.07	.10	-.11*	.13*	-.12*		
8. WLT correct[f]	.08	-.08	-.04	.13*	-.08	.08	-.67**	
9. FOT correct[g]	-.11*	-.34**	-.08	.18**	-.16*	.15**	-.21**	.22**

Notes: The sample size of possible responses is 370 after excluding 30 participants who reported some degree of beverage service experience in the inexperienced group and then excluding 7 participants who failed to draw a line in the glass. Sample size per cell may vary because of missing values. * $p < .05$, ** $p < .01$.

a. 0 = inexperienced (students and bus drivers), 1 = experienced (servers and bar tenders)
b. Men = 1, women = 2; other gender categories excluded
c. Age, physics (physics education), and bev service (beverage service industry experience) are reported in years.
d. Treated as a scale from 1 to 7, excluding 1 additional participant who selected "other."
e. Mean absolute error in degrees on the water level task
f. Performance on the water level task: incorrect = 0, correct = 1
g. Performance on the falling object task: incorrect = 0, correct = 1