Supplementary Information for

How expertise mediates the effects of numerical and textual communication on individual and collective accuracy

Nicholas Beauchamp, Sarah Shugars, Briony Swire-Thompson, and David Lazer

Corresponding Author: Nicholas Beauchamp (n.beauchamp@northeastern.edu)

This PDF file includes:

Figs. S1 to S2
Tables S1 to S6
SI References

List of Tables

List of Figures

Effects on standardized individual prediction values

Table S1 shows the results from OLS regression of standardized prediction values on the specified subject features (Models 1-3) and a multilevel regression with random coefficients for subjects and questions (Model 4). Models 5-6 also include expertise interacted with the anchor values, while models 7-8 include expertise interacted with reasoning skill. Note that these interactions are significant for expertise but not reasoning skill.

Table S1. Effects on standardized predictions values.

	Dependent variable:							
	Raw Prediction Value							
	OLS			*linear mixed-effects*	*OLS*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Anchor Value	0.326***	0.247***	0.266***	0.278***	0.172***	0.238***	0.258***	0.227***
	(0.027)	(0.041)	(0.047)	(0.047)	(0.053)	(0.061)	(0.066)	(0.080)
Anchor^2	−0.004	−0.004	0.002	−0.005	−0.017	0.038	−0.018	0.059
	(0.015)	(0.015)	(0.019)	(0.020)	(0.023)	(0.037)	(0.021)	(0.048)
Anchor^3	−0.020**	−0.019**	−0.012	−0.011	0.011	−0.024	−0.010	0.017
	(0.008)	(0.008)	(0.010)	(0.010)	(0.013)	(0.021)	(0.011)	(0.025)
Question Order		0.007**	0.007**	0.008**				
		(0.003)	(0.003)	(0.003)				
Confidence			0.055***	0.052***				
			(0.014)	(0.014)				
Self-judged ’Expertise’			0.004	0.003				
			(0.017)	(0.017)				
Anchor Value * Order Taken		0.011**	0.007	0.007				
		(0.004)	(0.005)	(0.005)				
Expertise (accuracy)					0.0004	−0.020		
					(0.019)	(0.022)		
Anchor * Expertise					−0.113***	−0.050		
					(0.032)	(0.044)		
Anchor^2 * Expertise						0.053*		
						(0.028)		
Anchor^3 * Expertise						−0.033**		
						(0.016)		
Reasoning skill							0.018	0.034*
							(0.017)	(0.019)
Anchor * R. Skill							0.006	0.022
							(0.025)	(0.034)
Anchor^2 * R. Skill								−0.036*
								(0.020)
Anchor^3 * R. Skill								−0.013
								(0.011)
Constant	−0.001	−0.054**	−0.204***	−0.192***	0.010	−0.013	−0.046	−0.079*
	(0.016)	(0.027)	(0.046)	(0.050)	(0.030)	(0.032)	(0.042)	(0.046)
Observations	2,800	2,800	2,262	2,262	1,252	1,252	1,325	1,325
R^2	0.070	0.074	0.083		0.094	0.098	0.072	0.075
Adjusted R^2	0.069	0.072	0.080		0.090	0.093	0.069	0.070
Akaike Inf. Crit.				4,975.201				
Bayesian Inf. Crit.				5,038.165				
Note:								*p<0.1; **p<0.05; ***p<0.01

Note that in these and the following models, N may be reduced due to only a subset of subjects either completing the reasoning task, or answering sufficient questions to allow for an accuracy-based expertise measure; N is further reduced in the split-sample models.

Effects on individual prediction accuracy

Table S2 shows the results from OLS and multilevel regression of standardized prediction accuracy (squared error, inverted) on the specified subject features. Linear mixed-effects models contain random effects for individuals and questions.

Table S2. The effects of treatments and user qualities on prediction accuracy.

	Dependent variable:						
	Accuracy (squared error, reversed)						
	OLS	*linear mixed-effects*			*OLS*		*linear mixed-effects*
	(1)	(2)	(3)	(4)	(5) High	(6) Low	(7)
Anchor Shown	0.238^{***}	0.239^{***}	0.251^{***}	0.468^{***}	0.179	0.351^{**}	0.159
	(0.069)	(0.062)	(0.077)	(0.177)	(0.139)	(0.149)	(0.138)
Question Order	0.020^{***}	0.014^{**}	0.012	0.012	0.038	0.041	0.041^{**}
	(0.007)	(0.007)	(0.008)	(0.008)	(0.025)	(0.029)	(0.017)
Reason Shown	0.007	-0.004	0.035	0.315^{*}	0.099	0.401^{***}	0.152
	(0.069)	(0.062)	(0.077)	(0.177)	(0.138)	(0.149)	(0.137)
Reason Requested	-0.185^{**}	-0.117^{*}	-0.109	-0.103	0.111	-0.539^{***}	-0.170^{*}
	(0.074)	(0.067)	(0.082)	(0.082)	(0.148)	(0.160)	(0.097)
Confidence	0.113^{***}	0.040	0.050	0.048	0.121^{*}	0.161^{**}	0.076^{*}
	(0.032)	(0.031)	(0.037)	(0.037)	(0.064)	(0.068)	(0.042)
Self-judged 'Expertise'	-0.170^{***}	-0.143^{***}	-0.056	-0.058	-0.081	-0.123	-0.076
	(0.039)	(0.037)	(0.046)	(0.046)	(0.088)	(0.076)	(0.051)
Time Taken	0.343^{***}	0.219^{***}	0.201^{***}	0.201^{***}	0.291^{***}	0.559^{***}	0.259^{***}
	(0.041)	(0.040)	(0.047)	(0.047)	(0.085)	(0.089)	(0.056)
Reasoning Ability			0.084^{**}	0.203^{***}			
			(0.037)	(0.065)			
Anchor Shown * R. Ability				-0.100			
				(0.074)			
Reason Shown * R. Ability				-0.129^{*}			
				(0.074)			
Expertise							0.216^{**}
							(0.100)
Anchor Shown * Expertise							-0.091
							(0.103)
Reason Shown * Expertise							-0.073
							(0.103)
Constant	-2.791^{***}	-2.078^{***}	-2.292^{***}	-2.547^{***}	-3.177^{***}	-4.247^{***}	-2.686^{***}
	(0.213)	(0.330)	(0.377)	(0.394)	(0.529)	(0.543)	(0.455)
Observations	4,467	4,467	2,384	2,384	1,021	997	2,026
R^2	0.026				0.023	0.065	
Adjusted R^2	0.024				0.016	0.058	
Akaike Inf. Crit.		19,302.680	9,842.834	9,848.650			8,704.086
Bayesian Inf. Crit.		19,373.130	9,906.376	9,923.745			8,777.066
Note:							*p<0.1; **p<0.05; ***p<0.01

Although it was not originally part of our experimental design, our results illuminate one way in which expertise causally determines success. Note that time taken is strongly associated with accuracy ($\beta = 0.34$, $p < 0.001$), though this association is somewhat weaker for experts. In general, of course, it is unclear whether time taken causes success, or is merely a correlate of it, but we also noted that asking subjects to provide a reason ("Reason Requested") both increases the time taken on a task, and decreases accuracy for the low-expertise subjects (low: $\beta = -0.54$, $p < 0.001$; high: $\beta = 0.11$, $p = 0.453$). This treatment was randomized per subject and functions as a handy instrument for time taken. By examining the effect on accuracy of requesting a reason with and without controlling for time taken, we find that for experts, requesting a reason increases the time they spend on a problem, and that extra time (instrumented by the requested reason) boosts their accuracy; however, this effect is offset by the request itself, which seems to harm accuracy. For non-experts, while the request similarly boosts time taken and thereby accuacy, the directly harmful effect is larger, with a conditionally negative effect of the request when controlling for time taken. These results are in line with work finding that explicit reflection can lead to lower accuracy , but our results also suggest that this direct reasoning cost is higher for lower-expertise subjects, and in fact if you take into account the additional time reasoning requires, the net benefit may be positive for higher-expertise subjects. In sum, experts are more sensitive to the quality of the information they receive, are more likely to discount low-quality information, and are only benefited by being forced to take more time or reason explicitly, while the opposite holds for non-experts. In practice, these findings suggest an accuracy-maximizing strategy of eliciting predictions and reasons from experts (who are not harmed by the request) and delivering that information preferentially to non-experts (who benefit most).

Effects of reason content on individual accuracy

Table S3 shows the effects of various NLP content on individual accuracy, also subdivided by expert (High) and non-expert (Low) subjects and interacted with expertise (Int). Table S4 shows the effect of a reason's rated accuracy. Rating values are the mean of all subjects who saw that reason, which generally ranges between 1 and 10 subjects. Table S5 shows the effects of the unseen anchor value made by the author of a treatment reason.

Table S3. Effects of reason NLP content on accuracy

	Dependent variable:			
	MSE Accuracy			
	All	High	Low	Int
Num. Words	0.006	-0.018	0.018^{*}	-0.004
	(0.006)	(0.013)	(0.009)	(0.009)
Avg. Word Length	-0.003	-0.009^{*}	0.002	-0.010^{**}
	(0.003)	(0.005)	(0.006)	(0.005)
Includes Numbers	0.284^{**}	0.757^{***}	0.152	1.044^{***}
	(0.131)	(0.284)	(0.216)	(0.288)
Repeated CAPS	-0.143	0.375	-0.168	0.197
	(0.267)	(0.659)	(0.405)	(0.362)
Has questions	-0.076	0.081	-0.714	-0.433
	(0.255)	(0.568)	(0.452)	(0.402)
Includes URL	0.426^{**}	0.154	0.393	0.552
	(0.172)	(0.387)	(0.272)	(0.385)
Expertise				0.034
				(0.104)
Avg. Word Length * Expertise				-0.003
				(0.007)
Includes Numbers * Expertise				0.458^{**}
				(0.209)
Includes URL * Expertise				0.104
				(0.290)
Constant	-1.282^{***}	-0.928^{***}	-1.290^{***}	-1.033^{***}
	(0.113)	(0.242)	(0.183)	(0.191)
Observations	2,249	468	536	861
R^2	0.007	0.024	0.021	0.029
Adjusted R^2	0.004	0.012	0.009	0.018

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table S4. Effects of (unseen) anchor values associated with reasons.

	Dependent variable:							
	Raw Prediction							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	High	Low	Int	All	High	Low	Int
Reason Quality					-0.003	0.004	-0.022	-0.007
					(0.018)	(0.034)	(0.034)	(0.036)
Reason's Prediction	0.069^{***}	0.060	0.145^{***}	0.033	-0.112^{**}	-0.197^{*}	0.077	-0.485^{***}
	(0.022)	(0.046)	(0.047)	(0.047)	(0.057)	(0.114)	(0.137)	(0.135)
Expertise								-0.045
								(0.077)
Expertise * R. Prediction				-0.064^{**}				-0.457^{***}
				(0.031)				(0.119)
Expertise * R. Prediction * Rating								0.143^{***}
								(0.043)
R. Prediction * Rating					0.077^{***}	0.117^{**}	0.029	0.204^{***}
					(0.022)	(0.047)	(0.048)	(0.051)
Expertise * Rating								0.006
								(0.027)
Constant	-0.055^{***}	-0.074^{**}	-0.029	-0.050^{**}	-0.045	-0.078	0.030	-0.056
	(0.018)	(0.035)	(0.034)	(0.025)	(0.051)	(0.096)	(0.099)	(0.103)
Observations	2,087	435	494	929	2,047	426	486	912
R^2	0.005	0.004	0.019	0.014	0.010	0.019	0.023	0.038
Adjusted R^2	0.004	0.002	0.017	0.012	0.009	0.012	0.017	0.030
Note:								*p<0.1; **p<0.05; ***p<0.01

Effects do not vary with subject pools

Table S5 shows that our estimates do not vary with subject type, which was drawn from two pools, volunteers and Mechanical Turk. Therefore all results shown elsewhere pool all subjects.

Table S5. Effects do not depend on subject type (volunteer vs Mechanical Turk).

	Dependent variable:
	MSE Accuracy
Anchor Shown	0.242***
	(0.088)
Question Order	0.012
	(0.008)
Reason Shown	0.099
	(0.087)
Reason Requested	-0.106
	(0.095)
Confidence	0.053
	(0.037)
'Expertise'	-0.004
	(0.054)
Time Taken	0.178***
	(0.054)
Female	0.041
	(0.086)
Age	0.003
	(0.004)
Education	0.017
	(0.030)
Knowledge	0.005
	(0.028)
Reasoning Ability	0.077**
	(0.039)
Volunteer vs MTurk	0.081
	(0.550)
Anchor Shown * VvsM	0.037
	(0.188)
'Expertise * VvsM	-0.176^{*}
	(0.092)
Reason Shown * VvsM	-0.287
	(0.187)
Reason Requested * VvsM	-0.019
	(0.199)
Time Taken * VvsM	0.091
	(0.113)
Constant	-2.506^{***}
	(0.420)
Observations	2,366
Akaike Inf. Crit.	9,815.702
Bayesian Inf. Crit.	9,936.850
Note:	*p<0.1; **p<0.05; ***p<0.01

The 16 prediction task questions.

Table S6. Stimuli: 16 prediction questions

Category	Question
politics	What percentage of the votes will Sauli Niinistö receive in the first round of the Finnish Presidential election on January 28, 2018?
politics	What percentage of vote will Nikolas Papadopoulos receive in the Cypriot Presidential election on the 28th of January?
politics	What will be the approval rate for the Russian government at the end of January?
politics	What will Donald Trump's Real Clear Politics average approval rating be on the 21st of January?
entertainment	By how many points will the Knicks defeat the Lakers in the Knicks vs. Lakers NBA game on January 21, 2018? (If the Lakers win, use negative points.)
entertainment	How many awards will "The Shape of Water" win at the 2018 BAFTAs (British Film Awards)?
entertainment	On Spotify's Global Top 50 chart, what place will Ed Sheeran's song "Shape of You" take on the 21st of January?
entertainment	What will the daily box office gross of "Star Wars: The Last Jedi" be in USD on the 21st of January?
economics	What will Amazon's stock price (AMZN) in USD be at the close of trade on the 21st of January?
economics	What will be the value of one bitcoin in USD at 11:59pm on the 21st of January?
economics	What will be the value of one US dollar in South African Rand at 11:59pm EST on the 21st of January?
economics	What will the silver price per ounce be in USD at the close of trade on the 21st of January?
health/weather	How many cases of flu will be recorded in Spain in the third week of January? This will be resolved using http://www.who.int/influenza/gisrs_laboratory/
health/weather	How many cases of MERS-CoV will be found in Asia between the 16th and the 30th of January? This will be resolved using http://empres-i.fao.org.
health/weather	How many earthquakes of magnitude 4.9 or stronger will occur worldwide between the 16th of January and 21st of January?
health/weather	What will the high temperature be in Doha, Qatar on the 21st of January in Fahrenheit?

Differential anchoring effects with questions

Subjects were asked to estimate future numbers or counts for the events shown in Figure S.1. We estimated the multilevel regression model shown in Figure 3, Model 2, with random effects for each question, which are plotted below along with 95% credible intervals. Events fell into four categories (politics, economics, entertainment, and weather/health). Note that anchoring effects appear lowest for economic, health, and sports topics which are presumably easier to estimate via internet search.

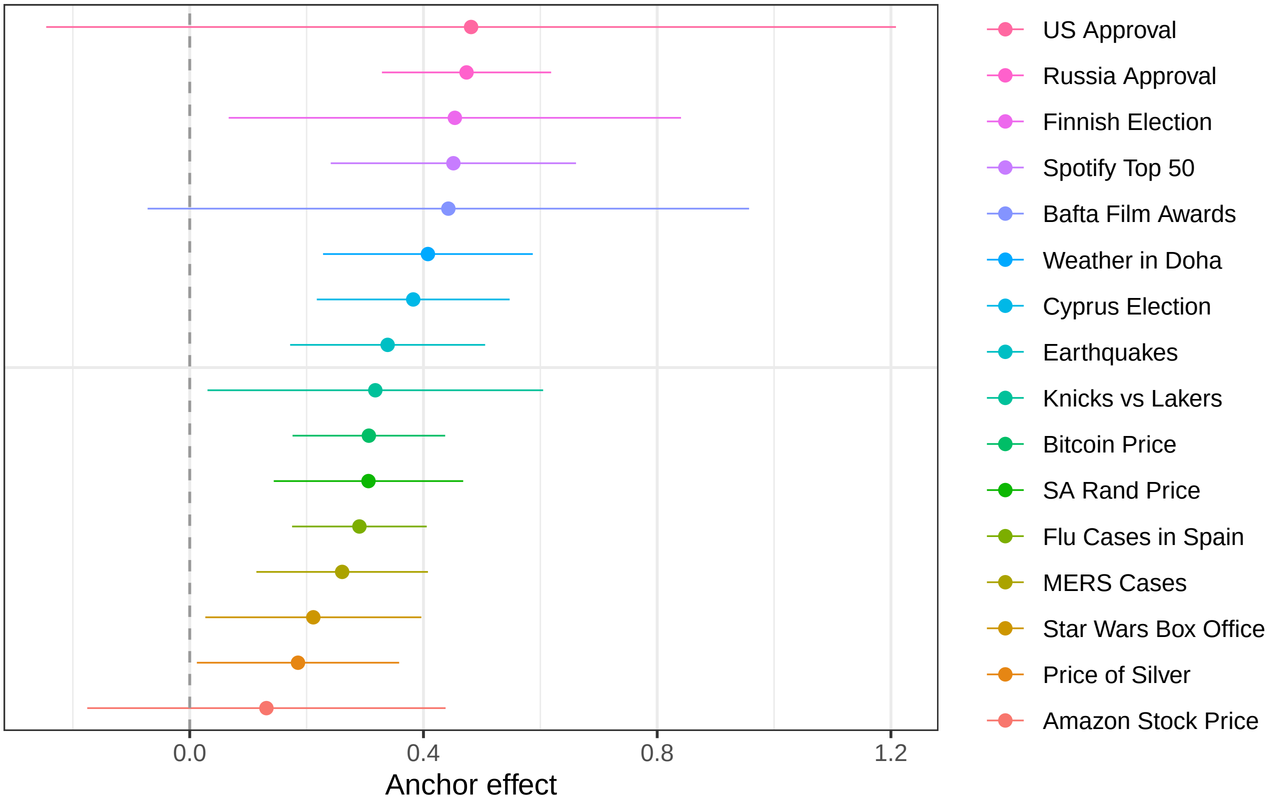

Fig. S1. Anchor effect per question from random effects model shown in Figure 3, Model 2.

Measuring expertise

To measure personal traits, we presented subjects with a three-question reasoning quiz and a political knowledge test, and also asked their education level. To measure self-judged ability, we asked subjects their degree of confidence in each question as well as their self-assessed "expertise" in that topic area. And finally, as the most direct measure of each subject's task-specific ability, we assessed each subject's accuracy in the first 50% of the (randomly ordered) prediction questions they were presented with, restricting this to the first half of each individual's responses. The personal traits were all mildly correlated with each other, ranging from 0.09 (reasoning skill and education) to 0.27 (knowledge and education). Similarly, the task-based measure of expertise was correlated similarly with the others (0.07, 0.11, 0.14 for education, knowledge and reasoning skill respectively). However, the self-assessed, per-question measures of confidence and expertise, while correlated at 0.41 with each other, were either uncorrelated or negatively correlated with the other personal traits, a result consistent with some previous work in this area (1, 2).

The negative correlation between accuracy and self-judged "expertise," however, is not a straightforward instance of the well-known Dunning-Kruger effect, where those that are most certain are often least skilled (3). Figure S2 shows that the relationship between true individual-level accuracy and self-judged "expertise" is quite non-linear: among those who consider themselves most expert (> 3 on a 5-point scale), there is a negative, Dunning-Kruger-like relationship between self-judged expertise and actual accuracy; but among those who rate themselves more poorly (< 2), there is a positive relationship. This non-linear relationship may explain some previous contradictory Dunning-Kruger studies (4–6).

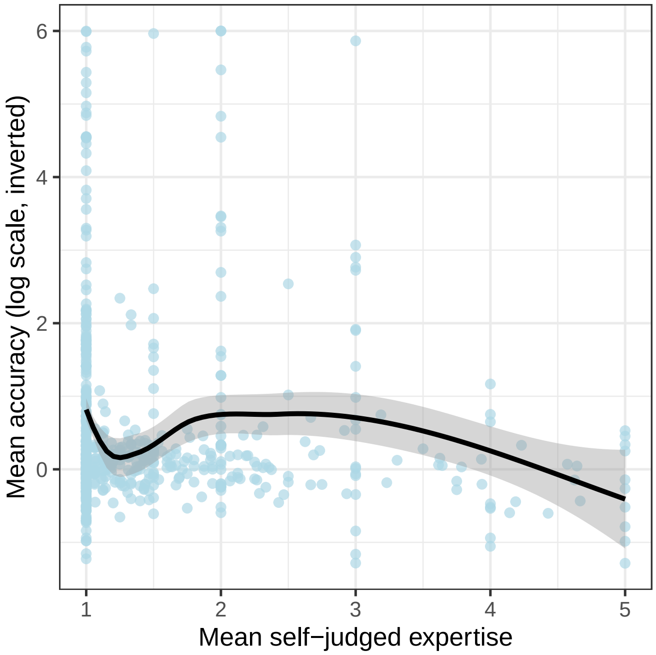

Fig. S2. Individual prediction accuracy (expertise) vs mean self-judged 'expertise,' with loess curve. Accuracy is mean squared error between prediction and outcome after standardizing by question, logged and then inverted so that higher values are better.

Simulation of expertise-dependent social communication

Our simulation experiment measures the effects of expertise-dependent sensitivity to peer information on collective accuracy. The pseudo-code is shown below. The number of participants tested N was from the set {5,10,100} and the expertise factors tested (Exp) was from the set {0,0.25,0.5}. Exp determines the total variance in expertise (and hence responsivity to peer information) among individuals in a trial. For each trial, information was exchanged N or $10N$ times.

```
1. Each individual $i$ is assigned expertise $e_i$ uniformly drawn from $U[1/2 - Exp, 1/2 + Exp]$. Note that the mean remains the same for all $Exp$ in $\{0, 0.25, 0.5\}$.
2. Each individual is assigned $\sigma_1^2 = 0.5 + (e_i - 0.5)/2$ and $\gamma = 1.5 - 2(e_i - 0.5)$. Lower $e_i$ (more expert) will thus have lower $\sigma_1^2$ and higher $\gamma$. See Equation 3 and following in main text.
3. Each individual begins with initial perception $p_i$ drawn from $N(\mu, e_i)$, where $\mu$ is drawn from $N(0, 1)$ for each trial.
4. For $N$ or $10N$ steps, repeat:
    (a) Draw at random receiver $i$ and sender $j$
    (b) $i$ updates their belief $p_i$ after seeing $p_j$ according to Equation 3.
5. Assess aggregate error $(\bar{p_i} - \mu)^2$
```

References

1. D Guilbeault, D Centola, Networked collective intelligence improves dissemination of scientific information regarding smoking risks. *PLOS ONE* **15**, 1–14 (2020).
2. Y Attali, D Budescu, M Arieli-Attali, An item response approach to calibration of confidence judgments. *Decision* **7**, 1 (2020).
3. D Dunning, K Johnson, J Ehrlinger, J Kruger, Why people fail to recognize their own incompetence. *Curr. directions psychological science* **12**, 83–87 (2003).
4. T Schlösser, D Dunning, KL Johnson, J Kruger, How unaware are the unskilled? empirical tests of the "signal extraction" counterexplanation for the dunning–kruger effect in self-evaluation of performance. *J. Econ. Psychol.* **39**, 85–100 (2013).
5. E Nuhfer, C Cogan, S Fleisher, E Gaze, K Wirth, Random number simulations reveal how random noise affects the measurements and graphical portrayals of self-assessed competency. *Numeracy: Adv. Educ. Quant. Lit.* **9** (2016).
6. E Nuhfer, S Fleisher, C Cogan, K Wirth, E Gaze, How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: Numeracy underlies better alternatives. *Numeracy: Adv. Educ. Quant. Lit.* **10** (2017).