CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Gender, confidence, and the mismeasure of intelligence, competitiveness, and literacy
Authors	Harrison, Glenn;Ross, Don;Swarthout, J. Todd
Publication date	2026-01-30
Original Citation	Harrison, G, Ross, D & Swarthout, J T 2026, 'Gender, confidence, and the mismeasure of intelligence, competitiveness, and literacy', Journal of Political Economy, vol. 134, no. 2, pp. 665-730. https://doi.org/10.1086/738345
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://www.scopus.com/pages/publications/105028792262 - 10.1086/738345
Rights	© 2026 The University of Chicago. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial reuse of the work with attribution. Rights for text and data mining and training of artificial intelligence technologies or similar technologies are reserved. Published by The University of Chicago Press. - https://creativecommons.org/licenses/by-nc/4.0/
Download date	2026-11-14 15:19:51
Item downloaded from	https://hdl.handle.net/10468/18606

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Gender, Confidence, and the Mismeasure of Intelligence, Competitiveness, and Literacy

Glenn W. Harrison

Georgia State University and University of Cape Town

Don Ross

University College Cork, University of Cape Town, and Georgia State University

J. Todd Swarthout

Georgia State University

The measurement of intelligence should identify and measure an individual's subjective confidence that a response to a test question is correct. Existing measures do not do that, nor do they use extrinsic financial incentive for truthful responses. We rectify both issues and show that each matters for the measurement of intelligence, particularly for women. Our results on gender and confidence in the face of risk have wider applications in terms of the measurement of "competitiveness" and financial literacy. Contrary to received literature, women are *more* intelligent than men, compete when they should in risky settings, and are *more* literate.

Most measures of intelligence score responses to questions as either right or wrong. If a test is to have any interesting discriminatory power, it needs to include some questions that some individuals do not answer

We are grateful to three referees, the editor, and participants at numerous presentations for constructive comments. This paper was edited by James Heckman.

Electronically published January 30, 2026

Journal of Political Economy, volume 134, number 2, February 2026.
© 2026 The University of Chicago. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial reuse of the work with attribution. For commercial use, contact journalpermissions@press.uchicago.edu. Rights for text and data mining and training of artificial intelligence technologies or similar technologies are reserved. Published by The University of Chicago Press.
https://doi.org/10.1086/738345

correctly. Incorrect answers reflect incorrect beliefs about the correct answer, if we assume that individuals are motivated to respond truthfully. We elicit the subjective beliefs of individuals over possible distributions of answers to a popular measure of (fluid) intelligence. Belief distributions matter for the measurement of intelligence in populations, but conventional scoring systems reflect only *modal* beliefs over discrete alternatives. As a result, we argue that conventional scoring systems lead to a mismeasure of intelligence, which can be corrected by ensuring that they elicit the confidence of beliefs. A corollary of eliciting belief distributions with incentives, providing controlled consequences for different possible reports, is that we must also attend to the risk preferences of individuals. In turn, a failure to attend to risk preferences in the face of choices conditioned on subjective beliefs has led to a mismeasure of other traits, such as "competitiveness."

We consider the extent to which this mismeasure of the traits of intelligence, competitiveness, and literacy changes the received thinking about the role of gender. Our results are striking and lead us to conclude that women process the confidence of their choices over risky responses better than men when it comes to intelligence, no worse than men when it comes to competitiveness, and more consistently when it comes to literacy. More generally, there are deeper reasons to be interested in the confidence with which individuals make knowledge claims over risky choices in tests and in life.

First, information on the modal belief, with no other information about the subjective belief distribution to put the mode in context, can misrepresent the knowledge claim of the individual. If the individual subjectively perceives some risk in choosing one option over another, it does a disservice to that perception to artifactually require it to masquerade as a riskless option. We should, arguably, be interested in the individual's risk perception. Have they at least managed to rule out "obviously dominated" options? Did the correct answer receive ε less subjective weight than the modal belief or no weight?

Second, in most models of decision-making under risk, the properties of the complete subjective belief distribution matter for decisions. Under subjective expected utility, it is the weighted average belief that matters for risk preferences, not the modal belief (Savage 1972). And under models of uncertainty or ambiguity aversion, the whole distribution typically matters (e.g., Klibanoff, Marinacci, and Mukerji 2005). So the manner in which intelligence translates into decisions depends, in general, on the distribution of beliefs about correct answers just as much as it depends on how the risks, uncertainties, or ambiguities are traded off.

Third, a subjective perception of "less than complete confidence" in choosing one option over another need not signal a deficit of intelligence, in a broader sense, if it serves as a trigger for the individual to seek out

some "cognitive scaffold" to help make a final decision. For now, think of scaffolds in terms of aids to discrimination, such as arranging files in a folder or counting on one's fingers—in principle, things that Robinson Crusoe might access (sec. IV.G provides a review of the origins of the concept of scaffolding). We include language here, since it serves cognitive functions going beyond communication (Bickerton 2014; Dennett 2017). Naming something stabilizes it in memory and for reidentification. It is an important question, for some, if one wants to call such enhanced capability "intelligence" or reserve that word for speed and power of processing in the brain. We prefer the broader interpretation, following Clark (2003). In any event, the availability of scaffolded responses may be expected to influence the subjective beliefs of agents about responses to questions on intelligence tests; our elicitation interface itself is a general scaffolding treatment.

Fourth, for a social animal such as homo sapiens, the set of relevant scaffolds extends to interactions with others. Obvious examples include accessing the internet, consulting an expert, conversing with a partner, or learning from a neighbor. Language is again a scaffold but a more powerful one: that a culture agrees on a name for something is highly informative (Dennett 2017; Planer and Sterelny 2021). Here we have an even broader sense of the word "intelligence," as collective and cultural, although the choices we observe on the test are still literally those of an individual agent.

Fifth, culture makes social scaffolds differentially available and salient to distinct demographic groups in populations. This is a crucial basis for caution about findings that attribute observable differences in cognitively mediated behavioral responses between genders and races, for example, to inherent properties of these groups. Such simplistic and socially problematic inferences might be undermined by varying the salience of scaffolds in experiments and reassessing gender and race effects in light of such experimental treatments. Our results provide a start at such reassessment, with a focus on gender, and offer some surprising conclusions.

Finally, the importance of "cognitive skills" for earnings, inequality, and other lifetime outcomes is well-documented, so there is a derived demand for more fundamental measurement of one core cognitive skill—fluid intelligence. Better measurement of cognitive skills in general should contribute to the debate over the relative importance of cognitive and noncognitive skills across the life cycle of individuals and households.[1]

[1] Hanushek and Woessmann (2008) review a literature showing interactions between the population distribution of cognitive skill levels and educational institutions. And Heckman (2008) reviews a rich literature showing different interactions of cognitive and noncognitive skills over the life-cycle of individuals, with significant implications for the impact of early-childhood programs on lifetime outcomes.

Moreover, this broader view of natural intelligence plays a critical part of the evolutionary story of our species; see Harrison and Ross (2025).

Following Smith (1982, 931–38),[2] we presume that financial incentives can be made large enough to be salient for subjects and to dominate incentives they might have to try to do something other than provide their most subjectively accurate response within the frames of questions. We provide some additional empirical tests of this presumption, complementing and confirming a wide range of findings that financial incentives do matter, on average, for better performance in tests of this kind.[3]

However, our focus is not on the role of incentives in general, which has been well-established across several literatures, but on the nature of the incentives needed for measures of intelligence to properly reflect confidence in responses. We augment existing measurement methods for measuring intelligence in order to capture a missing trait: self-awareness of the degree of confidence one has in a knowledge claim and a willingness to express that confidence.

The specific measure of intelligence we consider is the Raven Advanced Progressive Matrices (RAPM) test, documented by Raven, Raven, and Court (1998).[4] The primary component of this test is called set II;[5] it

[2] The five precepts proposed by Smith (1982) are nonsatiation in the reward, salience (the reward varies in a known way with performance and actions by the agent), dominance, privacy of rewards, and parallelism of laboratory results to field environments.

[3] Borghans et al. (2008), Borghans, Meijers, and Ter Weel (2008) Segal (2012), and Chen et al. (2020) survey evidence on the role of incentives on measures of cognitive and noncognitive ability, finding positive effects.

[4] Although the Raven suite of problems is widely used to measure fluid intelligence, we should not forget that it is just a constructed task. Heckman and Kautz (2012, 452) explain this as follows:

> Psychological traits are not directly observed. There is no ruler for perseverance, no caliper for intelligence. All cognitive and personality traits are measured using performance on "tasks," broadly defined. Different tasks require different traits in different combinations. Some distinguish between measurements of traits and measurements of outcomes, but this distinction is misleading. Both traits and outcomes are measured using performance on some task or set of tasks. Psychologists sometimes claim to circumvent this measurement issue by creating taxonomies of traits and by applying intuitive names to responses on questionnaires. These questionnaires are not windows to the soul. They are still rooted in task performance or behavior. Responding to a questionnaire is itself a task.

Similarly, the history of science informs us that concepts of female or ethnic intellect have often been constructed on shifting notions of what is determined to be the "nature" of the individual with that gender or race; see, e.g., Daston (1992).

[5] There is a set of 12 problems in the RAPM called set I, which are generally easier than those in set II. The set I problems are often used as a fast, coarse measure of intelligence, to determine whether the more discriminatory set II is needed for some subjects. One can also use set I to allow subjects to learn the nature of the task, and all of our subjects had completed set I in a prior experiment, as part of a series of hypothetical survey questions. There are other versions of the Raven tests. The Raven Standard Progressive Matrices (RSPM) test is for teenagers or less formally educated subjects (Raven, Raven, and Court

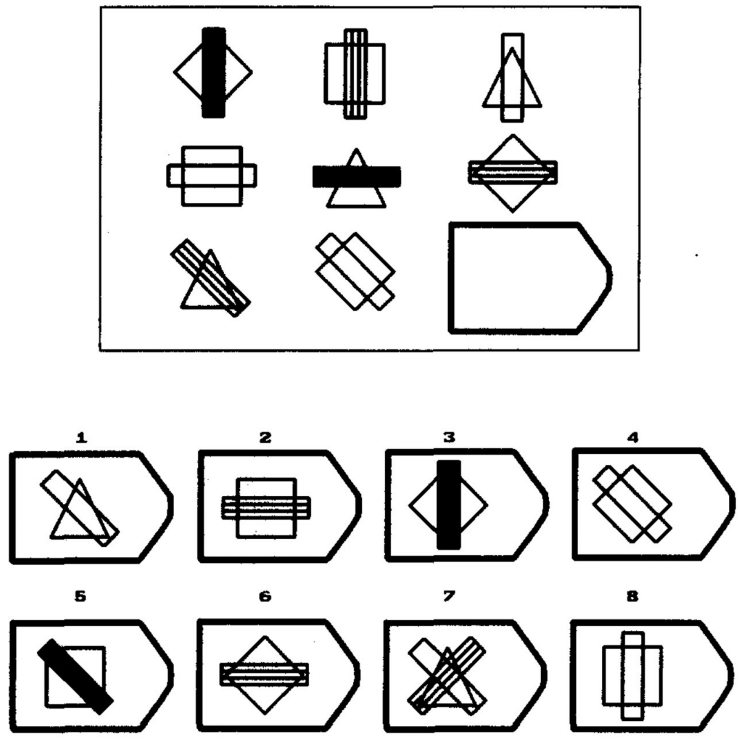


FIG. 1.—Ravenlike problem.

consists of 36 problems illustrated by an isomorph shown in figure 1.[6] Each problem consists of a block of nine images, arrayed as a 3×3 matrix, with one image deleted. The subject is presented with eight possible solutions to fill in the deleted image and is instructed to select the

2000), and the Raven Coloured Progressive Matrices test is for children or those with cognitive limitations, e.g., dementia (Raven, Raven, and Court 1962). The research literature on intelligence that uses these tests has focused on the RAPM and, to a lesser extent, the RSPM.

[6] It is inappropriate to display the actual test stimuli, to preserve test validity. The example in fig. 1 is used in our instructions, and is from Carpenter, Just, and Shell (1990). Moreover, one must pay for access to the test, as is common with many psychological batteries. There have been some efforts to write software to generate "Raven-like" problems that allow one to have an extended battery to select from, as well as avoid the fees for accessing the commercial version of the test; see Matzen et al. (2010), Wang and Su (2015), and Civelli and Deck (2018). And there are many instances in which "Raven-like" tests are developed and used, e.g., the eight questions developed by Borghans et al. (2009, 652) and used "to measure IQ."

correct image. With minor modification of the original instructions, the task presented in figure 1 is explained as follows:

> Only one of these pieces is perfectly correct. Pieces #2, #5, and #8 complete the problem correctly going down, in the sense that they have a square as the larger piece. Pieces #1, #4, and #5 are correct going across, in the sense that they have the right slope for the rectangle. Piece #5 is the correct bit, isn't it? It is correct going down as well as going across. So the answer is #5.

It is apparent that some possible answers are better than others, but only one is completely correct. It is also apparent, even from this relatively simple problem, that some possible answers can be eliminated relatively quickly.

The original RAPM was administered in a "pen and paper" fashion, with prepared answer sheets. The default version is untimed, in the sense that the individual is allowed any amount of time to complete the test. In practice, the test requires between 30 and 50 minutes to complete. There is also a popular timed version, in which individuals are told that they have only 40 minutes: this is often referred to as a test of "intellectual efficiency," recognizing the trade-off with time. It is very rare that financial incentives are offered, a point to which we return.

To orient discussion, consider the accuracy of responses in nonsalient versions of the RAPM. By "nonsalient," we mean where there are no rewards for better or worse performance. In our case, subjects were participating in an experimental session that had included salient rewards in a prior, unrelated task and were told that they would receive $5 for completing the RAPM. In the usual psychology setting, it is common for students of large introductory classes to be "required" to take these tests, whether or not they are referred to later in the class pedagogy.[7]

Figure 2 displays the fraction of correct answers from 55 subjects recruited from the Georgia State University (GSU) undergraduate population, as well as subjects recruited from comparable populations by Arthur and Day (1994) and Gignac (2018).[8] The "progressive" nature of problem difficulty is apparent from the decline in accuracy. Several nonmonotonic blips are evident and due to variations and interactions in the rules used to solve the problems and the use of slightly easier problems to help individuals see the progression of rules.[9] Our subjects generally

[7] We have ethical concerns with this practice, but that is a separate matter.

[8] For all of the applications of the RAPM, there are very few tabulations of actual performance across the set of 36 (or 48) questions.

[9] Carpenter, Just, and Shell (1990, table 1, 431) provide a taxonomy of five types of rules used in the RAPM and list which combinations of rules apply in each problem.

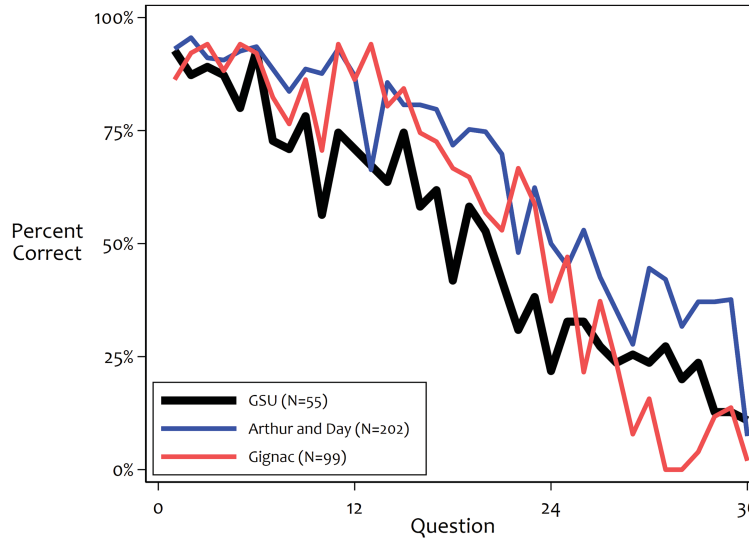


FIG. 2.—Raven scores with no salient rewards.

did worse on the first 25 problems. Our sample may well have been influenced by the nature of the sample selection process we used: our subjects were recruited to participate in experiments for financial rewards, and our lab has a reputation for providing "generous" rewards.[10] Although these subjects had participated in a task that had salient rewards, they may have reacted negatively to a task that offered "only" $5 for completion. For purposes of comparison with the treatments with salient rewards for performance, of course, this baseline is appropriate because the samples were recruited identically and assigned at random to the treatments.

The point of figure 2, for our purposes, is to flag that the RAPM provides a standard intelligence test that takes most individuals from a position of being very confident that they know the correct answer, and being correct in that confidence, to a position of not being confident about any correct answer. It therefore provides a gradual array of problems from "easy" to "hard," without a sharp discontinuity, at least at the aggregate level shown in figure 2.

Our approach is to adapt this nonsalient intelligence test so that we elicit subjective beliefs of individuals in a salient, incentivized manner,

[10] For the experiments we conduct, spanning thousands of students over 10 years, we regularly see earnings averaging roughly $30 for a 2-hour session (excluding participation payment).

in order to elicit and characterize the confidence that individuals have in their responses. The basic idea was proposed long ago by de Finetti (1965), Shuford, Arthur, and Massengill (1966), and Savage (1971). The broader normative case for this approach was explained elegantly by Savage (1971, 800):

> Proper scoring rules hold forth promise as more sophisticated ways of administering multiple-choice tests in certain educational situations. . . . The student is invited not merely to choose one item (or possibly none) but to show in some way how his opinion is distributed over the items, subject to a proper scoring rule or a rough facsimile thereof. Although requiring more student time per item, these methods should result in more discrimination per item than ordinary multiple-choice tests, with a possible net gain. Also they seem to open a wealth of opportunities for the educational experimenter. Above all, the educational advantage of training people—possibly beginning in early childhood—to assay the strengths of their own opinions and to meet risk with judgment seems inestimable. The usual tests and the language habits of our culture tend to promote confusion between certainty and belief. They encourage both the vice of acting and speaking as though we were certain when we are only fairly sure and that of acting and speaking as though the opinions we do have were worthwhile when they are not very strong.

The passage of more than 50 years has not, it seems, diminished the need to address the confusion and vices referred to.

Section I reviews the experimental design and theoretical basis for our elicitation of belief distributions. We focus throughout on two experimental conditions: salient, incentivized responses in which the individual can report only their modal belief and salient, incentivized responses in which the individual can report degrees of confidence in their belief. We also consider a treatment that deviates from the progressive presentation of Raven problems in increasing difficulty to one in which the order of difficulty is scrambled at random. This treatment allows us to identify the effect on performance of the scaffolding provided by the progressive order of difficulty, as distinct from the underlying cognitive challenge of each Raven problem considered in isolation.

Section II reviews the results from our experiments on the measurement of intelligence, which focus on the effect of salient incentives and on the interaction of incentives with the confidence of subjects. When confidence is taken into account, we explain why it is natural to consider

performance in terms of a concept familiar to economists—earnings efficiency, rather than raw accuracy. We flag one striking demographic result, that women and Blacks do much better when allowed to report their full distribution of beliefs rather than being constrained to report only their modal beliefs. These results overturn long-standing claims than women and Blacks do poorly on measurements of intelligence, simply by expanding the measurement to appropriately incentivize reports of confidence in responses. We expand extensively on the finding with respect to gender effects later, partly by reference to an experimental extension we administered. Deeper follow-up on the result concerning race effects is warranted but left for future work. Finally, we demonstrate that there is a significant welfare cost to the respondent from being forced to treat her modal belief as if it were her only belief, held with certainty, about some risky set of alternatives. In effect, the measurement of performance in terms of expected welfare generalizes the measurement of intelligence in terms of earnings efficiency, by allowing for the risk preferences of agents making risky, consequential reports about their beliefs.

Section III considers our striking result from section II about the superiority of women with the correct measurement of fluid intelligence and examines how general it is in two well-studied applications: measures of the "competitiveness" of women and measures of the financial literacy of women. In each case, there are claims that women exhibit insufficient confidence in their behavior, and we show that this conclusion is again due to measurement using interim, observable performance metrics that have nothing to do with welfare when facing risk.

Section IV considers a number of extensions of our approach, related literature, and open issues. We examine the connection between tests designed to measure traits and tests used for more traditional educational purposes. A related concern is the relationship to achievement tests that have social implications. The nature of intrinsic motivation, which may be related to the underlying mechanism driving the gender effects we observe, is reviewed. We also discuss different ways in which subjective beliefs may be elicited and the trade-offs of using one method or the other. Various metrics have been proposed for the ex post evaluation of beliefs, such as the notion of "calibration," and we relate those metrics to the ones we adopt. Normative implications of our approach are considered. Finally, we review the history of the concept of scaffolding, which plays a central organizing role in our approach.

Section V draws general conclusions, with a normative emphasis. Our conclusions on gender and confidence provide striking contrasts to received claims and sharply reorient the way in which gender should be evaluated in a wide range of domains. However, as important and urgent as the implications for gender are, we argue that there are broader normative implications.

I. Experimental Design

A. Eliciting Beliefs

We use a quadratic scoring rule (QSR) to elicit properly incentivized subjective belief distributions for probability mass functions defined over the eight alternative solutions for each RAPM problem, such as shown in figure 1. The QSR provides financial incentives for individuals to report beliefs, with theoretical properties discussed in Harrison et al. (2017).[11] There are alternative methods for eliciting subjective beliefs distributions; the specific method used to elicit beliefs is not central to our methodological point, although of course it matters for the specific results.[12]

Figures 3 and 4 illustrate the interface used for elicitation. On the left is the usual Raven stimulus, with the possible solutions displayed. On the right is a bar chart showing the eight possible solutions, allowing the individual to allocate 80 tokens across the eight solution "bins." The initial allocation, shown in the top panel of figure 3, is for zero tokens to be allocated to all bins. The individual moves the slider for each bin to allocate tokens to that bin, and as they are moved, the QSR payoffs (shown graphically by the height of each bar and numerically by a monetary amount above each bar) are instantaneously updated. These payoffs show the earnings that would be received for a given bin if that bin corresponds with the correct answer to the RAPM question. Only when all 80 tokens are allocated can the subject confirm and move on to the next question, and the subject may freely reallocate tokens for a given question before confirming. By allowing conditional QSR payout values to be observed interactively in real time during the decision process, we make the QSR payment logic more transparent for subjects; this is in contrast with QSR applications in many prior experiments.[13] We do allow individuals to return

[11] McDaniel and Rutström (2001) is a significant precursor. They considered the Tower of Hanoi puzzle, which is a staple of the Montessori classroom as a task that allows the student to learn about backward induction. Apart from providing financial incentives for better performance, measured in terms of the least number of moves required to solve the puzzle and time taken, they elicited subjective beliefs over the least number of moves needed. Their elicitation method used 10 intervals evenly spread between 0 and 100 moves and used a quadratic scoring rule for each interval. Hence, subjects were paid for 10 elicitations of a binary outcome (e.g., the true solution is between 31 and 40 moves), not one elicitation of the complete probability mass function over all possible outcomes. They also implicitly assumed that subjects were risk-neutral when drawing inferences about beliefs from reported allocations; Andersen et al. (2014a) show that this assumption can be problematic for the incentivized elicitation of probabilities over a binary event.

[12] A review of alternative approaches is provided in sec. IV.D.

[13] It is tempting to use the QSR and not provide subjects with real-time earnings feedback. Subjects might just be assumed to correctly understand the relatively complex QSR formula, trust the experimenter by way of instructions along the lines of "it is in your best interest to state your true beliefs," or to use static lookup tables, as in McKelvey and Page (1990, 1336) and Rutström and Wilcox (2009, 620). We regard these alternatives

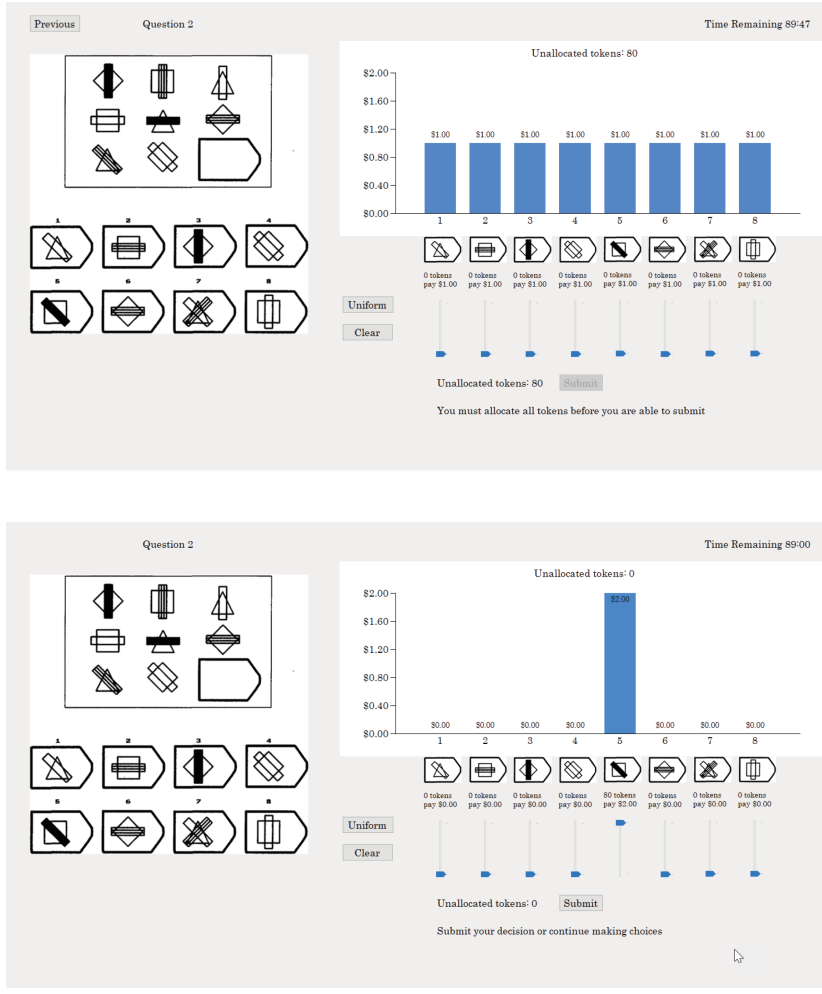


FIG. 3.—Belief elicitation interface with initial display and a completely correct response.

to a previously completed question and reallocate tokens, and only the final allocation for each question applies for earnings. The bottom panel of figure 3 shows a situation of complete confidence, where all 80 tokens are allocated to the correct answer, option 5.

Earnings accrue for each and every completed question, although the individual is not provided with any earnings information until all 36 problems have been completed and a confirmation is provided that

as being inferior to real-time appreciation of the salient rewards at play in QSR belief elicitation when contemplating alternative possible reports.

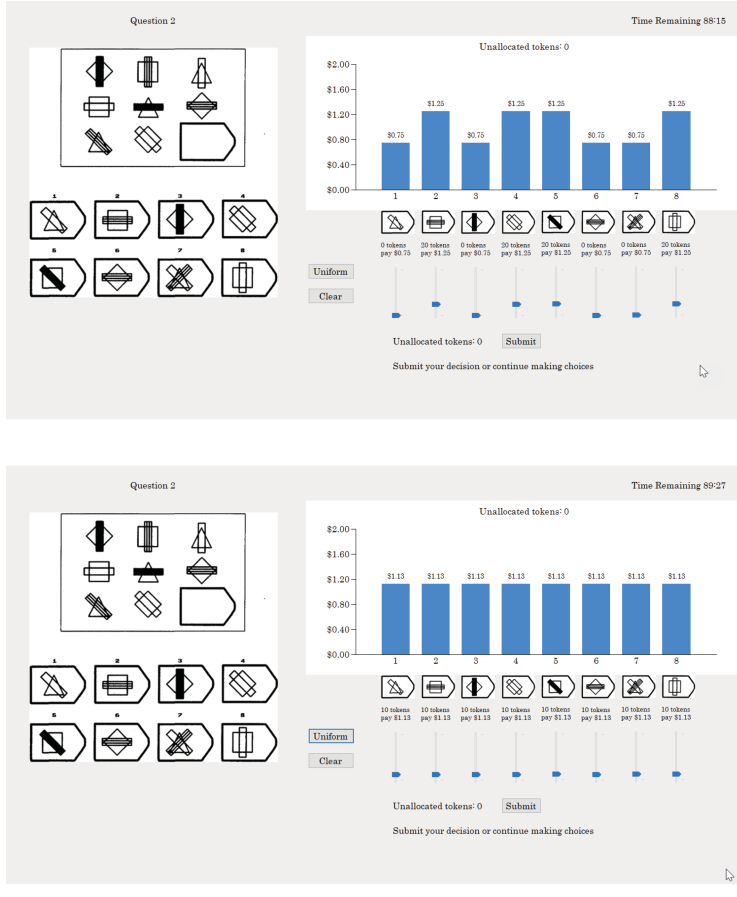


FIG. 4.—Belief elicitation interface with two possible displays of confidence.

there are no reallocations to be made. An exception occurs if the time limit is reached, in which case the individual knows that she would receive zero earnings for any question for which she has not submitted a token allocation.

The incentives shown in figures 3 and 4 imply that an individual would receive $2 if all 80 tokens were allocated to the correct answer, as shown in the bottom panel of figure 3. Possible "nondegenerate" allocations are shown in figure 4. The top panel of figure 4 shows a possible response to reflect this situation, as explained in the instructions:

> If you had decided that the correct answer was one of #2, #4, #5, or #8, but had not decided that #5 was actually the correct answer of these four possibilities, you might decide to allocate

> your tokens equally across the bars representing pieces #2, #4, #5, and #8 like this. . . .

Hence, the subject in this case has eliminated token assignment to possible answers that are less likely but not yet determined that only one of the possible answers is correct.

The bottom panel of figure 4 is important, because it reflects the optimal response if an individual has no clue which of the possible solutions might be correct. It also reflects an optimal response if the individual has some cautious apprehension that one or more possible solutions might be more likely but is too risk averse to want to allocate tokens that vary the payoffs over the eight possible solutions. If the subject allocates 10 tokens to each of the eight bins, then earnings are guaranteed to be $1.13 no matter what the correct solution is.

B. Treatments

We conduct a total of five treatments across subjects for our primary experiments on fluid intelligence and confidence, summarized in panel A of table 1. In the baseline treatment, the RAPM task is presented in the traditional manner: as a nonsalient, noncomputerized pen-and-paper task using the traditional printed test booklet and response sheet. Subject responses were typical multiple-choice format, and the QSR was not used. Subjects received $5 for completing the task, in addition to the general participation payment of $7. A total of 55 subjects participated in this treatment. The sole purpose of this treatment is to establish

TABLE 1
EXPERIMENTAL TREATMENTS

Treatment	Presentation of Questions	Sample Size	Response Format	Maximum Reward per Question ($)
	A. Fluid Intelligence, Confidence, and Scaffolds (secs. II and III)			
Baseline (traditional)	Progressive	55	No QSR; pen and paper	0
One Token Progressive	Progressive	95	QSR; software	2
80 Tokens Progressive	Progressive	82	QSR; software	2
One Token Scrambled	Scrambled	67	QSR; software	2
80 Tokens Scrambled	Scrambled	61	QSR; software	2
	B. Competitiveness, Confidence, and Welfare (sec. IV.A)			
One Token Piece Rate	Scrambled	43	QSR; software	2
One Token Tournament	Scrambled	43	QSR; software	8
One Token Select	Scrambled	43	QSR; software	2 or 8
80 Tokens Piece Rate	Scrambled	45	QSR; software	2
80 Tokens Tournament	Scrambled	45	QSR; software	8
80 Tokens Select	Scrambled	45	QSR; software	2 or 8

rough comparability to the traditional implementation of the Raven task.

The remaining four treatments in panel A of table 1 were computerized, and all computerized treatments were incentivized. One key treatment is whether the order of the questions for subjects was in the original progressive order of the RAPM or was scrambled at random. In the One Token Progressive and One Token Scrambled treatments, we give subjects only one token to allocate for each question. This token should optimally be allocated to the modal belief of the subject about the chances of the various answers being correct. The rationale for the One Token Progressive treatment is to have a computerized QSR treatment closely consistent with the baseline treatment, which constrains responses to just one answer in the traditional multiple-choice manner. In the 80 Tokens Progressive and 80 Tokens Scrambled treatments, we give subjects 80 tokens to allocate for each question. In all other aspects, the 80 Tokens Progressive and One Token Progressive treatments are identical, as are the 80 Tokens Scrambled and One Token Scrambled treatments. Each question is worth a maximum of $2, resulting in maximum possible earnings of $72 for the task. A total of 95, 92, 67, and 61 subjects participated in the One Token Progressive, 80 Tokens Progressive, One Token Scrambled, and 80 Tokens Scrambled treatments, respectively.

In general, we allow subjects 90 minutes to complete the RAPM but refer to this as untimed. The exact language used in the instructions was as follows: "You can have as much time as you want, although we have to be out of the room in 90 minutes." In practice, implementations of the RAPM in the literature that claim to have no time constraint probably have some such constraint, and we just wanted to be explicit about it. Only six subjects came within 5 minutes of that constraint, and the average time taken was only 45 minutes.[14]

One of our core findings from the primary experiments on fluid intelligence and confidence is that women perform much better than men when we measure intelligence with confidence, compared with traditional measurements that do not reflect confidence. We explore the generality of this important finding in a series of secondary experiments in section IV. One of these generalizations evaluates the claim that women "shy away" from competition when earnings are risky, and the other generalization evaluates the claim that women lack confidence when responding "do not know" to questions measuring financial literacy. Panel B of table 1 shows six treatments in which we reuse the Raven task as the

[14] In app. D (apps. A–D are available online), we report results from a pilot experiment in which we constrained subjects to have only 40 minutes for the 80 tokens task. There are no significant effects on performance, although suggestions that there might be significant effects for some demographic variables, e.g., gender.

basis for evaluating the claims about competition. The first three treatments—One Token Piece Rate, One Token Tournament, and One Token Select—are within-subject treatments in which the subject answers up to 12 scrambled Raven questions with specific payment rules and only one token to allocate. The last three treatments—80 Tokens Piece Rate, 80 Tokens Tournament, and 80 Tokens Select—are also within-subject treatments in which the subject answers up to 12 scrambled Raven questions with specific payment rules and 80 tokens to allocate. Subjects faced a 10-minute time constraint in all of these treatments, which are explained in section III.A.

II. Intelligence, Confidence, and Scaffolds

All subjects participated in sessions conducted in the ExCEN laboratory at Georgia State University during 2019 and 2022.[15] Subjects were recruited via email that contained no information about the task or expected earnings and were assigned randomly to treatments. Each subject received a participation payment of $7 in addition to task-specific earnings.

On arrival at the lab, subjects completed the Institutional Review Board (IRB) consent process and then received task instructions in both audio-video and print formats. By presenting the instructions with a prerecorded video, we minimize the amount of live speech during the session and better control for potential variation in spoken instructions across sessions. We exactly aligned the printed instructions with the video instructions and provided this alternative format for any participants who would rather read than watch instructions. Appendix A (apps. A–D are available online) contains printed instructions for all treatments.[16]

All subjects belonged to an IRB-approved panel that permits us to link responses across studies. Our 2019 subjects had participated in a prior experiment on an earlier date in which they completed a set of binary lottery choices, demographics, and the 12 questions from set I of the RAPM.[17] Our 2022 subjects undertook all tasks in the same session.

For convenience, we treat the reported token allocations as reflecting the subjective beliefs of the individual. This assumption is well-known to be true if the individual is risk neutral but also happens to be approximately true if the individual behaves consistently with expected utility theory (EUT) and has levels of risk aversion in the range normally found in laboratory experiments.[18] It is also literally true for most of the token allocations we

[15] There was a disruption in our ability to conduct in-person experiments in 2020 and 2021 due to the COVID-19 pandemic.

[16] Video instructions are available at https://cear.gsu.edu/gwh/raven/.

[17] For complete details of this prior experiment, see Harrison, Morsink, and Schneider (2022).

[18] This last theoretical result is established by Harrison et al. (2017). Essentially, risk-averse EUT subjects provide reports that are flattened with respect to their true beliefs,

observe in this instance, even when it is not true in general for belief elicitation. When the individual allocates all tokens to one answer, as many do for the early, easier questions in the RAPM, reports are beliefs. Similarly, when the individual reports a fully diffuse allocation of 10 tokens for each possible answer, as most do for the later, harder questions in the RAPM, reports are beliefs. And whenever the same number of tokens is allocated to two or more answers that receive any token allocations and zero to all others, risk preferences again literally play no role and reports are beliefs. It also follows that (plausible) risk preferences cannot matter significantly for allocations that approximately match these cases. The only stage when risk preferences might matter is the in-between cases where tokens are assigned sufficiently nonuniformly to some answers.[19]

At an aggregate level, pooling over the 36 questions for each individual, we are interested in accuracy and (earnings) efficiency. When someone must select only one option, accuracy is measured solely by the fraction of problems answered correctly. When degrees of confidence can be reported, this measure generalizes naturally to the fraction of tokens allocated to the correct answers. The concept of efficiency has long been used by experimental economists to measure the share of earnings that a subject realized divided by the earnings that could have been realized if all tokens had been allocated to the correct answer. In turn, earnings are defined by the reported QSR payoffs for the final token allocation submitted by a subject, which of course the subject saw, as illustrated in figures 3 and 4.

The reported QSR payoffs are directly linked to the QSR score itself. Let the decision-maker report her subjective beliefs in a discrete version of a QSR for continuous distributions. Partition the domain into K intervals, and denote as r_k the report of the likelihood that the event falls in interval $k = 1, \ldots, K$. Assume for the moment that the decision-maker is risk neutral and that the full report consists of a series of reports for each interval $\{r_1, r_2, \ldots, r_k, \ldots, r_K\}$ such that $r_k \geq 0 \ \forall \ k$ and $\Sigma_{i=1,K}(r_i) = 1$. If k is the interval in which the actual value lies, then the payoff score is defined by Matheson and Winkler (1976, 1088, eq. [6]): $S = (2 \times r_k) - \Sigma_{i=1,K}(r_i)^2$. So the reward in the score is a doubling of the report allocated to the true interval, and the penalty depends on how these reports are distributed across the K intervals. The subject is rewarded for accuracy, but if that accuracy misses the true interval, the punishment is severe. The punishment

so as to reduce variability over possible outcomes with positive subjective probability. This is a second-order adjustment in reports compared with the first-order adjustment in the binary case, where risk-averse individuals provide reports closer to one-half so as to reduce variability over the two possible outcomes.

[19] See Harrison et al. (2017) for formal statements of these intuitive results and figs. 8 and 9 for numerous examples from actual data.

includes all possible reports, including the correct one.[20] To ensure complete generality and avoid any decision-maker facing losses, allow some endowment, α, and scaling of the score, β. We then get the following scoring rule for each report in interval k that was actually used in our experiments, $\alpha + \beta[(2 \times r_k) - \Sigma_{i=1,K}(r_i)^2]$, where the QSR payoff score S had assumed $\alpha = 0$ and $\beta = 1$. We can specify $\alpha > 0$ and $\beta > 0$ to get the payoffs to any positive level and units we want. Hence, our use of the efficiency metric, based on expected earnings from the QSR, is monotonically linked to the QSR score itself.

A. *The Effects of Incentives for Accuracy and Confidence*

The first result displayed in figure 5 is that providing financial incentives significantly improves accuracy. The cleanest example here is to compare the baseline and One Token Progressive conditions, since subjects were forced to report their modal beliefs in both cases. Accuracy in the financially motivated test was 14.2 percentage points higher (p-value $< .001$).[21] We are here assuming that moving from the pen-and-paper version of the task to a computerized version has no effect on performance, at least with respect to accuracy. There have been several studies examining the effect of pen-and-paper versus computerized versions of the same test, although only a few controlling for measures of pretest ability. The available evidence suggests no differences in performance for students at a university level.[22]

[20] Take some examples, assuming $K = 4$. What if the subject has very tight subjective beliefs and allocates all of the weight to the correct interval? Then the score is $S = (2 \times 1) - (1^2 + 0^2 + 0^2 + 0^2) = 2 - 1 = 1$, and this is positive. But if the subject has tight subjective beliefs that are wrong, the score is $S = (2 \times 0) - (1^2 + 0^2 + 0^2 + 0^2) = 0 - 1 = -1$, and the score is negative. So we see that this score would have to include some additional "endowment" to ensure that the earnings are positive. Assuming that the subject has very diffuse subjective beliefs and allocates 25% of the weight to each interval, the score is less than 1: $S = (2 \times [1/4]) - ([1/4]^2 + [1/4]^2 + [1/4]^2 + [1/4]^2) = (1/2) - (1/4) = (1/4) < 1$. So the trade-off from the last case is that one can always ensure a score of one-quarter, but there is an incentive to provide less diffuse reports, and that incentive is the possibility of a score of 1.

[21] All estimates are from fractional regression models and reflect average marginal effects. Unless otherwise stated, demographic controls are included because of variations in demographics across sessions and the fact that demographics affect performance, as we discuss later. We include controls for gender, age in years above 18, whether one self-declares as Black, whether one has a business major, whether one declares no religious affiliation (referred to by us as "Atheist"), and employment status. The classification "Black" refers to someone, resident in the United States, who self-reports "Black or African American" or "African" in response to the question, "Which of the following categories best describes you?" We report exact p-values.

[22] Karay et al. (2015) find comparable performance among advanced undergraduates, although there were differences in the time taken (less time on the computer) and the way in which poor-performing individuals made their mistakes (more random guesses on the computer). Hardcastle, Hermann-Abell, and DeBoer (2017) find no differences for high school students but significant differences for elementary school students and

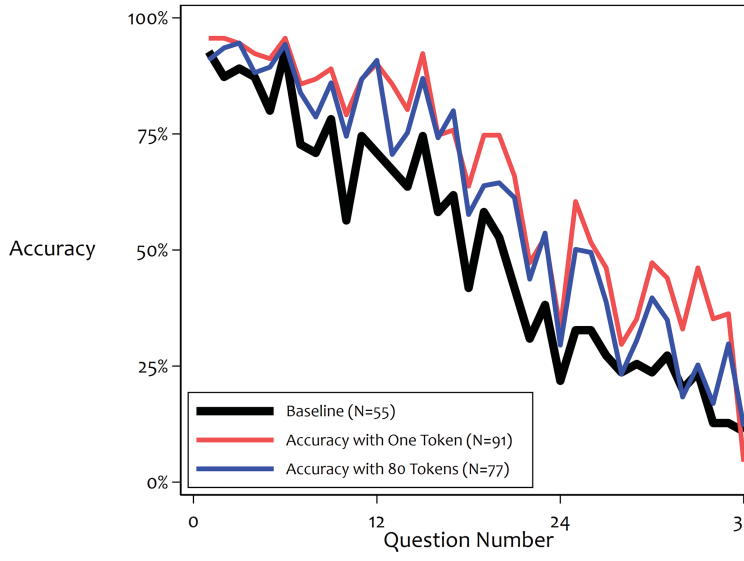


FIG. 5.—Average accuracy and incentives with progressive Raven problems. Accuracy is measured by the percent of tokens allocated to the correct answer. Beliefs are assumed to be the same as elicited reports.

Previous literature from psychology on the role of financial effects for the Raven task is surveyed by Gignac (2018). It does not view incentives, let alone financial incentives, the way that economists do. The older literature often relied on self-reported effort levels as a proxy for incentives to do well. And when financial rewards were used, they were always nonsalient: a small, fixed monetary amount for taking the test, irrespective of performance. A deeper presumption in the psychology literature is that expressions and productions of intelligence should be viewed as unaffected by rewards, whether intrinsic or extrinsic. We prefer to let data decide that simple question, rather than assume it.

Our result about the average effect of financial incentives complements the finding by Segal (2012) from a within-subject test using a coded speed test employed in the Armed Forces Qualifying Test, which gained notoriety as a measure of IQ due to Herrnstein and Murray (1994). Although this test is not a measure of fluid intelligence, as the term is normally used, it does measure the application of some cognitive abilities, as well as plain effort.[23] Her experimental design ingeniously allowed the

modest differences for middle school students. Although gender had no effect on the comparison, having English as a primary language did make some difference.

[23] There has been some unfortunate association of these "speed tests" with fluid intelligence in the economics literature. Heckman (1995, 1105) quotes Carroll (1993) as concluding that "there are three important sorts of cognitive abilities corresponding to fluid

separation of pure learning effects, due to repetition of the task three times, from the effects of salient incentives. And the within-subjects design allows her to metaphorically identify "Boy Scouts," who apply the same level of effort when told that it is a "test" as they do when provided piece-rate rewards for accuracy, and "Economists," who are motivated to apply effort only when there are explicit financial rewards. The average effect she identifies from financial incentives reflects the impact on the subsample of Economists.

The second result from figure 5, and perhaps something of a surprise, is that accuracy appears to be about the same when individuals are constrained to report their modal beliefs. In fact, accuracy is 3.7 percentage points higher when individuals are constrained to just report their modal belief, but this effect is not statistically significantly different from zero (p-value = .25). We did undertake a pilot experiment in which we progressively increased incentives for harder problems and observed statistically significant effects of incentives on accuracy in that case.[24] We say that the results in figure 5 may be a surprise, since our priors as economists tell us that constrained optimization should never generate better performance than unconstrained optimization.

Indeed, our priors are not violated if one looks at the right metric: in our incentivized setting, it is earnings that motivate, not accuracy as defined here. From figure 6, we see our third result, that earnings efficiency is indeed higher when individuals are not constrained to report their modal beliefs. The effect on earnings efficiency of the 80 Tokens Progressive condition compared with the One Token Progressive condition is +7.5 percentage points (p-value = .026). The bulk of the difference in earnings arises in the last dozen or so questions, where modal beliefs with high subjective probability of being true are scarce, but the One Token Progressive condition requires participants to report a modal belief. This result is exactly parallel to the familiar point in econometrics to explain why we care about likelihood values rather than "hit rates" when estimating (parametric) binary choice models: some mistakes matter more than others in terms of the metric that counts. Figure 6 shows,

---

intelligence (ability to solve problems quickly), crystallized intelligence (the ability to draw on old solutions to address new problems), and spatial and mechanical ability." This is incorrect: fluid intelligence has to do with the ability to solve novel problems. In fact, and most clearly in Carroll (1997), "general intelligence," called the g factor, is defined in terms of eight broad abilities: fluid intelligence, crystallized intelligence, general memory and learning, visual perception, auditory perception, retrieval ability, cognitive speediness, and processing speed.

[24] In app. D, we report results from the pilot experiment. We increased the earnings for allocating all 80 tokens to the correct solution from the default $2 for all questions to $2 for questions 1–12, $3 for questions 13–24, and $5 for questions 25–36. This increased potential aggregate earnings from $72 up to $120. There is a general improvement in accuracy and efficiency and marked improvements for Blacks and females.

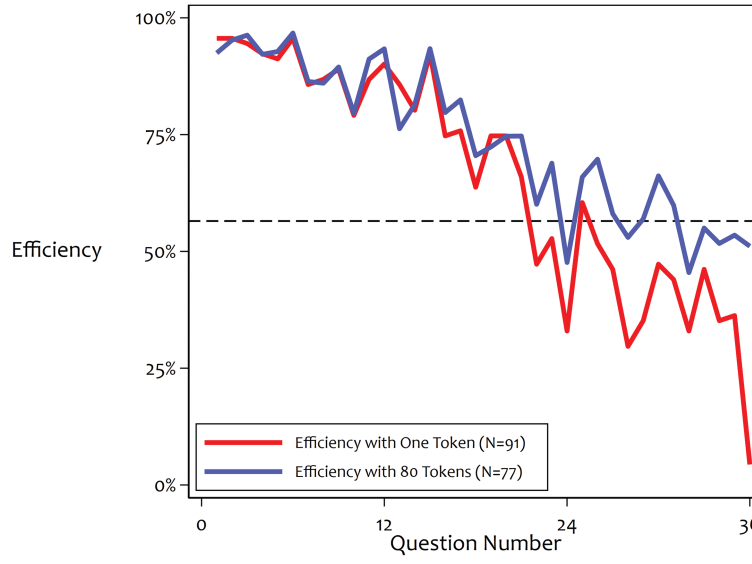


FIG. 6.—Average efficiency with progressive Raven problems. Efficiency is measured by the percent of realized earnings compared with maximum possible earnings. A uniform token allocation ensures earnings of $1.13 and efficiency of 56.5%. Beliefs are assumed to be the same as elicited reports.

by the horizontal dashed line, that a uniform report in the 80 Tokens Progressive condition generated an efficiency level of 56.5%.

B. Gender Effects

The fourth set of results have to do with demographic effects and, specifically, gender effects. Figure 7 displays these effects for accuracy and efficiency. These are average marginal effects, a point worth stressing since most previous research on gender and race effects has only looked at total effects. Both type of effect are valid, well-defined, and answer important questions, but they answer different questions.

The horizontal axes of figure 7 show the percentage point change in accuracy and efficiency, which themselves are measured as percentages. The horizontal lines, one for each demographic, show the point estimate and 95% confidence interval on that estimate.

We focus on the effects of demographics on accuracy first, in figure 7*A*. In the baseline condition, two effects stand out: the poorer performance of women and the significantly better performance of atheists. Notably, there is no significant effect for Blacks.

In figure 7*B* and figure 7*C*, we show the effect of the treatment condition on the accuracy or efficiency associated with each demographic. In

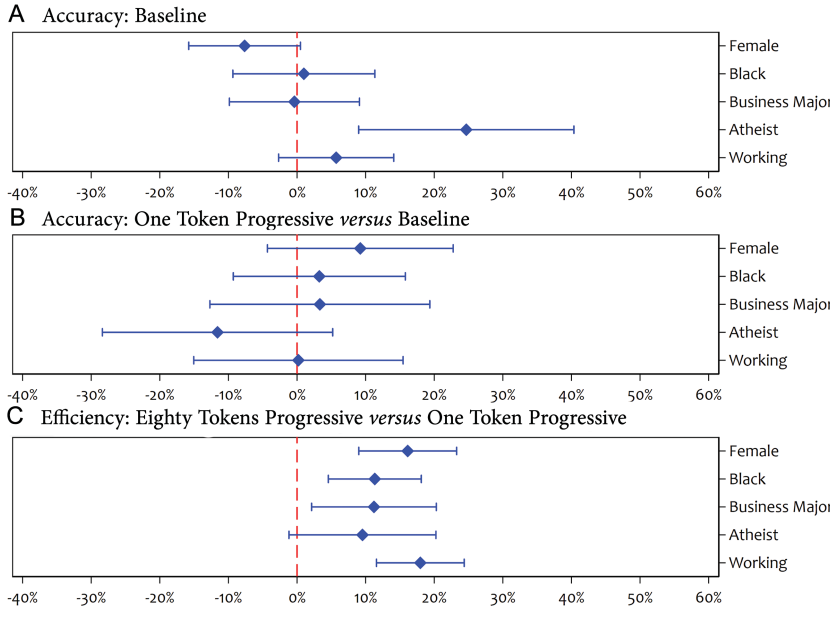


FIG. 7.—Effects of demographics with progressive Raven problems. All results used the original progressive order of the Raven problems. Average marginal effects are from fractional regression.

figure 7*B*, we continue to study accuracy and show the estimated effect from moving from the baseline condition to the One Token Progressive condition. Setting aside the computerization in the One Token Progressive condition, the treatment difference is the provision of financial incentives. When we add financial incentives but continue to constrain responses explicitly to the modal belief, we observe an improvement in accuracy for females, but it is not statistically significant at conventional levels. These results alert us to be wary of inferences about fluid intelligence with respect to gender and race that are based on measurements that rely on the hope that so-called intrinsic rewards exclusively motivate everyone participating in "intelligence tests." It bears stating clearly that this caution applies to a very large literature.

When we incentivize subjects to explicitly express their confidence, in the 80 Tokens Progressive condition, we observe that women do better than men, and Blacks do better than others. We also find that business majors do better, and those who work part-time or full-time do better. Figure 7*C* shows these effects on efficiency, comparing the 80 Tokens Progressive condition with the One Token Progressive condition. In this instance, compared with figure 7*B*, it is not just the use of financial incentives that matters but the ability to express confidence of belief.

Evans (2012) coined the notion of "risk intelligence" to capture the awareness of individuals of their nonextreme subjective beliefs and their willingness to express and act on them.[25] We do not seek to generalize beyond the population from which our sample was drawn, but the gender and race results are important.

In terms of accuracy, the first major review of gender differences in the Raven tests was by Court (1983). He concluded that there were no discernible differences, and this conclusion has been widely repeated in later references. However, Lynn and Irwing (2004) carefully note many limitations of the meta-analysis by Court (1983), not least that it was not based on any quantitative assessment of the evidence. They also offer their own meta-analysis, using more conventional measures and a more exhaustive literature review. They conclude that an advantage for males begins around the age of 15, and for adults 20 and over, it appears to be stable until the age of 79 at $+0.33d$, to use the conventional measure of "mean effects" in the psychology literature, Cohen's d.[26] The 95% confidence intervals on this estimate are between +0.28 and +0.37, reflecting pooled samples of 5,180 males and 4,451 females. We stress that these are total effects, not marginal effects after considering the effects of other demographics.

Our major insight is that women and Blacks benefit from having clear incentives and being able to explicitly express their confidence that some proposed solution is correct. We demonstrate that measurement tools that fail to facilitate incentivized expressions of confidence mischaracterize their fluid intelligence. The effects of allowing expressions of confidence are not at all a matter of women and Blacks being insufficiently confident as the problems become harder, but knowing when to *express* the appropriate level of confidence.[27] We stress the word "express," since someone might know that they are unsure but fail to be motivated to say so. Our design picks up both effects: knowing that one is

[25] His measure of risk intelligence used 50 hypothetical true/false survey questions (Evans 2012, app. 1).

[26] Cohen's d is just the difference in the averages for men and women, divided by the pooled sample standard deviation. Conventionally in this literature, a positive d is associated with better performance by men. The measures calculated by Lynn and Irwing (2004, table 2, 490) actually report corrected estimates of d and their 95% confidence intervals.

[27] We only ever use the expression "overconfidence" to refer to excessive certainty about the accuracy of one's belief. Harrison and Swarthout (2025) elicit subjective belief distributions about induced posterior distributions from observations of a sampling process and can therefore measure "appropriate" confidence with that Bayesian metric. Using samples from the same population considered here, they find general evidence for insufficient confidence, appropriate confidence, overconfidence, and overconfidence compared with the Bayesian metric after 1, 2, 3, and 4 rounds of accumulating data, respectively. Hence, evidence for overconfidence, in the sense that we use the expression here, can depend on what stage of Bayesian updating is being evaluated.

unsure and being prepared to say so. Hence, extrinsic incentives matter, as well as having a measurement tool that allows subjects to report confidence.

There is some value in examining the processes at work when subjects respond to the progressively more difficult problems and express their confidence in the correct answer. Figure 8 displays the allocation of tokens across all 36 RAPM problems, pooled over individuals in the 80 Tokens Progressive condition. We can readily identify at least three distinct phases of perceptions of the correct answer in the RAPM. One is a clear subjective sense of being able to identify the single correct answer with some high probability. The second is a subjective realization that these problems are becoming harder and that although there may be one single correct answer, and some possible answers are clearly wrong, it is not possible to state with much confidence which is the correct answer. The third is a subjective resignation that these problems are too hard and that it not even possible to identify some possible answers that are clearly wrong. In each case, we refer to the subjective sense, the subjective realization, and the subjective resignation: these are characterizations of the reports we observed, not of the validity of the subjective beliefs.

Figure 9 illustrates these distinct types of responses for Raven question 20, an ideal stage in terms of the progressive difficulty of the battery starting to bite. The token allocations for each of the first 20 individuals are displayed, and the overall pattern is clear. The top row alone shows one subject, number 2, who allocated all 80 tokens to the correct answer; one subject, number 1, who was overconfident with respect to a wrong answer, allocating all 80 tokens to answer 2; one subject, number 5, who had bimodal beliefs and allocated 40 tokens to each of two answers, one of which was correct; and two subjects, numbers 3 and 4, who reported completely diffuse beliefs.

Figure 10*A* provides a descriptive classification of the time trends of three types of allocation. One is a Certain allocation of all 80 tokens to one solution, whether or not it is the correct solution. Another is a Diffuse allocation of 10 tokens to each of the eight possible solutions. And the Unsure allocation is anything in-between the two extremes of Certain and Diffuse. As the share of beliefs that are Certain declines, there is a roughly equal increase in the other two types until the final 10 questions, at which point the Diffuse type dominates.

This variation in the type of allocation, when subjective confidence can reasonably vary from problem to problem, is important for performance on the RAPM. In other words, a key feature of good performance on the incentivized RAPM is "to know when to hold 'em and know when to fold 'em," referring colloquially to the skill of a successful poker player knowing when to stay in or drop out of a hand. Statistical analyses of the between-subjects determinants of better efficiency in the 80 Tokens

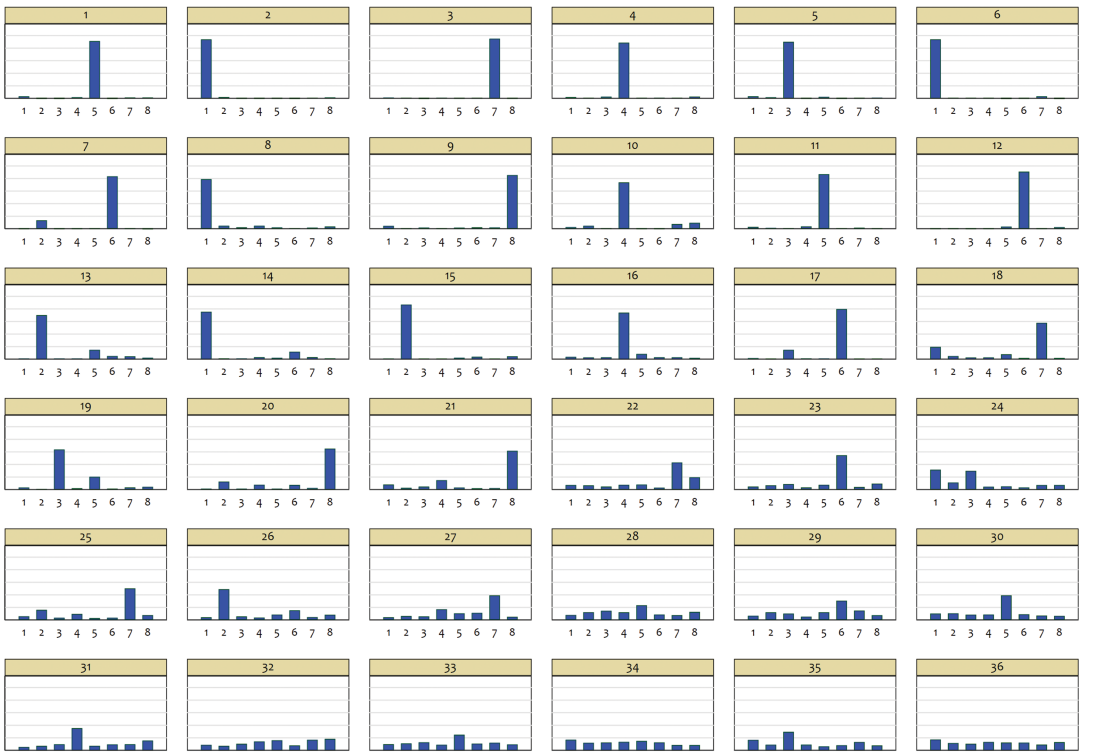


FIG. 8.—Allocation of tokens by Raven question, pooled over incentivized responses in the 80 Tokens Progressive condition.

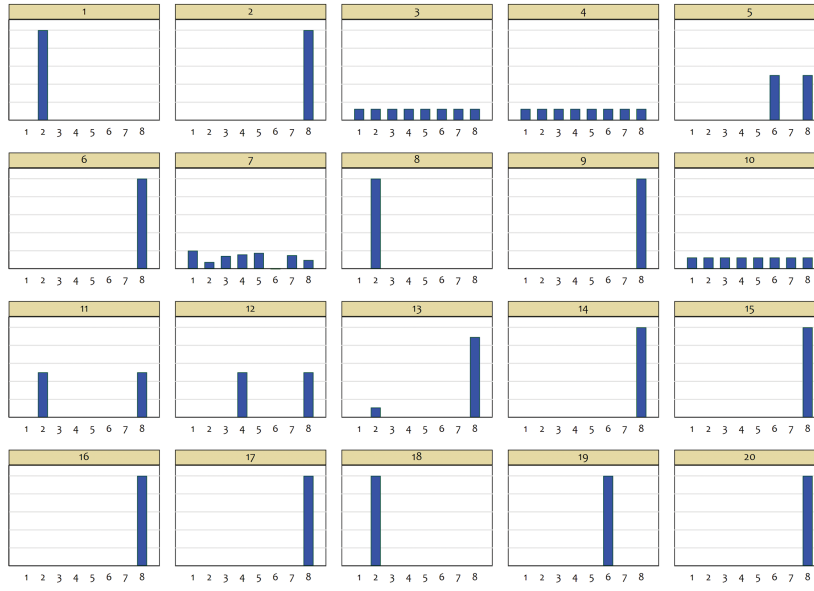


FIG. 9.—Allocation of tokens in Raven question 20. Incentivized responses in the 80 Tokens Progressive condition. The correct answer is 8.

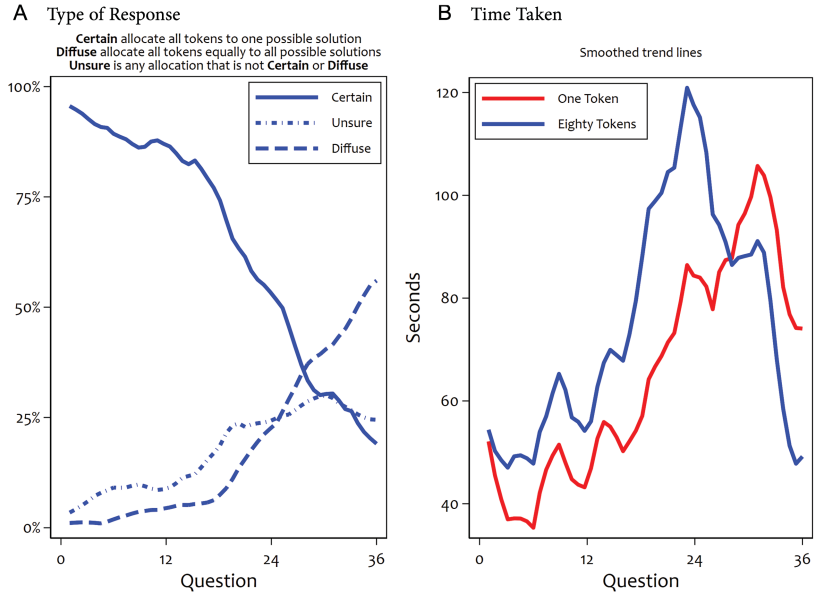


FIG. 10.—Time trends of time for response and for type of response.

Progressive condition compared with the One Token Progressive condition confirm that we see significant differences in the application of this skill across gender and race, for example. Although one might be tempted to view this as a (valuable) meta-cognitive skill, we view it as a core part of cognition and intelligence itself when faced with risks.

A related observation concerns the amount of time spent on each question. Figure 10*B* demonstrates the effect of incentive structure on the use of this input to the cognitive production process. Roughly two-thirds of the way into the set of progressively harder questions, subjects in the 80 Tokens Progressive condition significantly and steadily reduced their time allocation, consistent in figure 10*A* with increased use of the Diffuse allocation. On the other hand, and by comparison, figure 10*B* shows greater use of time in the One Token Progressive condition until much further into the set of harder questions. In this financial setting, the Diffuse allocation option is not available, and the rewards were then "all or nothing."[28] This pattern of responsiveness to time is evidence that subjects responded to incentives to report confidence in the manner anticipated by our design.

C. The Cognitive Scaffold of Progression

The discussion of striking gender and race effects in our standard presentation of the Raven Progressive Matrices task points to three dimensions of cognitive performance, suggested by our prior discussion of the temporal phases of deliberations. One dimension is the intrinsic difficulty of solving any one of the 36 problems, taken alone. A second dimension is the willingness to report imprecision of beliefs about the correct solution. And the third dimension is that the problems are presented in an ordered sequence, from easiest to hardest, and the evidence from accuracy and efficiency reflects this general pattern, as seen immediately in figures 5 and 6. In effect, the progressive presentation of problems could serve as a valid clue to cognition, quite apart from the first two dimensions of performance. It is perfectly natural that someone would exploit such a cognitive

[28] The patterns in fig. 6 and fig. 10 allow one to draw inferences about earnings per minute of time allocated to each question. Although fig. 6 shows realized earnings normalized by maximum earnings, those maximum earnings were the same across the One Token Progressive and 80 Tokens Progressive treatments, so efficiency is just a normalization of earnings. Earnings per minute in the One Token Progressive treatment were actually slightly larger than in the 80 Tokens Progressive treatment for the first third of questions. Of course, for risk-averse individuals, a slightly higher average earning is not per se attractive since it came with a much higher standard deviation of earnings. Then, in the middle third of questions, the average earnings per minute of the 80 Token Progressive treatment steadily became greater than for the One Token Progressive treatment. And this difference increased significantly for the last third of questions, reflecting the differential time allocations shown in fig. 10 for these questions.

scaffold, but clearly one would like to have a measure of fluid intelligence that did not assume that life's cognitive challenges are neatly arrayed from easiest to hardest.

An obvious implication of this concern is to evaluate performance in a scrambled version of the Raven task. The presentation and incentives are the same, with some obvious changes.[29] The attractive feature of this variant of the Raven task is that the individual should, ideally, engage in some meta-cognitive effort to identify the difficulty of the task, problem by problem. The difficulty of solving problem τ in this variant provides no added information about the difficulty of solving problem $\tau + 1$, other than insight into one own's ability to solve this general class of problems.

Figure 11 shows the results of applying this scrambled variant in the One Token Scrambled and 80 Tokens Scrambled treatments. We focus on efficiency; results on accuracy show a predictable drop compared with the standard progressive presentation. Figure 11 shows the same qualitative path as figure 6, with average performance arrayed by question in the original progressive order. Efficiency in the 80 Tokens scrambled treatment again converges on the diffuse report shown with a dashed, horizontal line. Efficiency in the One Token Scrambled treatment falls well below the 80 Tokens Scrambled treatment earlier in the original progressive order, around question 20 rather than around question 25.

We find much the same demographic effects on efficiency as with the progressive presentation of problems. Figure 12 shows these effects, in comparison with figure 7*C* and with the addition of measures of personality traits.[30] Again, we see a significant effect for women and Blacks, confirming that the measurement of fluid intelligence does not derive from their recognition of the progressive scaffold.

D. The Welfare Cost of Traditional Procedures

Our results allow an evaluation of the value of allowing individuals to reveal their confidence in answers to the Raven questions. We evaluate the

[29] The changes are shown in app. A, sec. C.

[30] These are from the 40-item Big Five Personality Inventory described by John and Srivastava (1999). Responses to these items are used to generate scores for five personality traits of the individual. One trait measures extraversion: whether someone is outgoing at one extreme or reserved at the other. Another trait measures the person's agreeableness: whether they are friendly toward others or tend to be quarrelsome and critical of others, at either extreme. Another trait, unfortunately called neuroticism, measures how nervous or resilient someone is when confronted with challenges. Another trait measures how conscientious someone is about tasks: whether they are organized at one extreme or careless at the other extreme. And a final trait measures openness to experience: whether they are imaginative, curious, and open to a wide range of interests. For each subject, we reduce the scores on each of these traits to a binary indicator, split as close to the median sample response as possible. These survey items were included in our 2022 experiments.

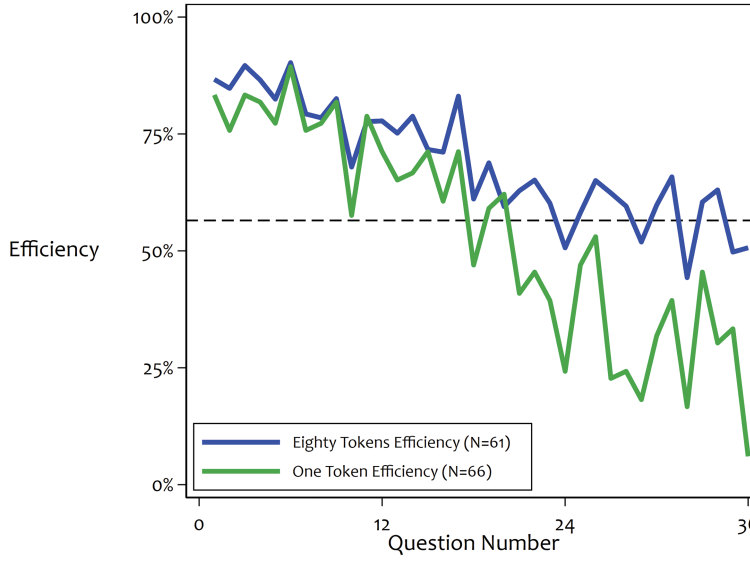


FIG. 11.—Average efficiency with scrambled Raven problems. Efficiency measured by the percent of realized earnings compared with maximum possible earnings. A uniform token allocation ensures earnings of $1.13 and efficiency of 56.5%. Question number is from the original progressive order. Beliefs are assumed to be the same as elicited reports.

expected welfare cost to a respondent of being forced to report their subjective beliefs in a way that requires that they report only their modal belief. In effect, this is the welfare cost to respondents of forcing them to use our two One Token instruments instead of the corresponding 80 Token instruments. This calculation uses the recovered subjective beliefs of each subject in each question, the QSR payoffs implied by their token allocation, and their risk preferences, to evaluate the risky lottery posed to each subject by such an intervention.[31] Using this information, we can calculate the certainty equivalent (CE) to the subject, and then report that as a percent of the CE of using the unconstrained instrument. The default CE is when the subject is allowed to allocate beliefs across all eight alternatives; the intervention CE assumes that the subject is forced to report that her beliefs are concentrated on her true modal belief. By construction, the welfare cost of the intervention is nonnegative, but of course it may be zero when the subject's beliefs are completely concentrated on one alternative. We do not assume in this calculation of welfare cost to the respondent that she is correct in the default reporting of beliefs, or even that she is unbiased.

[31] The procedures to recover beliefs are reviewed in sec. IV.D and app. B.

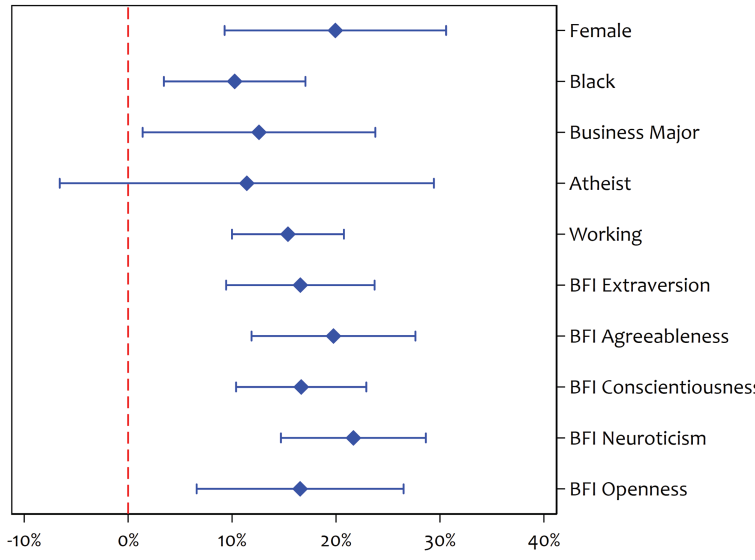


FIG. 12.—Effects of 80 tokens on efficiency with scrambled Raven problems. Average marginal effects from fractional regression, pooling 80 token and one token results, with a dummy variable for the 80 token condition and solely comparing results within the scrambled treatment.

We descriptively assume rank-dependent utility (RDU) instead of EUT as the model of individual risk preferences, since it allows for the empirical probabilities to be replaced by decision weights that reflect probability weighting. These calculations also form the basis for a normative evaluation of the observed choices, although some might want to use EUT for that purpose rather than RDU.[32] As it happens, there is little empirical difference in this setting between using EUT or RDU.

The resulting pattern of welfare costs rises with the original progressive question number. Figure 13*A* shows this pattern,[33] where the vertical axis ranges from \$0 to \$2 since these were the rewards at stake for responses to each question. As expected a priori, the cost is quite low for the first few questions, which are relatively easy to answer. For the progressive case, the cost starts at around \$0 and stays low for the first 15–20 question; for the scrambled case, the cost starts at around \$0.25 and stays there until around question 10. In both cases, the cost becomes much higher for the last set of questions, converging steadily to around \$0.80 for question 36. The scaffold provided by the ordering of problems by difficulty in the

[32] We report normative evaluations with EUT in app. B, fig. B13. The relative merits of EUT or RDU as a normative metric are discussed by Harrison and Ross (2023).

[33] Displayed using fractional polynomials with 95% confidence interval bands.

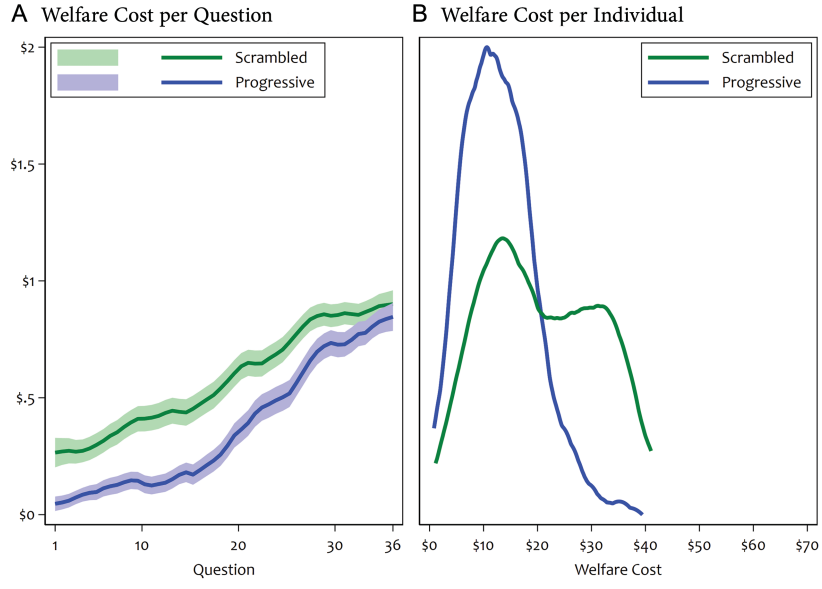


FIG. 13.—Expected welfare cost to respondent of being required to report only modal belief, evaluated using RDU risk preferences for each subject and using recovered beliefs from the 80 tokens task.

progressive treatment explains why the welfare cost is lower in that treatment for earlier questions. Average welfare losses per question are \$0.20 = \$0.58 − \$0.38 higher with the scrambled treatment, but about \$0.25 for the first 20 progressive questions and only \$0.08 for the last six questions.

We can also examine the welfare cost for each individual, rather than by individual question. Since we expect intelligence, however measured, to be an individual trait, this is a natural metric for welfare cost. Figure 13*B* displays kernel densities of these distributions. The bottom axis range reflects the potential earnings in this task between \$0 and \$72, to provide context on the significance of the welfare losses. Again we observe that the welfare losses of being forced to report only one's subjective modal belief as if it was the single true answer are more costly under the scrambled treatment. Average welfare losses per person are \$7.40 = \$20.88 − \$13.48 higher with the scrambled treatment. There is also an interesting mode just above \$10 for the welfare cost under the scrambled treatment, which is close to the mode under the progressive treatment. This suggests that there are some individuals who would not identify any cognitive gain from using the scaffold of progression, even when offered.

Figure 13*A* and figure 13*B* provide insight into the willingness of individuals to pay for access to scaffolds to aid cognition. First, over all

36 questions, individuals would be willing to pay up $7.40 to have the questions ordered in progression of difficulty, at least as difficulty is determined by the developers of the battery.[34] Second, there may be some significant fraction of individuals who would not see any value in a scaffold and not be willing to pay for it. Third, the value of a scaffold depends on the "cognitive context" of its use: if life is only full of questions like the devilishly difficult question 36, why pay for a scaffold? But if life is indeed like a box of cognitive chocolates, to paraphrase Forrest Gump, and you never know which of the 36 questions you are going to get, then the scaffold pays off.

III. Gender and Confidence, Reconsidered

Our striking result that women respond positively in terms of our measure of fluid intelligence leads us to consider whether this finding is more general. In the nature of "fluid" intelligence that is not dependent on one's "crystallized intelligence" about one context or another, one would a priori expect that it would generalize. We therefore reexamine two celebrated instances in which gender and confidence have been studied: willingness to engage in "competitive" risky behavior and measures of financial literacy.

One point of clarification is the simple use of the word "confidence" to mean two things. It is often used in these literatures in the colloquial sense of optimism ("overconfidence") or pessimism ("insufficient confidence"). We use the term to refer to the inverse of precision, the variance of beliefs over alternatives and degrees of subjective belief over alternatives. The two are each valid uses of the word confidence but must be kept clearly distinct. This point is well-known but often seems to be equally well forgotten; see Moore and Healy (2008).

The other point of clarification is that, in both cases, we find that thinking about how one measures performance, whether it be with an intelligence test, a measure of "competitiveness," or measure of literacy, makes a difference in terms of the welfare significance of the findings. Although the immediate focus is on gender, the methodological lessons generalize to other demographics and personality characteristics. We return to this deeper, normative reason for paying more attention to confidence, in the sense in which we use the term, in our concluding remarks in section V.

[34] This last qualification raises the interesting problem of identifying "difficulty" when there are several attributes that might be useful to produce a good cognitive response. For example, certain arithmetic tasks might be easy if one has excellent short-term memory (or pen and paper) and hard otherwise.

A. *Gender and Competitiveness*

Niederle and Vesterlund (2007) consider the effect of more or less competitive compensation schemes on the incentives that lead women to participate in a workforce characterized by those schemes. At one extreme is a piece-rate system, in which a fixed payoff is provided for every unit produced in a given period of time. At the other extreme is a tournament, in which only one in four participants will receive a positive payoff, based on having the greatest number of units produced in a given time period. The setting is one in which the production process requires real cognitive effort over some period of time. They observe from laboratory experiments that women tend to avoid tournaments in favor of piece rates, when asked to choose one or the other to be compensated, even when the production task has been selected to be one where men and women have roughly the same ability.[35] This tendency is regarded as a bad thing, as reflected by the first part of the title "Do Women Shy Away from Competition?" and myriad references in a large, subsequent literature to women not being willing to compete.

Our results with the Raven Progressive Matrices task shows that it is important to consider the metric of performance, even when some objectively correct answer exists, to allow for the fact that most of the time, in the field, we are concerned with belief assessments of alternative possible solutions. Only in a rare setting does the correct solution present itself in all instances. Of course, we find that if one uses one metric for performance, the number correctly answered, a very different measure of performance is obtained for women compared with men than if one uses a metric of performance that recognizes the cognitive difficulty of always being able to identify the single correct answer.

The same might be true of the tasks posed by Niederle and Vesterlund (2007) in the face of the time pressure of only having 5 minutes to come up with the largest number of produced solutions. And this might be exacerbated by the tension of competing against somebody else. Hence, there is value in examining the conclusions of Niederle and Vesterlund (2007) in the same manner, by asking whether they are evaluating men and women with the appropriate metric. We say "appropriate" and intend the word in both descriptive and normative senses. When men and women make a choice over piece rates or tournaments, they are evaluating two risky lotteries, where the risks are subjective.

[35] There are many settings in which the ability to complete more "production tasks" in a given period of time might vary by gender and the compensation scheme employed. Gneezy, Niederle, and Rustichini (2003) consider a puzzle consisting of mazes of varying difficulty as the production task. They find that men and women do solve roughly the same number of mazes under piece rates but that men solve many more than women under tournaments. The task in Niederle and Vesterlund (2007) was deliberately chosen to avoid this gender difference in actual ability and consists of adding up five 2-digit numbers correctly.

Our experimental design mimics Niederle and Vesterlund (2007), which has several elegant features to allow a decomposition of the determinants of observed behavior, well beyond the "reduced form" conclusions drawn from it about the apparent lack of competitiveness of women. There are six parts to our design. In part 1, the subject is asked to solve up to 12 scrambled Raven questions in 10 minutes and is rewarded by points. Each QSR allowed subjects to earn up to 200 points in each question if all tokens were allocated to the correct response. If this part was selected for payment, their reward would be based on their total earnings of points, multiplied by \$0.01, so each point was worth a penny, and subjects could earn up to \$2 per question, or \$24 for the 12 questions. This is the piece-rate compensation treatment. In part 2, the underlying cognitive production task was the same, with a different 12 scrambled Raven questions in 10 minutes. However, the compensation scheme was based on a tournament in which the point earnings of the subject would be compared with the point earnings of three other subjects selected at random. Only the subject with the best score would receive compensation if this part was selected for payment,[36] but the potential payoff was four times the payoff from the piece-rate part 1, so \$8 per question, or \$96 for the 12 questions.

In part 3, the focus is on the choice of compensation scheme. The subject is told that they will face yet another 12 scrambled Raven questions in 10 minutes. Without knowing their earnings, or the earnings of anyone else from parts 1 and 2, the subject must decide whether to have their earnings in part 3 determined by the piece-rate scheme of part 1 or the tournament scheme of part 2. If the tournament is selected, the new point earnings of the subject will be compared with the point earnings from part 2 of the other three members of her group. This is a key point of the design.

Assume for simplicity that the subject believes in part 3 that they will hit their previous piece-rate score from part 1 if they adopt the piece-rate scheme.[37] Payoffs are then equal to that expected number times \$0.01 per point, no matter what the number of points might be. Expectations for the tournament are nearly the same but with one twist. The individual knows what her own effort level was under the tournament conditions in part 2 but does not know the tournament outcome. The individual must additionally form some subjective belief that her earnings will

[36] Subjects were told that ties would be resolved randomly.

[37] In effect, this just assumes that the individual expects a data generating process for their efforts that has symmetric noise around their perceived previous outcome and then applies reduction of compound lotteries (ROCL) to that distribution. Under ROCL, they can have some subjective beliefs about how many they will complete but can "replace" that probability mass function with the expectation, which by construction here is their perceived previous outcome.

be the highest of the four in her tournament group. A natural prior would be to assume a one-quarter chance of being the winner of the tournament.[38] We come back to the actual subjective beliefs that the subject has in a later part but assume this diffuse prior for now.

Now just apply some simple economics to the choice of compensation scheme, taking into account the risk of winning or losing the tournament. Since the tournament gives a three-quarters chance of receiving \$0 and a one-quarter chance of winning a positive prize, it is much more risky than the piece-rate option if effort and performance in the cognitive task for the subject are the same under either compensation scheme. The experimental design of Niederle and Vesterlund (2007), indeed, selected \$2 to be exactly four times \$0.50, under this null hypothesis about empirical expectations on the chance of winning the tournament.[39] Thus one would expect all risk-averse agents to want to avoid the tournament, because the parameters of the experiment were designed to make a risk-neutral agent indifferent.

Assume everyone is risk averse, to varying degrees; we check this assumption, but it is a weak one. Why, then, are we concerned about the behavior of women if they make the decision to avoid the risky lottery that offers about the same earnings but with higher variance? Surely the problem here is that men tend to make the wrong risk-management decision, not that women tend to be too timid to compete.[40]

It is a simple matter to apply some EUT finger mathematics to the actual data from Niederle and Vesterlund (2007) to be more precise about these welfare effects. Appendix C details these calculations. Using of welfare gain or loss again, we find that the women who chose the piece-rate scheme gained in welfare terms by +\$3.94; the women who chose the tournament scheme lost in welfare terms by \$3.65; the men who chose the piece-rate scheme gained in welfare terms by +\$3.01; and the men who chose the tournament scheme lost in welfare terms by \$3.01. Hence, it is important to recognize that women tended to do the right thing, assuming that EUT describes their risk preferences, by selecting the tournament only 35% of the time. But men tended to do the wrong thing, by selecting the tournament 73% of the time. In principle, these calculations

[38] If indeed the individual has equal ability to others, and the design of Niederle and Vesterlund (2007) was built around a cognitive task in which the average man was expected to have the same ability as the average woman, then the probability of winning the tournament is one-quarter by design.

[39] One second-order empirical violation of the ceteris paribus assumption is that both men and women did ever so slightly better in their part 2 tournament production than in their part 1 piece-rate production, possibly due to learning effects from repetition. We can numerically take these statistically insignificant differences into account, with no change in the general conclusion.

[40] This inference has nothing to do with the claim that women are more risk averse than men. In the sequel, we let data fill in the risk preferences of individuals.

can be repeated at the level of the individual, and we do that using our own experimental data below. We can again also descriptively assume RDU instead of EUT, which allows for the empirical probabilities to be replaced by decision weights that reflect probability weighting. These calculations also form the basis for a normative evaluation of the observed choices.

This inference, that the simple economics of the choice task indicate that men, not women, have appeared to have made the wrong decision, is before we see the results of eliciting subjective beliefs. Those results confirm the reason for these wrong decisions in the experiments of Niederle and Vesterlund (2007): poorly calibrated subjective beliefs by men. In their experiments, women have reasonably well-calibrated subjective beliefs and just made the right decision based on them.

There are three final, brief steps to the experimental design, but they play a surprising role in helping evaluate observed behavior. In part 4, the subject decides whether she wants her performance in part 1, the initial piece-rate task, to be evaluated with the piece-rate compensation scheme or the tournament compensation scheme. This choice of compensation schemes differs from part 3, since it used the previously generated performance of the subject, rather than adding the "stress" of competing. Hence, part 4 elegantly teases apart two attributes of competition: being asked to *perform* some task, knowing that the earnings will depend on doing it better than others, and being *compared* with others to generate a binary payoff for the best performance. The second attribute involves a simple risk: the subjective probability that your score will be better than scores of others in your group. The first attribute might or might not add some risk: the variability of performance could depend on knowing ex ante that one is performing the task under these competitive conditions. This subjective risk is separate from the preference the individual might have for such head-to-head performance. Teasing these apart is exactly what a comparison of behavior from parts 3 and 4 does.[41] So part 4 controls for any stress from having to perform but leaves in place the risky lottery involved if the subject selects to be evaluated by a tournament.[42]

[41] Shurchkov (2012) recognizes that there are these additional attributes from competition but only replicates parts 1–3 of the design of Niederle and Vesterlund (2007). One of the attributes she considers is time pressure, which is a factor in our replication and extension of Niederle and Vesterlund (2007) but was not effectively a factor in our main experiments with the Raven task.

[42] Niederle and Vesterlund (2007) present part 4 as controlling for risk aversion, but it does not. Niederle (2015, sec. II.A) explains it in this manner:

> *Explanation 3: Gender differences in risk and feedback aversion*: These are dimensions different from taste for competition that also impact the choice between a tournament and a piece-rate incentive scheme. The tournament payment scheme not only is competitive, it is also more uncertain and provides more information about relative performance than the piece-rate scheme. For both risk aversion as well as preferences over receiving feedback about relative performance, there may be gender

The final steps of the experimental design elicit subjective beliefs that drive the choices in parts 3 and 4. In part 5, the subject reports beliefs about the rank of her piece-rate earnings in part 1 compared with the others in her tournament group of four. We use a QSR with four possible responses and provide 100 tokens for the subject to allocate. The subject could earn up to $30 by allocating all 100 tokens to the correct response. Finally, in part 6, the subject again reports beliefs about her point earnings rank, in this case with respect to the tournament earnings in part 2. The QSR setting and incentives are the same as in part 5. Of course, the response from part 6 provides exactly the individual subjective probability of winning the tournament that is needed to evaluate the choice in part 3 descriptively and normatively; we do not need to assume a diffuse probability of one-quarter of winning.[43]

As flagged above, following the discussion of part 3, the subjective beliefs elicited in part 6 surely play a crucial role in evaluating behavior. Niederle and Vesterlund (2010, 134) summarize their data as follows:

> Accounting for ties, at most 30% of men and women should guess that they are the best in their group of four. We find that 75% of men compared to 43% of women guessed that they were the best. While both men and women are overconfident, men are more overconfident than women.

differences. *Design solution*: Instead of directly controlling for risk and feedback aversion, participants make a decision between two incentive schemes which mimic both the uncertainty in payment and the provision of feedback without any actual competition taking place.

Feedback aversion, as defined here, is a part of what we call stress. But the subjective risk of winning a big prize or a low prize if the tournament is selected remains for part 4, so part 4 does nothing to control for risk aversion.

[43] The belief elicitation step is not in the written instructions for the experiment of Niederle and Vesterlund (2007) but is embedded in Ztree code available from Lise Vesterlund (private communication, May 2021). It asked the subject to "guess" their rank in part 2 or part 1, and pays $1 if the subject is correct or incorrect. Chen et al. (2024) note that beliefs elicited over past tournament rank are not the same as beliefs elicited over future tournament rank and that it is the latter that matter for the choices made in parts 3 and 4. On the other hand, eliciting beliefs over outcomes depending on actions by the agent herself generates issues of incentive compatibility: "Do I forecast a low rank and make sure I get that belief elicitation reward by tanking in the tournament?" Recognizing this, Chen et al. (2024, 16) infer beliefs over a future "lump-sum tournament" with payoffs that depend on winning but not on performance: "While we cannot construct informative identified sets on the subjective probability of winning a future tournament, we can on the subjective probability of winning a future lumpsum tournament. We suspect that these subjective probabilities are equal since they have the same condition for winning." We are not so sure that these are comparable tournaments, theoretically or behaviorally. In addition, Chen et al. (2024, 6) must assume risk neutrality in order to infer subjective beliefs; their own incentivized test for risk preferences shows risk aversion on average, as expected.

Recognizing that the word "overconfident" is used here to mean optimistic, this finding is crucial. The full title of Niederle and Vesterlund's (2007) paper is "Do Women Shy Away from Competition? Do Men Compete Too Much?" Surely these aggregate data say clearly that it is the latter, not the former.

What is needed is direct economics: a calculation of the welfare gains or losses to individuals given their subjective beliefs about their chances of being ranked best in the tournament, their expected payoffs given the possible outcomes, their risk preferences, and their choices in parts 3 and 4. We had 88 subjects undertake this task—45 with the 80 Token Scrambled treatment and 43 with the One Token Scrambled treatment. Table 2 lays out the calculations, focusing on the decision to compete in part 3. The calculations show the welfare effects for a risk-neutral individual for reference since that was assumed in the design of the tournament payoffs. The calculations then show the welfare effects for a risk-averse individual with EUT risk preferences and finally for a risk-averse individual with RDU risk preferences.

The numbers in panel A of table 2 are the average payoff in parts 1 and 2. The payoffs for the "Tournament win" row are the realized payoffs, which get paired in a lottery with some probability of getting zero. The numbers in panel B are the subjective probabilities of receiving each payoff. The piece-rate payoff is certain, of course, and the probability of a tournament win comes directly from part 6 of the experiment. The risk-preference parameters in panel C of table 2 are from estimated EUT or RDU models, both using a familiar Constant Relative Risk Aversion utility function $U(x) = x^{1-r}/(1 - r)$; the RDU model also assumes a Prelec (1998) probability weighting function (PWF).[44] Every subject completed a risk battery of 30 binary choices, selected at random from a larger battery of 67 gain-frame lottery choices from Harrison and Swarthout (2023). Although table 2 reports averages for men and women, subsequent calculations at the individual level use individual risk-preference parameter estimates from the Bayesian hierarchical model of Gao, Harrison, and Tchernis (2023) and confirm the general conclusions.[45]

Armed with this information, it is just a matter of finger mathematics to calculate the expected utility (EU) or RDU of the piece-rate lottery choice and the tournament lottery choice. The piece-rate lottery is a degenerate lottery, with the payoffs listed being received with certainty; the

[44] This probability weighting function exhibits considerable flexibility and is $\omega(p) = \exp\{-\eta(-\ln p)^{\varphi}\}$, defined for $0 < p \leq 1$, $\eta > 0$ and $\varphi > 0$. When $\eta = \varphi = 1$, $\omega(p) = p$, so there is no probability weighting.

[45] The average subject exhibited moderate risk aversion under EUT and RDU. The average RDU subject has a more concave utility function than under EUT and has a globally optimistic probability weighting function to offset that. The average RDU parameters are quite close to EUT, but this is not true at the individual level.

TABLE 2
WELFARE EVALUATION OF WILLINGNESS-TO-COMPETE DECISIONS

	Risk Neutral		EUT		RDU	
	Male	Female	Male	Female	Male	Female
	A. Payoffs for 80 Tokens					
Piece rate ($)	1.37	1.36	1.37	1.36	1.37	1.36
Tournament win ($)	5.66	5.72	5.66	5.72	5.66	5.72
Tournament lose ($)	0	0	0	0	0	0
	B. Probabilities for 80 Tokens					
Piece rate	1	1	1	1	1	1
Tournament win	.26	.28	.26	.28	.26	.28
Tournament lose	.74	.72	.74	.72	.74	.72
	C. Risk-Preference Parameters					
Utility parameter r	0	0	.6	.56	.71	.7
PWF parameter η	1	1	1	1	.97	.9
PWF parameter φ	1	1	1	1	1	.92
	D. Certainty Equivalents for 80 Tokens					
Piece rate ($)	1.37	1.36	1.37	1.36	1.37	1.36
Tournament ($)	1.47	1.60	.87	.98	.09	.18
	E. Expected Welfare Gain for 80 Tokens					
Piece-rate compensation ($)	−.10	−.24	+.50	+.38	+1.28	+1.18
Tournament compensation ($)	+.10	+.24	−.50	−.38	−1.28	−1.18
	F. Expected Welfare Gains for One Token					
Piece-rate compensation ($)	+.11	+.20	+.51	+.62	+.99	+1.09
Tournament compensation ($)	−.11	−.20	−.51	−.62	−.99	−1.09

NOTE.—Parameter values for payoffs, subjective probabilities, and risk preferences are documented in app. C.

tournament lottery is a well-defined simple lottery. Once these have been calculated, the CE of each lottery can be calculated as shown in panel D of table 2, since $u(\text{CE}) = \xi$ solves for $\text{CE} = [\xi \times (1 - r)]^{1/(1-r)}$ for $\xi \in \{\text{EU}, \text{RDU}\}$. If subjects are risk neutral, the CE for the piece-rate and tournament lotteries for men is $1.37 and $1.47, respectively, and $1.36 and $1.60 for women. So risk-neutral men and women should (just) choose the piece-rate option in part 3. The normative "should" here comes from the subjective beliefs and subjective risk preferences of the subjects. Panel D also shows the welfare calculations if risk preferences are characterized using EUT or RDU. Since both EUT and RDU show risk aversion, the piece-rate choice looks to be the best for both men and women.[46] In panel E, we show the expected welfare gain from selecting one or other compensation scheme, with the welfare gain

[46] The parameter values for η and φ show probability optimism for both men and women. Hence, the better (poorer) tournament lottery outcome is weighted more (less), ceteris

choice with positive values. Finally, panel F repeats the same calculations using the one token data, implicitly replacing the data shown in panels A and B. For the one token case, even risk-neutral men and women should choose the piece rate and avoid the tournament.

These calculations can be undertaken at the level of the individual. Figure 14 shows distributions of the welfare effect if each individual chose to compete in the two settings in which they have that choice available. These are ex ante welfare effects of making the choice to compete, whether or not that choice is justified by the simple economics of evaluating the risky options presented. A clear pattern emerges, consistent with table 2: the vast majority of individuals should never choose to compete. Somewhat more should compete under RDU risk preferences than EUT but never more than a fifth of the sample.

By construction, there would be no ex post welfare losses if every individual chose consistently with the ex ante welfare effect. If those with a negative welfare effect in figure 14 chose not to compete and those with a positive welfare effect in figure 15 chose to compete, we would observe nothing but welfare gains from the data. The size of the gain would depend on payoffs under the piece-rate and tournament conditions, subjective probabilities of being the top rank in the tournament, and individual risk preferences. Of course, from figure 14, we see that the vast majority of individuals should have chosen not to compete in the tournament, as expected from the simple economics of the risky choice. However, figure 15 shows the extent of the departure from this ideal, with significant welfare losses for some individuals, since quite a few chose to compete when their risk preferences and subjective risk perceptions indicated that they should not.

What is driving these losses? In our data we observe virtually identical average payoffs for men and women in the piece-rate and tournament tasks in parts 1 and 2, whether these were one token or 80 token conditions. Certainly there are some men and some women who score better than others, but performance in the task is not generally different for men and women. We find that women are slightly more optimistic about their chances of winning the tournament than men when there were 80 tokens and about the same when there was only one token. Figure 16 displays the elicited beliefs for all ranks, by men and women, in each of the conditions. If we just focus on beliefs about winning the tournament,[47] we find that men and women are slightly pessimistic in comparison to the diffuse prior of one-quarter when there was only one token. On average,

paribus generating risk-loving behavior. But the ceteris paribus assumption is violated, with modestly concave utility functions, so the overall effect is modest risk aversion.

[47] It is also worth noting the "Lake Wobegon" effect, in which everybody believes that there is only a very low chance that they will come in last.

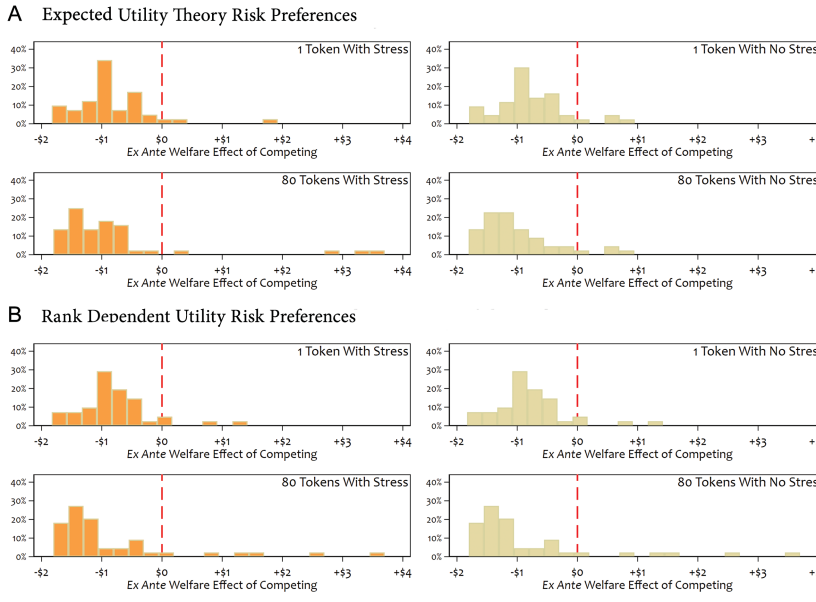


FIG. 14.—Ex ante welfare effects of the decision to compete.

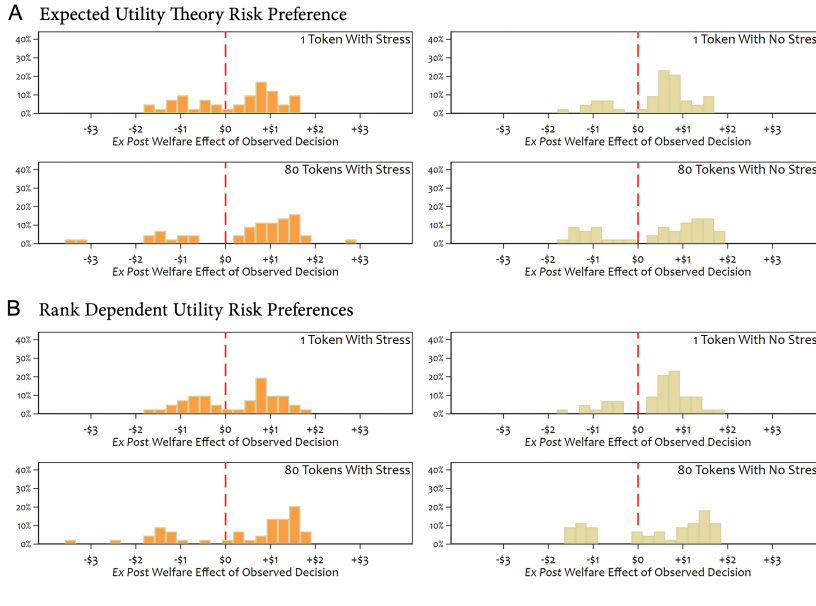


FIG. 15.—Ex post welfare effects of the decision to compete (or not).

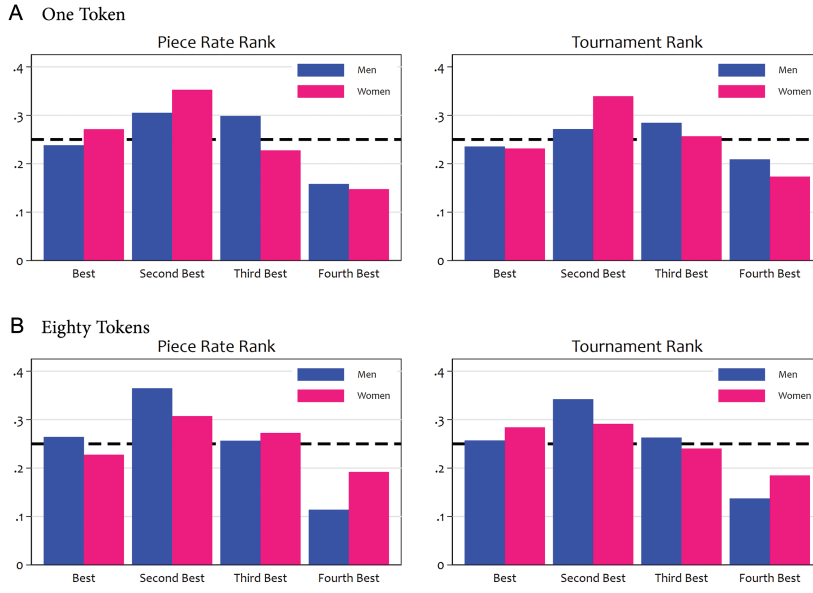


FIG. 16.—Recovered beliefs on the probability of personal performance ranks by men and women.

men hold beliefs that are right at the diffuse prior of one-quarter when there were 80 tokens, and women were modestly optimistic on average when there are 80 tokens.[48] From figure 12, we know that (different) women did much better in the scrambled condition with 80 tokens than in the comparable condition with one token, so this relative optimism is generally well-founded. Still, this optimism was taken into account in the simple economics of table 2, and the risk aversion of women was more than enough to offset these optimistic subjective beliefs.

Roughly 69% of all ex post welfare effects were improvements, with more welfare gains when the choice to compete involved no stress.[49] There was no (unconditional) difference between men and women in terms of the fraction of positive welfare effects. Nor was there any (conditional) difference between men and women when controlling for other demographics, personality traits, or binary indicators of the "belief optimism" or "performance excellence" of the individual. Belief optimism was an indicator

[48] This modest optimism by women was actually a marked optimism in terms of the observed belief reports. The beliefs in fig. 16 have been recovered from those observed reports, reflecting the RDU risk preferences of each individual. Figure B16 in app. B shows the results with observed belief reports.

[49] Detailed results are presented in app. C, sec. 3. All estimates refer to average marginal effects from probit regressions.

that the reported belief of winning the tournament exceeded one-quarter, and performance excellence was an indicator that the actual performance of the individual to be used in the tournament was in the top 25% over all subjects. Thus there appears to be no "gender problem" when it comes to these decisions to compete or not, when correctly evaluated in terms of welfare effects.

There is some evidence of systematic determinants of the welfare effects of the decision to compete in the no-stress setting of part 4. The first is that atheists tended to be 15 percentage points more likely to realize a welfare gain from their decisions (p-value = .09). The second is that belief optimists tended to be 18 percentage points less likely to realize a welfare gain (p-value = .07), and the third is that those who exhibited performance excellence tended to be 29 percentage points less likely to realize a welfare gain (p-value < .001). The last two results point to individuals who have some cause to expect to do well in the tournament, but possibly failed to factor in the extent of their risk aversion, even in the face of having "better odds than most."

One implication of the welfare evaluations in table 2 and figure 14 is that risk-neutral or expected-payoff-maximizing choices are not reliable measures of the welfare effects allowing for risk preferences. Moreover, these risk-neutral welfare calculations might also be relying on poorly calibrated subjective beliefs, which by themselves could generate welfare losses. Hence, one must have grave concern that various interventions to "get women to compete" could easily generate welfare losses for many individuals and even for the average individual. For example, Balafoutas and Sutter (2012) and Niederle, Segal, and Vesterlund (2013) consider a variety of "affirmative action" policies to reduce the gender gap with respect to choosing to compete but offer no evidence that it improves anyone's welfare. Alan and Ertac (2019) consider interventions to teach young children about the "value of grit" and determination, and that seems to reduce a gender gap in choosing to compete, but it is far from obvious that it improves welfare for risk-averse agents.[50] On the other hand, the complete design we applied, an extension of Niederle and Vesterlund (2007), coupled with some basic economics of risk, does point to the way to evaluate the welfare effects of these and other policies, to help establish credible priors that our interventions are, in expectation, doing no harm.[51]

[50] We explain the calculations of the "efficiency" and "expected cost" of these policies in app. C, sec. 4, and why they should not be viewed as evaluating expected welfare.

[51] One key difference in our design, which is not difficult to add, is the use of a battery of choices that allow one to characterize the risk preferences of individuals. Many of the studies in this subliterature have used minimal or questionable measures of risk aversion, primarily to have some covariate to add to regressions. Quite apart from any conceptual or empirical validity of these measures, they do not readily convert to risk-preference parameters that allow one to evaluate the CE of risky lottery choices that are central to the design and the normative evaluation of welfare.

B. Gender and Literacy

The literature on financial literacy stresses that women tend to be less literate than men. This conclusion is derived from a wide collection of hypothetical survey responses in which individuals are asked questions and given multiple-choice options to select from. In many cases the surveys allow "do not know" and "refuse to answer" as options. Reviewing a wide array of responses over many countries, Bucher-Koenen et al. (2017, 257) conclude that, "Not only are female respondents less likely to answer financial literacy questions correctly but they are also more likely to state that they do not know the answers to the questions."[52] The same theme is echoed by Lusardi and Mitchell (2008) and Bucher-Koenen et al. (2025). The latter, indeed, see the lack of confidence that women exhibit as a call for action, with a reference to "fearless women" in their title, explained in reference to

> *Fearless Girl*—a bronze statue of a girl that was placed in front of the Charging Bull on Wall Street in New York City on March 7, 2017 (one day before International Women's Day). The intent was to raise awareness and encourage women's leadership. Its symbolic placement sparked a debate about women's roles, particularly in financial professions, and pointed to the importance of confidence, especially in the fields of finance and investing. A fearless girl will become a fearless woman.

Stirring as this metaphor is, and important as the social debate is, we believe the metaphor to be dangerously misplaced. We encourage changing the conversation and measurement of confidence so that we think about "appropriate confidence" rather than simply more or less confidence. There are risks that one might want to be averse to, and even to fear, such as a charging bull. The issue is whether confidence is appropriate or not.

Maybe the greater use of the "do not know" response by women is simply a reflection of them not having a 100% belief that any one of the available answers is correct and cooperatively communicating that to the researcher. In other words, the response "do not know" could just be intended to reflect the belief that "I do not know with enough confidence to select it if you force me to select one option," and men and

[52] The conclusion includes several massive, nationally representative surveys. Focusing on the "inflation" question, we examine below, they show that the "do not know" option was used by 18.4% of the women and 9.8% of the men in the United States in 2009, by 16.9% of the women and 10.1% of the men in the Netherlands in 2010, and by 21.0% of the women and 12.4% of the men in Germany in 2009 (Bucher-Koenen et al. (2017, 260–62).

women might vary in their culturally conditioned willingness to be open in survey and experimental responses to this lack of complete confidence. This is a point stressed earlier for our Raven responses: we cannot identify whether the responses of women (and Blacks) reflect less confidence in their beliefs over the true answers or just a greater willingness to admit that. Exactly the same could be true of the literacy responses underlying the apparent gender disparity in financial literacy.

One approach is to construct a statistical model to view these "do not know" responses as a latent reflection of lack of confidence, allowing the model to reallocate probability mass to other options. This model requires comparable data in which the "do not know" option is not available to the respondent, and those data are available from the Netherlands and used in this manner by Bucher-Koenen et al. (2025). They also asked subjects who did not have this option to report, on a Likert scale, how confident they were in their response.

Inventive as this approach is, a more direct approach would be to simply change the mode of responses to such literacy questions to allow, and incentivize, respondents to report full belief distributions over possible responses, exactly as we have done for the mode of response to Raven questions. In fact, this idea was already proposed and implemented for literacy measurement by Di Girolamo et al. (2015) and Harrison, Morsink, and Schneider (2022).

We illustrate the effects of using this mode of response with a literacy question that is a direct adaptation of one of the "Big Three" from Lusardi and Mitchell (2011),[53] to test understanding of the effects of inflation:

> Imagine that you have \$100 in a savings account and the annual interest rate on your savings account was 1% per year, and annual inflation was 2% per year. After one year, how much purchasing power would you have on the initial \$100?

The correct answer to this question is $\$99.02 = \$100 \times 1.01 \div 1.02$. We asked this question of 776 GSU undergraduates, using the same incentivized QSR procedure as in our Raven experiments.[54] Subjects were

[53] The original question is, "Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, would you be able to buy more than, exactly the same as, or less than today with the money in this account?" Multiple-choice options provided were: "more than today," "exactly the same," "less than today," "do not know," and "refuse to answer." The correct answer is "less than today."

[54] In this case, these questions were part of a larger battery of questions on general financial literacy and inflation literacy, undertaken for the Federal Reserve Bank of Atlanta in 2013, 2014, and 2016. One belief question was selected at random for payment, and the QSR parameters selected so that the maximum reward for allocating all tokens to the correct response would earn \$50.

given response options in dollar amounts, shown on the horizontal axes of the displays in figure 17. One treatment was to shift the labels so that the correct answer was in a different belief report bin interval: hence, the horizontal axis displays on the left and right of figure 17 are, by design, slightly different even if both include the correct response.

The top panels of figure 17 (*A*, *C*) show the pooled beliefs of men in response to this question, and the bottom panels (*B*, *D*) show the pooled beliefs of women. In each panel, we report the average response μ and the standard deviation of responses σ. We have a direct measure of absolute bias as $|\mu - 99.02|$, and we observe that the bias is higher for women than it is for men, varying slightly by the range of options on the left or right panels. One might infer that women exhibit greater bias than men in this task and, hence, are less literate. But the displays of the distributions of beliefs, as well as the reported standard deviations, make it clear that these biases are tiny in relation to the level of confidence in these beliefs. Since it is a common error, we explicitly caution that the standard error of the statistic μ is not the same as the standard deviation σ of the underlying distribution. It could be that the estimated μ values are statistically significantly different from \$99.02, but that confuses the precision of an estimate of a statistic with the precision (i.e., inverse of variance) of the sample data distribution.[55] This is an error in statistics and not the same as the distinction between statistical and economic significance. Hence, we infer from figure 17 that both men and women exhibit biased beliefs about the correct response to this literacy question but report a lack of confidence in their belief that makes that bias statistically irrelevant.

More useful insights come from examining belief distributions of individuals, which our approach generates by construction. Figure 18 displays what we need to be paying attention to in terms of the likely welfare consequences of illiteracy. These are displays from actual subjects from our sample. Figure 18*A* shows the "poster child" of literacy: no bias and perfect confidence with all tokens allocated to the correct response. Figure 18*B* shows a common modal type of response for this question, a subject that has beliefs either side of the correct response but is unwilling to commit to allocating all tokens to her modal response. Figure 18*A* is important and interesting: the subject exhibits a large bias but is reporting such a diffuse belief distribution that we have to respect that the subject is telling us that they do not have a clue. Here is why this matters. If we went back to this subject, as a financial planner might and suggested that they look into this a bit more since it really matters for financial planning,

[55] This issue is highlighted by Harrison et al. (2022, 817) in the context of evaluations of the bias of subjective beliefs about the mortality effects of COVID-19.

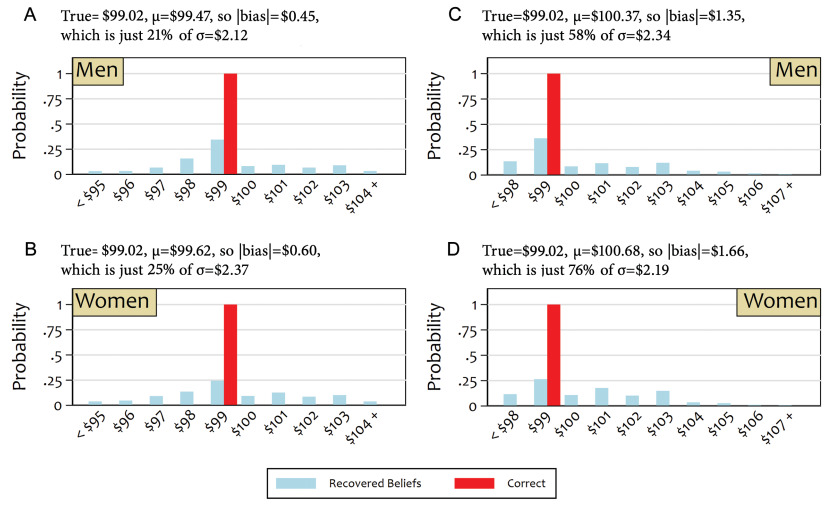


FIG. 17.—Bias and confidence in the literacy of men and women on the inflation question.

the diffuse belief suggests to us that this is a subject who is likely to start searching for a scaffold to make a better decision. Whether that scaffold is the advice of the financial planner, the internet, or a partner, the evidence of the diffuse belief should only be taken as evidence of "naked, Robinson Crusoe without bars on his phone" illiteracy.

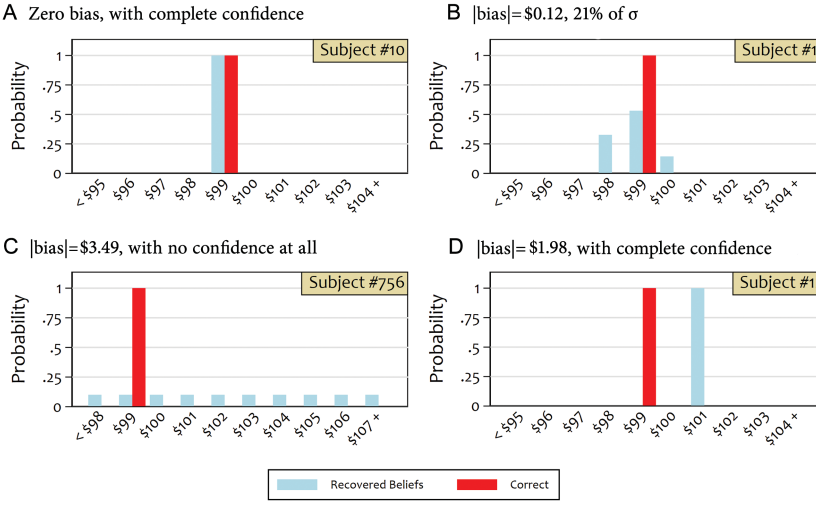


FIG. 18.—Literacy and confidence: the good, the bad, and the ugly.

The actual subject in figure 18*D* is the one we worry about. Here we see evidence of bias but complete confidence in the belief that generates that bias. This is the subject who we fear would "bet the house" on their belief and not see the need for seeking out scaffolds to aid their literacy. This is the subject type we particularly seek to identify, to evaluate the potential welfare gains of informational policy treatments or regulations.

IV. Extensions, Related Literature, and Open Issues

Winkler (1996, sec. 2) crisply distinguishes the role of scoring rules for ex ante assessment of subjective beliefs and the role of scoring rules for the ex post evaluation of those being assessed. Many issues arise when these two are confused rather than being seen as complementary.

A. Probabilistic Testing in Education

There is a long psychology literature considering variants on the standard scoring rules used in educational testing in order to elicit "partial knowledge": for example, Rippey (1968, 1970), Koehler (1971, 1974), Kansup and Hakistan (1975), Ben-Simon, Budescu, and Nevo (1997), Bereby-Meyer, Meyer, and Budescu (2003), and Bickel (2007, 2010). All recognize the value of measuring confidence and the potential role of proper scoring rules. However, from the ex ante perspective of assessing beliefs, most of the alternatives considered are openly ad hoc variants of existing scoring procedures in tests, and the metrics of evaluation are not well-motivated if the objective is reliable elicitation of confidence. All of the empirical studies rely on intrinsic rewards and in most cases provide a loose characterization to the subject of the operation and consequences of the scoring rule.

The literature on probabilistic testing in education has much more to say on the ex post perspective of evaluating reports and beliefs using scoring rules. And here the various reasons for conducting tests matter greatly. In the educational context, tests can be used to ordinally compare applicants, to prod to study, and as a measurement tool for some trait (Wainer, Bradlow, and Wang 2007, chap. 1). Each objective naturally leads to different criteria for tests to meet.

An important example is the issue of the scoring rule score depending on the complete distribution of reports and not just on the correct, true report. In testing applications, it is usually the case that one true alternative is defined. Shuford, Arthur, and Massengill (1966) proved that most of the proper scoring rules, such as the QSR, were symmetric in the manner in which they treated false response. If a subject reports a 70% token allocation to the true answer and allocates 20% to one of the false answers and 10% to another of the false answers, it does not matter for the QSR

score which false answers get the 20% or 10% allocations. But if another subject reports 70% for the true answer and 30% to just one of the false answers, the QSR score is lower than for the first subject, since the sum of the quadratic components of the QSR are different. This difference in score bothers some as unfair (e.g., Winkler 1996, 15 ff.; Bernardo and Smith 2000, 72; Bickel 2007, 2010). The implication is then to use a proper scoring rule that uses only responses to the true alternative, the logarithmic rule.[56]

There is general confusion in the literature on scoring rules for testing purposes on how to account for nonlinear utility of test respondents or more general forms of risk preference (e.g., Winkler 1996, 19ff., 52ff.; Bickel 2007, sec. 4). We consider this issue in sec. IV.C below.

B. Achievement Tests

Our overall objective with respect to the measure of intelligence has been to augment existing measurement methods in order to capture self-awareness of the degree of confidence a person has in a knowledge claim *and* a willingness to express that confidence. Heckman and Kautz (2012, 451) make a parallel point with respect to standardized achievement tests:

> Achievement tests miss, or more accurately, do not adequately capture, soft skills—personality traits, goals, motivations, and preferences that are valued in the labor market, in school, and in many other domains. The larger message . . . is that soft skills predict success in life, that they produce that success, and that programs that enhance soft skills have an important place in an effective portfolio of public policies.

The concept of a "soft skill" is evidently used in this passage to refer to entire broad areas of competence.[57] We argued and demonstrated constructively that one can rigorously elicit measures of confidence that augment traditional intelligence tests. Economic theory tells us how to do this: it is not a matter of eliciting some proxy for a trait, seeing if it improves statistical fit on some target behavior, and then declaring it a valid measure because of that improved statistical fit.

[56] Shuford, Arthur, and Massengill (1966, 137) present a "truncated" variant of the logarithmic rule that elegantly avoids the problem of unbounded payoffs for reports of 0.

[57] In cognitive science, the idea of a "soft skill" refers to a skill that lacks invariant criteria for successful application. Thus, interesting cases of soft skills must be narrower than the examples above, since every very broadly delineated general skill, including mathematical skill, meets the scientific definition of "soft." Confidence as understood here, however, is neither broadly delineated nor soft, but that is not the point.

We stress that our augmented intelligence test should not be viewed as a "magic bullet" correction for all of the limitations of traditional intelligence tests, particularly when interpreted and applied as general-purpose achievement tests. For example, the extended case studies in Heckman, Humphries, and Kautz (2014) of the General Educational Development (GED) testing program, and specifically the GED exam, show that GED recipients performed in the labor market more like high school dropouts than high school graduates. At one level, this might be dismissed as just claiming more for an achievement test than one should. But at another level, these achievement tests can take on an institutional, lobbying, and political life that risks selling false hope to vulnerable populations for decades, generating a derived demand for critical evaluation of the "performance of performance standards" erected around them (Heckman, Heinrich, and Smith 2002).

If someone is designing a test of (fluid) intelligence or an achievement test, they have a choice of having the respondent incentivized in a risk-neutral manner, incentivized by the respondent's own risk preferences, or even incentivized to induce some risk preferences on responses. In section IV.D, we explain how each of these alternatives have been implemented in the literature on the elicitation of beliefs and the trade-offs involved in adopting each approach. Here we ask why one might want to use one or other of these options.

Having subjects respond as if risk neutral has two advantages. One is that reports of beliefs are in fact the subjective beliefs of the subject, at least under the theory of the scoring rules employed. This means that reports are beliefs, and beliefs are data and not estimates. Hence, they can be legitimately used directly in subsequent statistical analyses as data. The other advantage is that it controls for a potential "nuisance parameter" when trying to compare measures across individuals. If one wants to make a claim that one person is more intelligent than another, in terms of a specific cognitive test such as the RAPM, then one does not need to attend to how those individuals managed the risk of their responses.

Having subjects respond with their own risk preferences is based on a rejection of risk preferences as just being a nuisance parameter. It recognizes that for some general "principal-agent" settings, the principal may have a keen interest in how well the agent manages the risk of choosing over alternatives that are difficult to evaluate cognitively. In other words, the ability to manage risk, when it matters to the agent and involves making cognitively challenging inferences, may be exactly what the principal wants to evaluate. It is certainly what we want to have our surgeons and ER physicians evaluated on, to take some striking examples; in fact, there is a myriad of executive tasks undertaken by agents, where good management of risk in the face of cognitive challenge is as important as the "raw processing ability" of meeting the challenge. An immediate application

of this approach with our data was the evaluation of the welfare cost to the respondent of forcing them to submit their modal belief as if it represented their entire belief.[58]

C. *Intrinsic Motivation*

As economists, in conducting experiments, we focus on measures of intelligence as demonstrated under controlled incentives. This is often contrasted in the literature with a construct referred to as "intrinsic motivation." The idea is that people's choices may be influenced by goals or values that they bring into the lab and that continue to motivate them after they learn about the incentives provided by the experimenter. For an example directly relevant here, subjects confronted with cognitive puzzles might prefer reporting correct answers to incorrect ones regardless of the experimental protocol that determines their earnings.

The everyday semantics of "intrinsic" might misleadingly suggest that such motivations are fundamental, in the sense of being more basic than the experimental incentives, and in that sense strictly independent of experimentally controlled incentives. The cognitive science literature, however, is largely unfriendly to this general model of human motivation (Chater 2018). There may be no fundamental motivations in an intensely socially malleable species such as humans. The relevant sense of intrinsic motivation for our purposes is simply an incentive we do not observe, and do not control, that might interact with those we do manipulate through the QSR.

The possible influence of such motivations might explain our main finding with respect to the comparative efficiency of women's and men's behavior in our experiment. Suppose that many subjects both want to maximize their earnings and report uniquely "correct" answers. In that case, it is an open possibility that, perhaps due to gendered socialization histories, men are more strongly influenced by the second motivation, relative to the first motivation, than women are. This hypothesis cannot be evaluated based on our experiment. However, the hypothesized mechanism is observable in principle and could be examined in a follow-up experiment with treatments to separate the financial motivation from the speculative "intrinsic" motivation.

D. *The Elicitation of Beliefs*

The ex ante elicitation of subjective beliefs is at the core of the approach to measurement of intelligence, competitiveness, and literacy that we

[58] The need for careful treatment of nonlinear utility in the use of scoring rules is also stressed by eminent commentators Kadane, Lindley, and Murphy of Winkler (1996, 37, 38 ff., 42).

propose. As it happened, the pattern of RAPM responses we observed could generally be safely evaluated without attending to risk preferences. But in general this is not the case, so we must recognize how to deal with risk preferences when recovering beliefs.[59] There are three ways to elicit beliefs that address, or explicitly ignore, risk preferences, each with strengths and weaknesses for different applications.

Following Manski (2004), belief distributions might be elicited using hypothetical surveys. The key feature here is the absence of any incentives for responses. The advantage of this approach is that it is cheap, easier to explain to subjects, easier to implement in software, allows questions about events that cannot be verified, and does not need to attend to risk preferences since there are no (risky) consequences. Considerable attention has been paid to the manner in which belief questions are presented, particularly in field settings; see Delavande, Giné, and McKenzie (2011). The disadvantage of this approach is that the results might exhibit hypothetical bias, and it is easy to document that this bias can be a significant one (Harrison 2025). What "hypothetical bias" means here is that one gets different results, usually on a between-subjects basis, when asking the same question with no incentives compared to asking with incentives. In the nature of subjective beliefs, there is no true answer that one can look to: fidelity with some objective outcome is no metric of value here, unless one wants to impose some "rational expectations" constraint on beliefs, which we do not. So the expression "hypothetical bias" is just a shorthand for the results being different with and without incentives and the prior that incentivized responses are likely to reflect more effort and willingness to respond truthfully than nonincentivized responses. Although this is our prior, we appreciate that others might not share it.

For us, the primary remaining advantage of hypothetical elicitation is the ability to ask questions about nonverifiable events. For example, if someone is interested in longevity risk, one could ask a series of incentivized questions about how long the subject thinks people in their country will live, or someone with their gender, or someone with their gender and race, or someone with their gender, race and income level, and so on. Each of these could be incentivized with respect to official mortality tables. But then a question about how long the specific subject will live cannot be so incentivized and is what we might be interested in. To take this example, we would ask all of the incentivized questions and then one nonincentivized question and evaluate the bias and confidence of all but the last in relation to verifiable data. This would then give us a prior on

[59] Striking examples of the potential importance of the effect of risk preferences on recovered beliefs are easy to find for the beliefs about ranks in the two "competitiveness" tasks in which the subject's earnings depend on having the best piece-rate score or tournament score. Figures C1–C6 in app. C document this point for two subjects.

the bias and confidence for the final hypothetical question. In this manner, we see hypothetical and incentivized belief questions as complementary.

A second approach to belief elicitation starts by recognizing that risk preferences affect the rational responses of subjects to the popular scoring rules but that risk-neutral subjects should rationally report their beliefs. Hence, one can append an experimental payment procedure, called the binary lottery procedure (BLP), to "risk-neutralize" the responses of the subject and view reports directly as beliefs. The BLP was developed by Smith (1961) and has been widely used in various settings in which risk preferences are a confound, such as the bargaining experiments of Roth and Malouf (1979). The first statements of this mechanism, joining the QSR and the BLP, appear to be Allen (1987) and McKelvey and Page (1990). Hossain and Okui (2013) and Schlag and ven der Weel (2013) examine the same extension of the QSR, along with certain generalizations, renaming it a "binarized scoring rule" and "randomized QSR," respectively. Harrison et al. (2015) evaluated the QSR with the BLP in terms of the elicitation of subjective belief distributions, not just some summary statistic of that distribution. Harrison and Phillips (2014) applied this method to elicit the beliefs about financial risk by chief risk officers of major companies, all of whom had advanced training in statistical and actuarial methods.

The clear strength of this approach is that it allows the reports of the subject to be evaluated directly as beliefs, under the maintained assumption that the BLP works as advertised by theory. The disadvantage is that it adds an additional layer of understanding for subjects when it comes to translating earnings in the QSR into cash.[60] In our experience, documented in Harrison, Martínez-Correa, and Swarthout (2014), Harrison and Phillips (2014), and Harrison et al. (2015), this is relatively easy to overcome: one simply defines all earnings from the QSR in terms of points rather than a natural currency and then adds a paragraph at the end

[60] Although there are some studies showing the apparent failure of the BLP in other settings, that evidence is not as compelling as many claim; see Harrison et al. (2013) for a critical literature review. The claim by Danz, Vesterlund, and Wilson (2022) that the QSR using the BLP fails metrics of "behavioral incentive compatibility" is premature. They explicitly assume that subjects correctly understand the objective likelihoods the experimenter induces. They correctly note that the real "challenge for examining whether information on the mechanism's quantitative incentives encourages truth telling is that we do not know participants' true beliefs." But they then take the extreme metric of "reports of the objective prior as truthful (subjective beliefs). This true/false terminology is chosen for clarity and does not imply that all participants are assumed to understand that the objective prior is the true likelihood" (Danz, Vesterlund, and Wilson 2022, 2853). So they are testing some variant of a rational expectations model of subjective beliefs jointly with the incentive compatibility of the QSR using the BLP applied to the prior subjective beliefs subjects might hold, and they are not just testing the latter hypothesis.

explaining how points get turned into cash using the BLP.[61] Nonetheless, there are many settings in which one does not want to assume understanding of the BLP mechanism, particularly for less formally literate field populations who, in our experience, can be wary of "sophisticated," indirect payment protocols.

The third approach is to just implement some scoring rule, such as the QSR, pay the subjects the stated rewards in a natural currency, and use independent measures of risk preferences to rigorously recover the beliefs that led to those reports (Andersen et al. 2014a) and Harrison, Monroe, and Ulm 2022). The strength of this approach is that it allows the subject to see the monetary bets that her reports are generating, just as if she was placing bets with a series of bookies at some sporting event. This framing of the reports as bets is made more transparent with the use of real-time interfaces of the kind we use, developed by Harrison et al. (2017). Early implementations of the QSR relied on subject's understanding algorithms or squinting at long numerical tabulations of potential payoffs, but those have not been used for decades.

The weakness of this approach is that one must account for the effect of risk preferences on stated reports. This requires one additional task be conducted to elicit risk preferences, some econometrics to infer risk preferences at the individual level, and the recovery of beliefs once one has those estimated risk preferences in hand, as illustrated by Harrison, Monroe, and Ulm (2022).

In summary, we see the second and third approaches as the most attractive, and complementary, and have used both in different settings. They trade off the "extra work" needed to rigorously identify subjective beliefs. The second approach effectively puts that burden on the subject, and the third approach effectively puts that burden on the experimenter.

E. Calibration and the Evaluation of Beliefs

When evaluating the beliefs of a decision-maker ex post, it is common to propose metrics based on some notion of the "long run" consistency of subjective beliefs with objective realizations of events.[62] Referred to as

[61] We avoid the use of "experimental currencies" and a stated exchange rate between those currencies and some natural currency. This device is often used to just scale up rewards in some framed experimental currency, with the hope that this motivates subjects better. This procedure risks the confound of money illusion: if the subjects do not suffer from money illusion, the procedure adds nothing to incentives in terms of the natural currency, but if they do suffer from money illusion, the experimenter has lost control of incentives.

[62] Consistency is usually defined for some binary event E, e.g., the prediction of rain tomorrow, as in Budescu and Johnson (2011). Then conditional consistency is evaluated by comparing the forecast probability by individual i of the event j occurring, $\mathrm{P}(\mathrm{E}_{ij}|C_{ij} = c)$, where C_{ij} is the forecast probability, E_{ij} is the event j forecast by individual i, and c is some

"calibration," the idea is that individuals who are less calibrated have, in some sense, inferior forecasting abilities. In relation to our approach, calibration can be viewed as promoting the idea of accuracy as an appropriate metric for evaluation rather than, as we propose, efficiency derived from expected earnings of reports, or ideally welfare.

Calibration is usually introduced in the context of a binary event, such as the trusty weather forecaster being asked whether it will rain in a given city tomorrow. Presumably "rain" is when there is some minimal, agreed precipitation level over some agreed, specific locations. One immediate problem with binary events is that it is not possible to tease apart bias from an inappropriate variance, where "appropriate" for us means a comparison to the Bayesian posterior variance.[63] There are many other, well-known conceptual reasons for wanting to avoid notions of calibration, even in the binary case, and particularly to avoid the normative notion of using information on miscalibration to correct beliefs (e.g., Seidenfeld 1985) and Kadane and Fischhoff 2013).

For subjective beliefs elicited over continuous events, it is possible to tease apart bias from an inappropriate variance. For this reason, such settings are ideal for tests of the behavioral evidence for Bayesian updating, allowing "confidence" to be defined in terms of the variance of beliefs rather than the presence of a bias (e.g., Harrison and Swarthout 2025). Alpert and Raiffa (1982) extended the notion of calibration to this setting, by examining whether or not the interquartile range of beliefs occurred 75% of the time and whether there were any realizations outside the range of beliefs with positive subjective probability.

In our case, subjective beliefs are probability mass functions defined over three or more alternatives that are not even ordered. One could still construct some calibration metric, such as evaluating whether modal beliefs occurred with the subjective probability assigned to the mode(s) or whether any alternative occurred that had been assigned zero subjective probability. But at this point, the idea that some simple "calibration metric" can be devised becomes problematic. And the general conceptual reasons for wanting to avoid notions of calibration remain.

realized value for the forecast. By collecting a number of forecasts for a given individual i for the value c, or nearby values, miscalibration is revealed when $\mathrm{P}(\mathrm{E}_{ij}|C_{ij} = c) \neq c$. Unconditional miscalibration can also be evaluated and revealed when $\mathrm{P}(\mathrm{E}_{ij}) \neq C_{ij}$.

[63] A Bernoulli distribution is defined by the parameter p for the probability of success, with mean p and variance $p(1 - p)$. With Bayesian methods, the estimate for p is a random variable with a mean and variance, typically characterized by a beta distribution. But this does not lead to the variance of the Bernoulli realizations being independent of the mean of the realizations. The same is true for the Binomial distribution defined over multiple trials.

F. Normative Implications

Although motivated responses are the only ones we care about, the possible effect of financial motivations has important implications for the way in which intelligence scores are interpreted and used. Our perspective is that the responses we see in intelligence tests come from a cognitive production function that takes as input effort, prior knowledge of heuristics, experience with logic, subjective valuation of time, and risk-management skills.[64] The fact that there is always some opportunity cost to applying effort or time is enough for us to see that financial incentives might matter. Any effect of financial incentives is viewed as confounding by some, because they would like to measure intelligence in a context-free manner, in the hope that the measure in question might apply invariantly across contexts.[65] This ambition would be feasible if intelligence were a simple function of something like "raw cortical processing power," with effects that are not conditional on varying environmental affordances. But cognitive science rejects that view (Flynn 2007; Sloman and Fernbach 2017). We submitted it to test.

The deeper normative issue, for us, has to do with what one provides incentives for, rather than whether one should provide incentives at all. Our primary contribution is to expand the notion of incentives, as applied in the measurement of (fluid) intelligence, and then to the measurement of literacy and the measurement of the willingness to compete. We expand the notion of incentives to explicitly include incentives to report the appropriate level of confidence in some proposition that reflects subjective beliefs of the individual. And we expand the notion of incentives to explicitly account for the role of risk preferences in making decisions that involve subjective risk.

We do not assume that the individual is unbiased in responses about beliefs or even has the appropriate level of confidence suggested, say, by the correct application of Bayes' rule. On the contrary, we want to study any such biases or deviations from Bayesian posterior distributions in terms of variance (or higher-order moments) in order to normatively evaluate observed behavior toward risks.

The broader implications for the evaluation of policy programs of rigorously extending traditional tests of intelligence (or literacy or competitiveness) to allow for confidence remain to be documented. Evidence

[64] Camerer and Hogarth (1999) provide an explicit statement of this perspective on "produced" cognition. Major elaborations to estimate the multistage production function for cognitive and noncognitive skills have been provided by Cunha and Heckman (2007, 2008) and Cunha, Heckman, and Schennach (2010).

[65] Cawley et al. (1997) directly challenge claims about the immutability of measures of cognitive ability, stemming from the debates over *The Bell Curve* of Herrnstein and Murray (1994). And Cawley, Heckman, and Vytlacil (2001) stress how problematic it is to rigorously tease apart the tight correlation of schooling and measures of "innate" cognitive ability.

has already been accumulating about the importance of augmenting traditional measures of intelligence with measures of skills that

> are often defined and measured in terms of work habits, such as effort, discipline, and determination, or in terms of behavioral traits, such as self-confidence, sociability, and emotional stability. No single factor has emerged in the psychological literature and it is unlikely that one will be found, given the diversity of traits [involved]. (Ter Weel 2008, 729)

The evidence that these additional skills, even with traditional measures of intelligence, help to understand labor market outcomes and behavior over the life cycle is now established; see, for example, Borghans, Ter Weel, and Weinberg (2008), Ter Weel (2008), Heckman, Pinto, and Savelyev (2013), and Heckman and Kautz (2014). Again, we do not see our proposal to rigorously augment measures of intelligence (or literacy or competitiveness) as a substitute for the measurement of these additional skills but as a complement.

Extending this idea, it is formally possible to induce any level of risk preferences that an agent might face.[66] In this manner, the principal can evaluate how the agent manages risk preferences that the principal cares about, which need not be risk neutrality or the homegrown risk preferences of the respondent. Again, the reason for this choice is that the principal cares about how the agent manages risk, since that is a part of communicating beliefs about the best choices when facing cognitive risk. The same logic extends to considering all of the economic consequences to the principal when designing scoring rules for managerial decision-making (Murphy 1966; Pearl 1978).

G. Origins of the Concept of Scaffolding

Scaffolding has been defined generally by Clark (1997, 45) as "exploitation of external structure" in information processing by an agent. Clark (2003, 2011) provided the most influential consolidation, which has been taken up by most cognitive scientists, of the idea that animals with relatively complex central nervous systems carry out some, or indeed much, of their thinking by manipulating aspects of their physical environments. This is partly a means of experimentation to refine conjectures about relationships, but more fundamentally it directly amplifies

[66] The idea is simple but difficult to implement. Berg et al. (1986) extended the BLP, which risk-neutralizes responses, to allow for a nonlinear exchange rate between the "points" that define the binary probability lottery chances of a bigger prize. This extension allows the principal or experimenter to induce risk-averse or risk-loving preferences with concave or convex exchange rate functions, albeit with added complexity for subjects.

an agent's intelligence by formatting presentations of information in ways that make it easier to cognitively model (Dennett 1996). In a classic example due to Kirsh and Maglio (1994), people cannot solve problems in the video game Tetris without actively engaging with the game. Thus if Tetris play is used, as it occasionally has been, to proxy intelligence as pattern recognition, a test subject equipped with a working interface would be assigned a much higher score than a test subject who lacked access to this scaffolding. External structure for scaffolding is not limited to physical tools and includes language and socially evolved concepts, as stressed by Vygotskij (1962) and Bruner (1968). Dennett (1991, 1996) argued that language is the most ubiquitous and fundamental scaffolding technology for humans: each person encounters it as an already evolved technology that she must learn to use and without which manifestation of basic social intelligence is impossible.

The first clear statement of what is now thought of as scaffolding comes from the critiques by Dreyfus (1965, 1972) of the so-called "classical" approach to artificial intelligence (AI). The attack on AI was aimed at the idea that we could design AI systems that contained all of the information they needed and just had to "look up" this information when making a decision. Intelligent systems, Dreyfus argued, must be embedded in, and learn on the basis of, dynamic interaction with external environments.

Bruner (1968) picked up the ideas of Dreyfus (1965) and applied them to developmental learning by children. This extension was an important widening of the notion of scaffolding to include social interactions, although not requiring that the external context be social. This extension is important for another reason: it allows us to see how scaffolds need not always be Pareto improvements, even if they are costless to implement. For example, consider the use of the word "boat." Suppose the socially accepted definition of a boat required that it be made of wood, drawing on historical precedent from when all boats were wooden. Then when someone comes along to suggest making a steel boat, they might encounter a chorus of complaints that "that is not a boat!" This restricted concept of a boat, as a scaffold, might, at least for a time, inhibit anyone from drawing inferences from the history of wooden boats to help inform the design of steel boats.

The idea of "embodied cognition" from Dreyfus, now central to cognitive science generally (Shapiro 2014), refers to two different aspects, only one of which is related to scaffolding. The unrelated item is a generalization of the idea of "muscle memory." The related item is knowledge that we build into artifacts, so that we can then "forget" it. For example, users once knew how to program their word processors, but most no longer do because the knowledge is "embodied" in the interface. So this sense of embodied cognition is a form of scaffolding.

Almost all controlled experiments in economics entail the use of some scaffolds, because of the use of language to convey questions and possible responses. A particularly striking example in time discounting experiments arises when one tells a subject the interest rates implied by their choices between "smaller, sooner" amounts of money and "larger, later" amounts of money (e.g., Andersen et al. 2014b), who study the effect of removing this scaffold on elicited discount rates). Individuals have varying field experience with what interest rates are and, hence, treat that information as a scaffold to varying degrees. Indeed, the history of behavioral economics, particularly when focused on framing anomalies, is largely about manipulating, in a controlled manner, the nature of the scaffolding provided to subjects. An open issue, which is a major theme of the comparison of lab and field experiments, is whether behavior in the context of artifactually provided scaffolds is the same as behavior in the context of natural scaffolds, including scaffolds that are endogenously sought out by agents (e.g., checking Google or Wikipedia). In a related vein, there is some evidence that humans process probabilities better when they are presented as "natural frequencies" rather than as probability statements, reflecting the intuition that humans have used frequencies more generally over time as a scaffold than they have used probabilities (e.g., Gigerenzer and Hoffrage 1995).

Scaffolding can both promote and detract from efficiency, depending on the context. Hutchins (1995) provides the most richly detailed and extended case study, of a trained crew operating an aircraft carrier. But socially intelligible thought—and therefore most of what anyone has meant by "human intelligence" on any rigorous conceptualization—is impossible without it.

V. Conclusions

Intelligence tests, and measures of cognition more generally, are typically applied under the maintained assumption that salient incentives do not matter or are somehow always sufficient in the form of uncontrolled incentives to encourage responses that reliably reflect individual attributes. Perhaps it is implicitly assumed that people are naturally and typically motivated to demonstrate their full cognitive potential whenever they are told it is being probed; we return to this assumption below. Remarkably, there is very little evidence on these issues for complete measures of fluid intelligence such as the Raven Advanced Progressive Matrices test, despite the widespread use by economists of cognitive attributes "on the right-hand side" of regressions.[67] We provide controlled evidence that the level

[67] As noted earlier, this general point has been recognized in literature reviewed by Borghans et al. (2008), Borghans, Meijers, and Ter Weel (2008), Segal (2012), and Chen

of salient incentives matter for those measures. Individuals exhibit more fluid intelligence in the test we use when financially motivated. We have no prior that the effect of incentives on performance is likely to be identical across individuals.

Critically, intelligence tests and measures of cognition do not account for the confidence with which individuals report responses. At best, they reflect modal beliefs about the possible alternative responses. We have no prior that the mode is likely to reflect the belief distribution or even the central tendency of the distribution.

There has long been evidence that individuals' general confidence in their attributional judgments varies across cultures with the extent to which culture-specific learning encourages people to attend to context (Gudykunst and Nishida 1986). Context invariably includes social scaffolding and typically some asocial scaffolding as well. Variance in such cultural learning is widely regarded as among the sources of cognitive inequality within cultures that motivates increased social investment in early childhood education for children from disadvantaged households.[68]

In fact, our elicitation interface itself is our general scaffolding treatment, as noted earlier. Dennett (1991, 218–24) argues that the most important and basic scaffolding for the kind of intelligence that is distinctive to humans is the regular requirement we face to format our thoughts for presentation to, and for comprehension by, other people. When we allow more than one token to be assigned, our interface requires subjects to format their thoughts in terms of confidence reports. This makes them smarter on average—the interface is not merely making "preexisting" inner intelligence observable to us.

The effects of our interface, particularly for women and Blacks, is sharp evidence of the manner in which different scaffolds and context contribute to the cognitive inequality in cultures. Our striking result about the superiority of women with the correct measurement of fluid intelligence generalizes to measures of the "competitiveness" of women and measures of their financial literacy. In each case, there are claims that women exhibit insufficient confidence in their behavior, and we show that these conclusions are due to measurement being based on observable performance metrics that have nothing to do with earnings efficiency and welfare.

Fundamentally, we encourage recognition that the production of intelligence and cognition in general requires attention to the confidence of beliefs, as stressed by Savage (1971, 800). If someone is to have a derived demand for costly cognitive scaffolds with which to produce intelligence

et al. (2020). We do not know of any evidence for complete batteries in the domain of fluid intelligence.

[68] For example, see Heckman (2013) and Levitt et al. (2016).

and cognition, they need to be aware that scaffolds could be valuable to them.

Data Availability

Code replicating the tables and figures in this article, as well as documentation for the software used to conduct the experiments, can be found in Harrison, Ross, and Swarthout (2025) in the Harvard Dataverse, https://doi.org/10.7910/DVN/IZQOI6.

References

Alan, Sule, and Seda Ertac. 2019. "Mitigating the Gender Gap in the Willingness to Compete: Evidence from a Randomized Field Experiment." *J. European Econ. Assoc.* 17 (4): 1147–85.

Allen, Franklin. 1987. "Discovering Personal Probabilities When Utility Functions Are Unknown." *Management Sci.* 33 (4): 452–54.

Alpert, Marc, and Howard Raiffa. 1982. "A Progress Report on the Training of Probability Assessors." In *Judgment under Uncertainty: Heuristics and Biases*, edited by D. Kahneman, P. Slovic, and A. Tversky. New York: Cambridge Univ. Press.

Andersen, Steffen, John Fountain, Glenn W. Harrison, and E. Elisabet Rutström. 2014a. "Estimating Subjective Probabilities." *J. Risk and Uncertainty* 48:207–29.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2014b. "Discounting Behavior: A Reconsideration." *European Econ. Rev.* 71: 15–33.

Arthur, Winfred, Jr., and David V. Day. 1994. "Development of a Short Form for the Raven Advanced Progressive Matrices Test." *Educ. and Psychological Measurement* 54 (2): 394–403.

Balafoutas, Loukas, and Matthias Sutter. 2012. "Affirmative Action Policies Promote Women and Do Not Harm Efficiency in the Laboratory." *Science* 335 (6068): 579–82.

Ben-Simon, Anat, David V. Budescu, and Baruch Nevo. 1977. "A Comparative Study of Measures of Partial Knowledge in Multiple-Choice Tests." *Appl. Psychological Measurement* 21 (1): 65–88.

Bereby-Meyer, Yoella, Joachim Meyer, and David V. Budescu. 2003. "Decision Making Under Internal Uncertainty: The Case of Multiple-Choice Tests with Different Scoring Rules." *Acta Psychologica* 112:207–20.

Berg, Joyce E., Lane A. Daley, John W. Dickhaut, and John R. O'Brien. 1986. "Controlling Preferences for Lotteries on Units of Experimental Exchange." *Q.J.E.* 101 (May): 281–306.

Bernardo, José M., and Adrian F. M. Smith. 2000. *Bayesian Theory*. Chichester, UK: Wiley.

Bickel, J. Eric. 2007. "Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules." *Decision Analysis* 4 (2): 49–65.

———. 2010. "Scoring Rules and Decision Analysis Education." *Decision Analysis* 7 (4): 346–57.

Bickerton, Derek. 2014. *More than Nature Needs: Language, Mind and Evolution*. Cambridge, MA: Harvard Univ. Press.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *J. Human Resources* 46 (4): 972–1059.

Borghans, Lex, Bart H. H. Golsteyn, James J. Heckman, and Huub Meijers. 2009. "Gender Differences in Risk Aversion and Ambiguity Aversion." *J. European Econ. Assoc.* 7 (2–3): 649–58.

Borghans, Lex, Huub Meijers, and Bas ter Weel. 2008. "The Role of Noncognitive Skills in Explaining Cognitive Test Scores." *Econ. Inquiry* 46 (1): 2–12.

Borghans, Lex, Bas ter Weel, and Bruce A. Weinberg. 2008. "Interpersonal Styles and Labor Market Outcomes." *J. Human Resources* 43 (4): 815–58.

Bruner, Jerome. 1968. "Processes of Cognitive Growth: Infancy." Clark Univ. Press Heinz Werner Lectures, https://commons.clarku.edu/heinz-werner-lectures/20.

Bucher-Koenen, Tabea, Rob J. Alessie, Annamaria Lusardi, and Maarten van Rooij. 2025. "Fearless Woman: Financial Literacy, Confidence and Stock Market Participation." *Management Sci.* 71 (9): 7223–8095.

Bucher-Koenen, Tabea, Annamaria Lusardi, Rob J. Alessie, and Maarten van Rooij. 2017. "How Financially Literate Are Women? An Overview and New Insights." *J. Consumer Affairs* 51 (2): 255–83.

Budescu, David V., and Timothy R. Johnson. 2011. "A Model-Based Approach for the Analysis of the Calibration of Probability Judgments." *Judgment and Decision Making* 6 (8): 857–69.

Camerer, Colin F., and Robin M. Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework." *J. Risk and Uncertainty* 19:7–42.

Carpenter, Patricia A., Marcel Just, and Peter Shell. 1990. "What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test." *Psychological Rev.* 97 (3): 404–31.

Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies.* New York: Cambridge Univ. Press.

———. 1997. "The Three-Stratum Theory of Cognitive Abilities." In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, edited by D. P. Flanagan and L. Harrison. New York: Guilford.

Cawley, John, Karen Conneely, James Heckman, and Edward Vytlacil. 1997. "Cognitive Ability, Wages, and Meritocracy." In *Intelligence, Genes, and Success*, edited by B. Devlin, S. E. Fienberg, D. P. Resnick, and K. Roeder. New York: Springer.

Cawley, John, James Heckman, and Edward Vytlacil. 2001. "Three Observations on Wages and Measured Cognitive Ability." *Labour Econ.* 8 (4): 419–42.

Chater, Nick. 2018. *The Mind Is Flat.* New York: Penguin.

Chen, Yuanyuan, Shuaizhang Feng, James J. Heckman, and Tim Kautz. 2020. "Sensitivity of Self-Reported Noncognitive Skills to Survey Administration Conditions." *Proc. Nat. Acad. Sci. USA* 117 (2): 931–35.

Chen, Yvonne Jie, Deniz Dutz, Li Li, Sarah Moon, Edward Vytlaci, and Songfa Zhong. 2024. "Eliciting Willingness-to-Pay to Decompose Beliefs and Preferences that Determine Selection into Competition in Lab Experiments." *J. Econometrics* 243 (1–2): 105652.

Civelli, Andrea, and Cary Deck. 2018. "A Flexible and Customizable Method for Assessing Cognitive Abilities." *Rev. Behavioral Econ.* 5:123–47.

Clark, Andy. 1997. *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA: MIT Press.

———. 2003. *Natural-Born Cyborgs.* New York: Oxford Univ. Press.

———. 2011. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford Univ. Press.
Court, J. H. 1983. "Sex Differences in Performance on Raven's Progressive Matrices: A Review." *Alberta J. Educ. Res.* 29:54–74.
Cunha, Flavio, and James J. Heckman. 2007. "The Technology of Skill Formation." *A.E.R. Papers and Proc.* 97 (2): 31–47.
———. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *J. Human Resources* 43 (4): 738–82.
Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.
Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." *A.E.R.* 112 (9): 2851–883.
Daston, Lorraine. 1992. "The Naturalized Female Intellect." *Sci. Context* 5 (2): 209–35.
de Finetti, Bruno. 1965. "Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item." *British J. Math. and Statis. Psychology* 18 (May): 87–123.
Delavande, Adeline, Xavier Giné, and David McKenzie. 2011. "Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence." *J. Development Econ.* 94:151–63.
Dennett, Daniel. 1991. *Consciousness Explained*. New York: Little, Brown.
————. 1996. *Kinds of Minds*. New York: Basic.
————. 2017. *From Bacteria to Bach and Back*. New York: Norton.
Di Girolamo, Amalia, Glenn W. Harrison, Morten I. Lau, and J. Todd Swarthout. 2015. "Subjective Belief Distributions and the Characterization of Economic Literacy." *J. Behavioral and Experimental Econ.* 59:1–12.
Dreyfus, Hubert L. 1965. "Alchemy and Artificial Intelligence." https://www.rand.org/pubs/papers/P3244.html.
———. 1972. *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: Harper and Row.
Evans, Dylan. 2012. *Risk Intelligence: How to Live with Uncertainty*. New York: Free Press.
Flynn, James R. 2007. *What Is Intelligence? Beyond the Flynn Effect*. New York: Cambridge Univ. Press.
Gao, Xiaoxue Sherry, Glenn W. Harrison, and Rusty Tchernis. 2023. "Behavioral Welfare Economics and Risk Preferences: A Bayesian Approach." *Experimental Econ.* 26:273–303.
Gigerenzer, Gerd, and Ulrich Hoffrage. 1995. "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats." *Psychological Rev.* 102 (4): 684–704.
Gignac, Gilles E. 2018. "A Moderate Financial Incentive Can Increase Effort, but Not Intelligence Test Performance in Adult Volunteers." *British J. Psychology* 109 (3): 500–516.
Gneezy Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences." *Q.J.E.* 118:1049–74.
Gudykunst, William B., and Tsukasa Nishida. 1986. "Attributional Confidence in Low- and High-Context Cultures." *Human Communication Res.* 12:525–49.
Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *J. Econ. Literature* 46 (3): 607–68.
Hardcastle, Joseph, Cari F. Hermann-Abell, and George E. DeBoer. 2017. "Comparing Student Performance on Paper-and-Pencil and Computer-Based

Tests." Paper presented at the AERA Annual Meeting, https://files.eric.ed.gov/fulltext/ED574099.pdf.
Harrison, Glenn W. 2025. "Causal Inference Over Unobservables." Working Paper no. 2025-02, CEAR, Georgia State Univ.
Harrison, Glenn W., Andre Hofmeyr, Harold Kincaid, Brian Monroe, Don Ross, Mark Schneider, and J. Todd Swarthout. 2022. "Subjective Beliefs and Economic Preferences During the COVID-19 Pandemic." *Experimental Econ.* 25:795–823.
Harrison, Glenn W., Jimmy Martínez-Correa, and J. Todd Swarthout. 2013. "Inducing Risk Neutral Preferences with Binary Lotteries: A Reconsideration." *J. Econ. Behavior and Org.* 94:145–59.
———. 2014. "Eliciting Subjective Probabilities with Binary Lotteries." *J. Econ. Behavior and Org.* 101 (May): 128–40.
Harrison, Glenn W., Jimmy Martínez-Correa, J. Todd Swarthout, and Eric R. Ulm. 2015. "Eliciting Subjective Probability Distributions with Binary Lotteries." *Econ. Letters* 127:68–71.
———. 2017. "Scoring Rules for Subjective Probability Distributions." *J. Econ. Behavior and Org.* 134:430–48.
Harrison, Glenn W., Brian Monroe, and Eric Ulm. 2022. "Recovering Subjective Probability Distributions: A Bayesian Approach." Working Paper no. 2022–03, CEAR, Robinson Coll. Bus., Georgia State Univ.
Harrison, Glenn W., Karlijn Morsink, and Mark Schneider. 2022. "Literacy and the Quality of Index Insurance Decisions." *Geneva Risk and Insurance Rev.* 47 (1): 66–97.
Harrison, Glenn W., and Richard D. Phillips. 2014. "Subjective Beliefs and the Statistical Forecasts of Financial Risks: the Chief Risk Officer Project." In *Contemporary Challenges in Risk Management*, edited by T. J. Andersen. New York: Palgrave Macmillan.
Harrison, Glenn W., and Don Ross. 2023. "Behavioral Welfare Economics and the Quantitative Intentional Stance." In *Models of Risk Preferences: Descriptive and Normative Challenges*, edited by G. W. Harrison and D. Ross, eds. Bingley, UK: Emerald.
———. 2025. *The Gambling Animal: Humanity's Evolutionary Winning Streak—and How We Risk It All.* London: Profile.
Harrison, Glenn W., Don Ross, and J. Todd Swarthout. 2025. "Replication Data for: 'Gender, Confidence, and the Mismeasure of Intelligence, Competitiveness and Literacy.'" Harvard Dataverse, https://doi.org/10.7910/DVN/IZQOI6.
Harrison, Glenn W., and J. Todd Swarthout. 2023. "Cumulative Prospect Theory in the Laboratory: A Reconsideration." In *Models of Risk Preferences: Descriptive and Normative Challenges*, edited by G. W. Harrison and D. Ross. Bingley, UK: Emerald.
———. 2025. "Belief Distributions, Bayes Rule and Apparent Confidence." Working Paper no. 2020–11, CEAR, Robinson Coll. Bus., Georgia State Univ.
Heckman, James J. 1995. "Lessons from the Bell Curve." *J.P.E.* 103 (5): 1091–120.
———. 2008. "Schools, Skills, and Synapses." *Econ. Inquiry* 46 (3): 289–324.
———. 2013. *Giving Kids a Fair Chance.* Cambridge, MA: MIT Press.
Heckman, James J., Carolyn Heinrich, and Jeffrey Smith. 2002. "The Performance of Performance Standards." *J. Human Resources* 37 (4): 778–811.
Heckman, James J., John Eric Humphries, and Tim Kautz, eds. 2014. *The Myth of Achievement Tests: The GED and the Role of Character in American Life.* Chicago: Univ. Chicago Press.
Heckman, James J., and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Econ.* 19:451–64.
———. 2014. "Fostering and Measuring Skills: Interventions that Improve Character and Cognition." In *The Myth of Achievement Tests: The GED and the Role of*

Character in American Life, edited by J. J. Heckman, J. E. Humphries, and T. Kautz, Chicago: Univ. Chicago Press.
Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *A.E.R.* 103 (6): 2052–86.
Herrnstein, Richard J., and Charles Murray. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.
Hossain, Tanjim, and Ryo Okui. 2013. "The Binarized Scoring Rule." *Rev. Econ. Studies* 80 (3): 984–91.
Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.
John, Oliver P., and Sanjay Srivastava. 1999. "The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives." In *Handbook of Personality: Theory and Research*, vol. 2, edited by L. A. Pervin and O. P. John. New York: Guilford.
Kadane, Joseph B., and Baruch Fischhoff. 2013. "A Cautionary Note on Global Recalibration." *Judgment and Decision Making* 8 (1): 25–27.
Kansup, Wanlop, and A. Ralph Hakistan. 1975. "A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-Choice Tests: I. Scoring Procedures." *J. Educ. Measurement* 12:219–30.
Karay, Yassin, Stefan K. Schauber, Christoph Stosch, and Katrin Schüttpelz-Brauns. 2015. "Computer Versus Paper—Does It Make Any Difference in Test Performance?" *Teaching and Learning Medicine* 27 (1): 57–62.
Kirsh, David, and Paul Maglio. 1994. "On Distinguishing Epistemic from Pragmatic Action." *Cognitive Sci.* 18 (4): 513–49.
Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. "A Smooth Model of Decision Making Under Ambiguity." *Econometrica* 73 (6): 1849–92.
Koehler, Roger A. 1971. "A Comparison of the Validities of Conventional Choice Testing and Various Confidence Marking Procedures." *J. Educ. Measurement* 8:297–303.
———. 1974. "Overconfidence on Probabilistic Tests." *J. Educ. Measurement* 11:101–8.
Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff. 2016. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." *American Econ. J. Econ. Policy* 8 (4): 183–219.
Lusardi, Annamaria, and Olivia S. Mitchell. 2008. "Planning and Financial Literacy: How Do Women Fare?" *A.E.R. Papers and Proc.* 98 (2): 413–17.
Lynn, Richard, and Paul Irwing. 2004. "Sex Differences on the Progressive Matrices: A Meta-Analysis." *Intelligence* 32:481–98.
Manski, Charles F. 2004. "Measuring Expectations." *Econometrica* 72 (5): 1329–76.
Matheson, James E., and Robert L. Winkler. 1976. "Scoring Rules for Continuous Probability Distributions." *Management Sci.* 22 (10): 1087–96.
Matzen, Laura E., Zachary O. Benz, Kevin R. Dixon, Jamie Posey, James K. Kroger, and Ann E. Speed. 2010. "Recreating Raven's: Software for Systematically Generating Large Numbers of Raven-Like Matrix Problems with Normed Properties." *Behavior Res. Methods* 42 (2): 525–41.
McDaniel, Tanga M., and E. Elisabet Rutström. 2001. "Decision Making Costs and Problem Solving Performance." *Experimental Econ.* 4:145–61.
McKelvey, Richard D., and Talbot Page. 1990. "Public and Private Information: An Experimental Study of Information Pooling." *Econometrica* 58 (6): 1321–39.

Moore, Don A., and Paul J. Healy. 2008. "The Trouble with Overconfidence." *Psychological Rev.* 115 (2): 502–17.
Murphy, Alan. 1966. "A Note on the Utility of Probabilistic Predictions and the Probability Score in the Cost-Loss Ratio Decision Situation." *J. Appl. Meteorology* 5:534–37.
Niederle, Muriel. 2015. "Gender." In *Handbook of Experimental Economics*, vol. 2, edited by J. Kagel and A. E. Roth. Princeton, NJ: Princeton Univ. Press.
Niederle, Muriel, Carmit Segal, and Lise Vesterlund. 2013. "How Costly Is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness." *Management Sci.* 59 (1): 1–16.
Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Q.J.E.* 122:1067–101.
———. 2010. "Explaining the Gender Gap in Math Test Scores: The Role of Competition." *J. Econ. Perspectives* 24 (2): 129–44.
Pearl, Judea. 1978. "An Economic Basis for Certain Methods of Evaluating Probabilistic Forecasts." *Internat. J. Man-Machine Studies* 10:175–83.
Planer, Ronald J., and Kim Sterelny. 2021. *From Signal to Symbol: The Evolution of Language*. Cambridge, MA: MIT Press.
Prelec, Drazen. 1998. "The Probability Weighting Function." *Econometrica* 66:497–527.
Raven, J., J. C. Raven, and J. H. Court. 1962. *Coloured Progressive Matrices*. London: HK Lewis.
———. 1993. *Raven Manual: Section 1. General Overview*. Oxford: Oxford Psychologist Press.
———. 1998. *Raven Manual: Section 4. Advanced Progressive Matrices*. Oxford: Oxford Psychologist Press.
———. 2000. *Raven Manual: 3. Standard Progressive Matrices*. Oxford: Oxford Psychologist Press.
Rippey, Robert M. 1968. "Probabilistic Testing." *J. Educ. Measurement* 5:211–15.
———. 1970. "A Comparison of Five Different Scoring Functions for Confidence Tests." *J. Educ. Measurement* 7:165–70.
Roth, Alvin E., and Michael W. K. Malouf. 1979. "Game-Theoretic Models and the Role of Information in Bargaining." *Psychological Rev.* 86:574–94.
Rutström, E. Elisabet, and Nathaniel T. Wilcox. 2009. "Stated Beliefs versus Empirical Beliefs: A Methodological Inquiry and Experimental Test." *Games and Econ. Behavior* 67:616–32.
Savage, Leonard J. 1971. "Elicitation of Personal Probabilities and Expectations." *J. American Statis. Assoc.* 66 (December): 783–801.
———. 1972. *The Foundations of Statistics*. 2nd ed. New York: Dover.
Schlag, Karl H., and Joël van der Weele. 2013. "Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk-Neutrality." *Theoretical Econ. Letters* 3:38–42.
Segal, Carmit. 2012. "Working When No One Is Watching: Motivation, Test Scores, and Economic Success." *Management Sci.* 58 (8): 1438–57.
Seidenfeld, Teddy. 1985. "Calibration, Coherence and Scoring Rules." *Philosophy Sci.* 52:274–94.
Shapiro, Lawrence, ed. 2014. *The Routledge Handbook of Embodied Cognition*. London: Routledge.
Shuford, Emir H., Albert Arthur, and H. Edward Massengill. 1966. "Admissible Probability Measurement Procedures." *Psychometrika* 31 (June): 124–45.
Shurchkov, Olga. 2012. "Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints." *J. European Econ. Assoc.* 10 (5): 1189–213.

Sloman, Steven, and Philip Fernbach. 2017. *The Knowledge Illusion: Why We Never Think Alone*. New York: Riverhead.
Smith, Cedric A. B. 1961. "Consistency in Statistical Inference and Decision." *J. Royal Statis. Soc.* 23:1–25.
Smith, Vernon L. 1982. "Microeconomic Systems as an Experimental Science." *A.E.R.* 72 (5): 923–55.
ter Weel, Bas. 2008. "The Noncognitive Determinants of Labor Market and Behavioral Outcomes." *J. Human Resources* 43 (4): 729–37.
Vygotskij, Lev S. 1962. *Thought and Language*. Cambridge, MA: MIT Press.
Wainer, Howard, Eric T. Bradlow, and Xiaohui Wang. 2007. *Testlet Response Theory and Its Applications*. New York: Cambridge Univ. Press.
Wang, Ke, and Zhendong Su. 2015. "Automatic Generation of Raven's Progressive Matrices." In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, edited by Qiang Yang, 903–9. Buenos Aires: AAAI Press.
Winkler, R. L. 1996. "Scoring Rules and the Evaluation of Probabilities." *Test* 5 (1): 1–35.