

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

by

Glenn W. Harrison, Don Ross and J. Todd Swarthout[†]

July 2024

ABSTRACT

The measurement of intelligence should identify and measure an individual's subjective confidence that a response to a test question is correct. Existing measures that use multiple-choice answers are constrained in measuring responses that reflect, at best, only modal beliefs, due to the use of surveys with no extrinsic financial incentive for a truthful response *and* by not eliciting measures of confidence. We rectify both issues, and show that each matters for the measurement of intelligence. We use a canonical measure of fluid intelligence, the Raven Advanced Progressive Matrices test. We show that awareness of the confidence of responses, and being able to express them in a properly incentivized manner, plays a critical role when interpreting intelligence measures. Despite evidence that women tend to do worse than men in historically cited intelligence measures using the Raven test, they do much better than men when responding to incentivized measures of confidence. We also show that our results on gender and confidence in the face of risk have wider applications in terms of “competitiveness” and financial literacy. It is not the case that women lack the confidence to take on the risks of competition: they respond exactly as any risk averse agent should. When confidence in response to questions about financial literacy is measured correctly, “do not know” responses are not evidence that women are afraid of answering questions about financial knowledge.

Keywords: Intelligence, Belief, Risk, Confidence, Awareness, Economics of Gender
JEL Codes: D81, D83, C81, J16

[†] Maurice R. Greenberg School of Risk Sciences and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Harrison). School of Society, Politics, and Ethics, University College Cork, Cork, Ireland; School of Economics, University of Cape Town, Cape Town, South Africa; and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Ross). Department of Economics and Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University, USA; and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Swarthout). Harrison is also affiliated with the School of Economics, University of Cape Town. E-mail contacts: gharrison@gsu.edu, don.ross931@gmail.com and swarthout@gsu.edu. We are grateful to three referees, the editor, and participants at numerous presentations for constructive comments.

Table of Contents

1. Experimental Design	-9-
A. Eliciting Beliefs	-9-
B. Treatments	-11-
2. Intelligence, Confidence and Scaffolds	-13-
A. The Effects of Incentives for Accuracy and Confidence	-16-
B. Gender Effects	-18-
C. The Cognitive Scaffold of Progression	-23-
D. The Welfare Cost of Traditional Procedures	-25-
3. Gender and Confidence, Reconsidered	-27-
A. Gender and Competitiveness	-28-
B. Gender and Literacy	-39-
4. Extensions, Related Literature, and Open Issues	-43-
A. Probabilistic Testing in Education	-43-
B. Achievement Tests	-44-
C. Intrinsic Motivation	-46-
D. The Elicitation of Beliefs	-48-
E. Calibration and the Evaluation of Beliefs	-51-
F. Normative Implications	-53-
G. Origins of the Concept of Scaffolding	-55-
5. Conclusions	-58-
References	-74-
Appendix A: Instructions (Online Working Paper)	-A1-
1. Non-Salient “Pen and Paper” Treatment	-A1-
2. Eighty Tokens Treatment	-A4-
3. Single Token Treatment	-A9-
4. Scrambled Treatment	-A13-
Appendix B: Subjective Beliefs and Risk Preferences (Online Working Paper)	-A14-
Appendix C: Gender and Competitiveness (Online Working Paper)	-A19-
Appendix D: Pilot Results with Variants on the Raven Task (Online Working Paper)	-A50-
1. Time Constraints	-A50-
2. Progressively Increasing Incentives	-A51-
3. Instructions: Time Constraint Treatment	-A57-
4. Instructions: Increased Incentives Treatment	-A63-

Most measures of intelligence score responses to questions as either right or wrong. If a test is to have any interesting discriminatory power, there need to be some questions that some individuals do not answer 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 only reflect *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 dis-service to that perception to artefactually 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, under 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 (§4.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] and Dennett [2017]). Naming something stabilizes it in memory and for re-identification. 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 (Planer and Sterelny

[2021] and Dennett [2017]). 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. And better measurement of cognitive skills in general should contribute to the debate over the relative importance of cognitive and non-cognitive skills across the life-cycle of individuals and households.¹

Following Smith [1982; p. 931-938],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

¹ Hanushek and Woessmann [2008] reviews 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 non-cognitive skills over the *life-cycle of individuals*, with significant implications for the impact of early childhood programs on lifetime outcomes.

² The five precepts proposed by Smith [1982] are non-satiation 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.

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 component⁵ of this test is called Set II, and consists of 36 problems illustrated by an isomorph⁶ shown in Figure 1. Each problem

³ Borghans, Duckworth, Heckman and ter Weel [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 non-cognitive 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. As explained by Heckman and Kautz [2012; p.452], “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, for example, 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 if 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 [2000]), and the Raven Coloured Progressive Matrices test is for children or those with cognitive limitations, such as 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. 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 Civelli and Deck [2018], Matzen et al. [2010] and Wang and Su [2015]. And there are many instances in which “Raven-like” tests are developed and used, such as the 8 questions developed by Borghans et al. [2009] and used “to measure IQ” (p. 652).

consists of a block of 9 images, arrayed as a 3×3 matrix, with one image deleted. The subject is presented with 8 possible solutions to fill in the deleted image, and is instructed to select the 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 “paper and pen” 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 only have 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 non-salient versions of the RAPM. By “non-salient” 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

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

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 non-monotonic blips are evident, and are 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 did worse in 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 “only” offered \$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 non-salient intelligence test so that we elicit subjective beliefs of individuals in a salient, incentivized manner, 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,

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] provide a taxonomy of five types of rules used in the RAPM, and list (p. 431) which combinations of rules apply in each problem.

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).

Arthur and Massengill [1966] and Savage [1971]. The broader normative case for this approach was explained elegantly by Savage [1971; p. 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 1 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 only report 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 2 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 only report their modal beliefs. These results overturn long-standing claims that 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 3 considers our striking result from Section 2 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 4 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 5 draws general conclusions, with a normative emphasis. Our conclusions on gender and confidence provide striking contrasts to received claims, and sharply re-orient 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.

1. 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 8 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 8 possible solutions, allowing the individual to allocate 80 tokens across the eight solution “bins.” The initial

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 *ten* 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 §4.D.

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 to a previously completed question and re-allocate 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 there are no re-allocations 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 “non-degenerate” allocations are shown in Figure 4. The top panel of Figure 4 shows a possible response to

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; p. 1336] and Rutström and Wilcox [2009; p. 620]. We regard these alternatives as being inferior to real-time appreciation of the salient rewards at play in QSR belief elicitation when contemplating alternative possible reports.

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 8 possible solutions. If the subject allocates 10 tokens to each of the 8 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 non-salient, non-computerized “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 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 1 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 **Eighty Tokens Progressive** and **Eighty Tokens Scrambled** treatments we give subjects 80 tokens to allocate for each question. In all other aspects the **Eighty Tokens Progressive** and **One Token Progressive** treatments are identical, as are the **Eighty Tokens Scrambled** and **One Token Scrambled** treatments. Each question is worth a maximum of \$2, resulting in a maximum possible earnings of \$72 for the task. A total of 95, 92, 67 and 61 subjects participated in the **One Token Progressive**, **Eighty Tokens Progressive**, **One Token Scrambled** and **Eighty 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: “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 6 subjects came within 5 minutes of that constraint, and 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 to traditional measurements that do not reflect confidence. We explore the generality of this important

¹⁴ In Appendix D (online) we report results from a pilot experiment in which we constrained subjects to only have 40 minutes for the **Eighty Tokens** task. There are no significant effects on performance, although suggestions that there might be significant effects for some demographic variables, such as gender.

finding in a series of secondary experiments in Section 4. 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 re-use the Raven task as the 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, **Eighty Tokens Piece Rate**, **Eighty Tokens Tournament** and **Eighty 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 §3.A.

2. 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 emails that contained no information about the task or expected earnings, and assigned randomly to treatments. Each subject received a participation payment of \$7 in addition to task-specific earnings.

Upon arrival at the lab, subjects completed the IRB consent process and then received task instructions in both audio-video and print formats. By presenting the instructions with a pre-recorded 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

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

watch instructions. Appendix A (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 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 happen 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 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 non-uniformly to some answers.¹⁹

At an aggregate level, pooling over the 36 questions for each individual, we are interested in

¹⁶ Video instructions are available at <https://cear.gsu.edu/gwh/raven/>.

¹⁷ For complete details of this prior experiment see Harrison et al. [2022c].

¹⁸ This last theoretical result is established by Harrison et al. [2017]. Essentially, risk averse EUT subjects provide reports which are flattened with respect to their true beliefs, so as to reduce variability over possible outcomes with positive subjective probability. This is a second-order adjustment in reports compared to the first-order adjustment in the binary case, where risk averse individuals provide reports closer to $\frac{1}{2}$ so as to reduce variability over the two possible outcomes.

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

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, and measures 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 $k = 1, \dots, K$. 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, $\{r_1, r_2, \dots, r_k, \dots, r_K\}$ such that $r_k \geq 0 \forall k$ and $\sum_{i=1}^K (r_i) = 1$. If k is the interval in which the actual value lies, then the payoff score is defined by Matheson and Winkler [1976; p.1088, equation (6)]: $S = (2 \times r_k) - \sum_{i=1}^K (r_i)^2$. 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 includes all possible reports, including the correct one.²⁰ To ensure complete generality, and avoid any decision maker facing losses, allow some

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 tradeoff from the last case is that one can always ensure a score of $1/4$, but there is an incentive to provide less diffuse reports, and that incentive is the possibility of a score of 1.

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, $\alpha + \beta [(2 \times r_k) - \sum_{i=1 \dots K} (r_i)^2]$, 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 < 0.001).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 paper-and-pen *versus* computerized versions of the same test, although only a few controlling for measures of pre-test ability. The available evidence suggests no differences in performance for students at a university level.²²

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

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 an 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 et al. [2017] find no differences for high-school students, but significant differences for elementary-school students and modest differences for middle-school students. Although gender had no effect on the comparison, having English as a primary language did make some difference.

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 non-salient: 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 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 sub-sample 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

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

statistically significantly different from zero ($p = 0.25$). 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 **Eighty Tokens Progressive** condition compared to the **One Token Progressive** condition is +7.5 percentage points (p -value = 0.026). 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, by the horizontal dashed line, that a uniform report in the **Eighty 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

²⁴ In Appendix D (online) 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 be \$2 for questions 1 through 12, \$3 for questions 13 through 24, and \$5 for questions 25 through 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.

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 panel A of Figure 7. 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 panels B and C we show the *effect of the treatment condition* on the Accuracy or Efficiency associated with each demographic. In panel 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 **Eighty 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. Panel C of Figure 7 shows these effects on Efficiency, comparing the **Eighty Tokens Progressive** condition to the **One Token Progressive** condition. In this instance, compared to panel 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 non-extreme 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 [2005] 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 5180 males and 4451 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 mis-characterize 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*

²⁵ His measure of risk intelligence [2012; Appendix 1] used 50 hypothetical true/false survey questions.

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 [2005; Table 2, p.490] actually report corrected estimates of d and their 95% confidence intervals.

*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 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 **Eighty 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, #2, who allocated all 80 tokens to the correct answer; one subject, #1, who was overconfident with

²⁷ We only ever use the expression “overconfidence” to refer to excessive certainty about the accuracy of one’s belief. Harrison and Swarthout [2022] 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 to 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.

respect to a wrong answer, allocating all 80 tokens to answer 2; one subject, #5, who had bimodal beliefs and allocated 40 tokens to each of two answers, one of which was correct; and two subjects, #3 and #4, who reported completely diffuse beliefs.

Panel A of Figure 10 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 8 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 **Eighty Tokens Progressive** condition compared to 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. Panel B of Figure 10 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 **Eighty Tokens Progressive** condition significantly and steadily reduced their time allocation, consistent in panel A of Figure 10 with increased use of the **Diffuse** allocation. On the other hand, and by comparison, panel B of Figure 10 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 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

²⁸ The patterns in Figure 6 and Figure 10 allow one to draw inferences about earnings per minute of time allocated to each question. Although Figure 6 shows realized earnings normalized by maximum earnings, those maximum earnings were the same across the **One Token Progressive** and **Eighty 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 **Eighty 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 **Eighty 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 Figure 10 for these questions.

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 12 shows the results of applying this **Scrambled** variant in the **One Token Scrambled** and **Eighty Tokens Scrambled** treatments. We focus on Efficiency; results on Accuracy show a predictable drop compared to 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 **Eighty 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 **Eighty 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 to panel C of Figure 7, 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.

²⁹ The changes are shown in §C of Appendix A (online).

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 5 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 *agreeableness* of the person: whether they are friendly towards 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.

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 expected welfare cost *to a respondent* of being forced to report their subjective beliefs in a way that requires that they only report 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 **Eighty 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 8 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 non-negative, 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.

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. Panel

³¹ The procedures to recover beliefs are reviewed in §4.D and Appendix B (online).

³² We report normative evaluations with EUT in Appendix B: see Figure B13. The relative merits of EUT or RDU as a normative metric are discussed by Harrison and Ross [2018][2023].

A of Figure 13 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 to 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 **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 6 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. Panel B of Figure 13 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.

Panels A and 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

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

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.

3. Gender and Confidence, Reconsidered

Our striking result that women respond positively in terms of our measure of fluid intelligence leads us to consider if 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 re-examine 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

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 paper and a pen), and hard otherwise.

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 §5.

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

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 5 two-digit numbers correctly.

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 to 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 if 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.

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 are 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 to the point earnings of

3 other subjects selected at random. Only the subject with the best score would receive compensation³⁶ if this part was selected for payment, but the potential payoff was 4 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 to the point earnings from part 2 of the other 3 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 one penny 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 be the highest of the 4 in her tournament group. A natural prior would be to assume a $\frac{1}{4}$ 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

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.

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 $\frac{1}{4}$ by design.

account the risk of winning or losing the tournament. Since the tournament gives a $\frac{3}{4}$ chance of receiving \$0 and a $\frac{1}{4}$ 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 4 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 (online) details these calculations. Using CE measures 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

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.

the tournament 73% of the time. In principle, these calculations 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 if she wants her earnings from 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 previously-generated earnings by 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* to 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

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 to the others in her tournament group of 4. 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 $\frac{1}{4}$ 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; p.134] summarize their

⁴¹ Shurchkov [2012] recognizes that there are these additional attributes from competition, but only replicates parts 1, 2 and 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], and 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 [2014; §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 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.

data as follows:

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

Recognizing that the word “overconfident” is used here to mean optimistic, this finding is crucial. The full title of Niederle and Vesterlund [2007] 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 **Eighty 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 CRRA utility function $U(x) = x^{1-\tau}/(1-\tau)$; the RDU model also assumes a Prelec [1998] probability weighting function.⁴⁴ Every subject completed a risk battery of 30

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.

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 tournament lottery is a well-defined simple lottery. Once these have been calculated, the Certainty Equivalent (CE) of each lottery can be calculated as shown in panel D of Table 2, since $u(CE) = \xi$ solves for $CE = [\xi \times (1-r)]^{1/(1-r)}$ for $\xi \in \{EU, RDU\}$. 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* choice in bold. Finally, panel F repeats the same calculations using the 1 token data, implicitly replacing the data shown in Table in panels A and B. For the 1 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

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.

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 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.

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 1 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 1 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 $\frac{1}{4}$ when there was only 1 token. On average men hold beliefs that are right at the diffuse prior of $\frac{1}{4}$ 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 1 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 that the reported belief of winning the tournament exceeded $\frac{1}{4}$, 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

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.

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

⁴⁹ Detailed results are presented in §3 of Appendix C (online). All estimates refer to average marginal effects from probit regressions.

more likely to realize a welfare gain from their decisions (p -value = 0.09). The second is that belief optimists tended to be 18 percentage points *less likely* to realize a welfare gain (p -value = 0.07), 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 < 0.001). 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, Niederle, Segal and Vesterlund [2013] and Balafoutas and Sutter [2012] 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 §4 of Appendix C (online), 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 sub-literature 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, Lusardi, Alessie and van Rooij [2017; p.257] conclude⁵² 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.” The same theme is echoed by Lusardi and Mitchell [2008] and Bucher-Koenen, Alessie, Lusardi and van Rooij [2021]. 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

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 (p. 260, 261, 262, respectively).

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 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 if 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 re-allocate 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, Alessie, Lusardi and van Rooij [2021]. 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, Harrison, Lau and Swarthout [2015] and Harrison et al. [2022c].

We illustrate the effects of using this mode of response with a literacy question that is a direct adaptation⁵³ of one of the “Big Three” from Lusardi and Mitchell [2011], to test understanding of the

⁵³ The original question is, “Imagine that the interest rate on your savings account was 1 percent per year and inflation was 2 percent 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

effects of inflation:

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

The correct answer to this question is $\$98.98 = \$100 \times 1.01 \times 0.98$. We asked this question of 776 GSU undergraduates, using the same incentivized QSR procedure as in our Raven experiments.⁵⁴

Subjects were 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 show the pooled beliefs of men in response to this question, and the bottom panels 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 $|\mu - 98.98|$, 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 \$98.98, 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

today, Exactly the same, Less than today, Do not know, 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.

⁵⁵ This issue is highlighted by Harrison, Hofmeyr, Kincaid, Monroe, Ross, Schneider and Swarthout [2022; p. 817] in the context of evaluations of the bias of subjective beliefs about the mortality effects of

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. The top left panel shows the “poster child” of literacy: no bias, and perfect confidence with all tokens allocated to the correct response. The top right panel of Figure 18 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. The top left panel of Figure 18 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, the diffuse belief suggests to us that this is a subject who is likely to start seaching 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.

The actual subject in the bottom right panel of Figure 18 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.

4. Extensions, Related Literature, and Open Issues

Winkler [1996; §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, Ben-Simon, Budescu and Nevo [1997], Bereby-Meyer, Meyer and Budescu [2003], Bickel [2007][2010], Kansup and Hakistan [1970], Koehler [1971][1974] and Rippey [1968][1970]. 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 a measurement tool for some trait (Wainer, Bradlow and Wang [2007; ch.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; p.15ff.], Bernardo and Smith [2000; p. 72] and Bickel [2007][2010]). The implication is then to use a proper scoring rule that only uses 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 non-linear utility of test respondents, or more general forms of risk preference (e.g., Winkler [1996; p.19ff., 52ff.], Bickel [2007; §4]). We consider this issue in §4.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; p.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.

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

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.

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 drop-outs 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 [2008]).

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 to induce some risk preferences on responses. In §4.D we explain how each of these alternatives have been implemented in the literature on the elicitation of beliefs, and the tradeoffs involved in adopting each approach. Here we ask why one might want to use one or other of these options.

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.

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

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

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 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 which address, or explicitly ignore, risk preferences, each with strengths and weaknesses for different applications.

Following Manski [2014], 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 [2016]). 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 short-hand 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

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 through C6 in Appendix C (online) document this point for two subjects.

truthfully than non-incentivized 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 non-verifiable 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 non-incentivized 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 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]. Schlag and van der Weel [2013] and Hossain and Okui [2013] examine the same extension of the QSR, along with certain generalizations, renaming it a “randomized QSR” and “binarized scoring rule,” 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 [2014b] 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 advertized 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 et al. [2014a][2014b][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 add a paragraph at the end 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 et al. [2022b]). The

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 (p. 2853) 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.” 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.

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.

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 et al. [2022b].

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 “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

⁶² Consistency is usually defined for some binary event E such as 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, $P(E_{ij} \mid C_{ij} = c)$, where C_{ij} is the forecast probability, E_{ij} is the event j forecast by individual i , and c is some realized value for the forecast. By collecting a number of forecasts for a given individual i for the value c , or nearby values, mis-calibration is revealed when $P(E_{ij} \mid C_{ij} = c) \neq c$. Unconditional mis-calibration can also be evaluated, and revealed when $P(E_{ij}) \neq C_{ij}$.

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 if 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 mis-calibration 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 [2022]). Alpert and Raiffa [1982] extended the notion of calibration to this setting, by examining if the interquartile range of beliefs occurred 75% of the time or not, and if there were any realizations outside the range of beliefs with positive subjective probability.

In our case, subjective beliefs are probability mass functions defined over 3 or more alternatives that are not even ordered. One could still construct some calibration metric, such as evaluating if modal beliefs occurred with the subjective probability assigned to the mode(s), or if any alternative occurred that had been assigned zero subjective probability. But at this point, the idea that some simple

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.

“calibration metric” can be devised becomes problematic. And the general conceptual reasons for wanting to avoid notions of calibration remain.

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] and 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

⁶⁴ Camerer and Hogarth [1999] provide an explicit statement of this perspective on “produced” cognition. Major elaborations to estimate the multi-stage production function for cognitive *and* non-cognitive skills have been provided by Cunha and Heckman [2007][2008] and Cunha, Heckman and Schennach [2010].

⁶⁵ Cawley, Conneely, Heckman and Vytlačil [1997] directly challenge claims about the immutability of measures of cognitive ability, stemming from the debates over *The Bell Curve*. And Cawley, Heckman and Vytlačil [2001] stress how problematic it is to rigorously tease apart the tight correlation of schooling and measures of “innate” cognitive ability.

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 towards 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 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; p.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 ter Weel [2008], Borghans, ter Weel and Weinberg [2008], Heckman, Pinto and Savelyev [2013] and Heckman and Kautz [2014], for example. 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, p. 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 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

⁶⁶ The idea is simple, but difficult to implement. Berg, Daley, Dickhaut and O’Brien [1986] extended the BLP, which risk-neutralizes responses, to allow for a non-linear 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.

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 artefactually 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.

5. 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 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

⁶⁷ As noted earlier, this *general* point has been recognized in literature reviewed by Borghans, Duckworth, Heckman and ter Weel [2008], Borghans, Meijers and Ter Weel [2008], Segal [2012] and Chen et al. [2020]. We do not know of any evidence for complete batteries in the domain of fluid intelligence.

childhood education for children from disadvantaged households.⁶⁸

In fact, our elicitation interface itself is our general scaffolding treatment, as noted earlier. Dennett [1991; p. 218-224] 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 “pre-existing” 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; p. 800]. If someone is to have a derived demand for costly cognitive scaffolds with which to produce intelligence and cognition, they need to be aware that scaffolds could be valuable to them.

⁶⁸ For example, see Heckman [2013] and Levitt, List, Neckermann and Sadoff [2016].

Figure 1: Raven-Like Problem

Source: P.A. Carpenter, M.A. Just and P. Shell,
What One Intelligence Test Measures: A Theoretical Account of
the Processing in the Raven Progressive Matrices Test,
Psychological Review, 97(3), 1990, 404-431

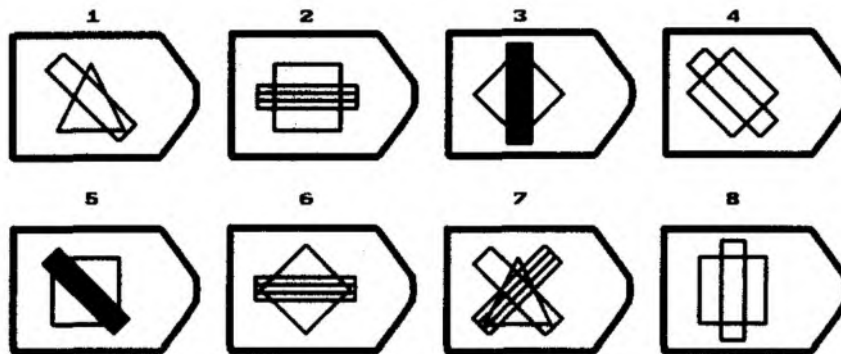
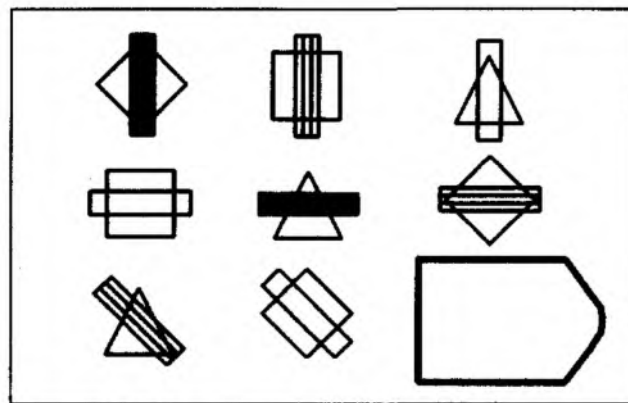


Figure 2: Raven Scores with No Salient Rewards

Non-GSU data from W. Arthur, Jr., and D.V. Day,
Development of a Short Form for the Raven Advanced Progressive Matrices Test,
Educational and Psychological Measurement, 54(2), Summer 1994, 394-403,
and

G.E. Gignac, A Moderate Financial Incentive
Can Increase Effort, But Not Intelligence Test Performance in Adult Volunteers,
British Journal of Psychology, 109(3), 2018, 500-516

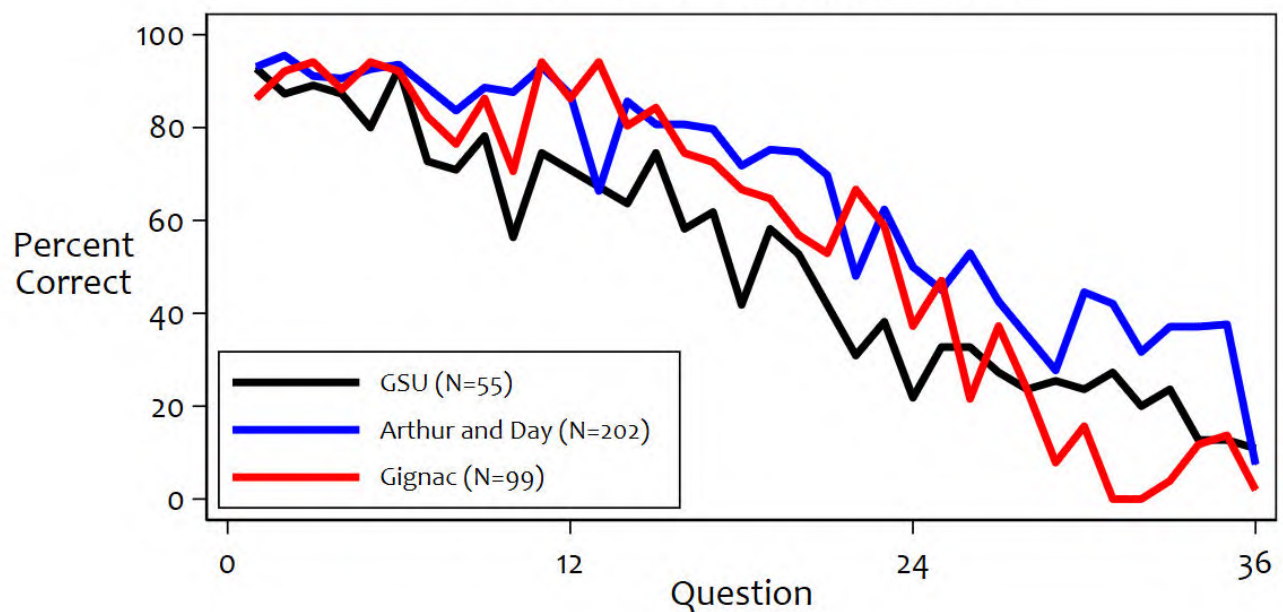


Figure 3: Belief Elicitation Interface:
Initial Display and Completely Correct Response

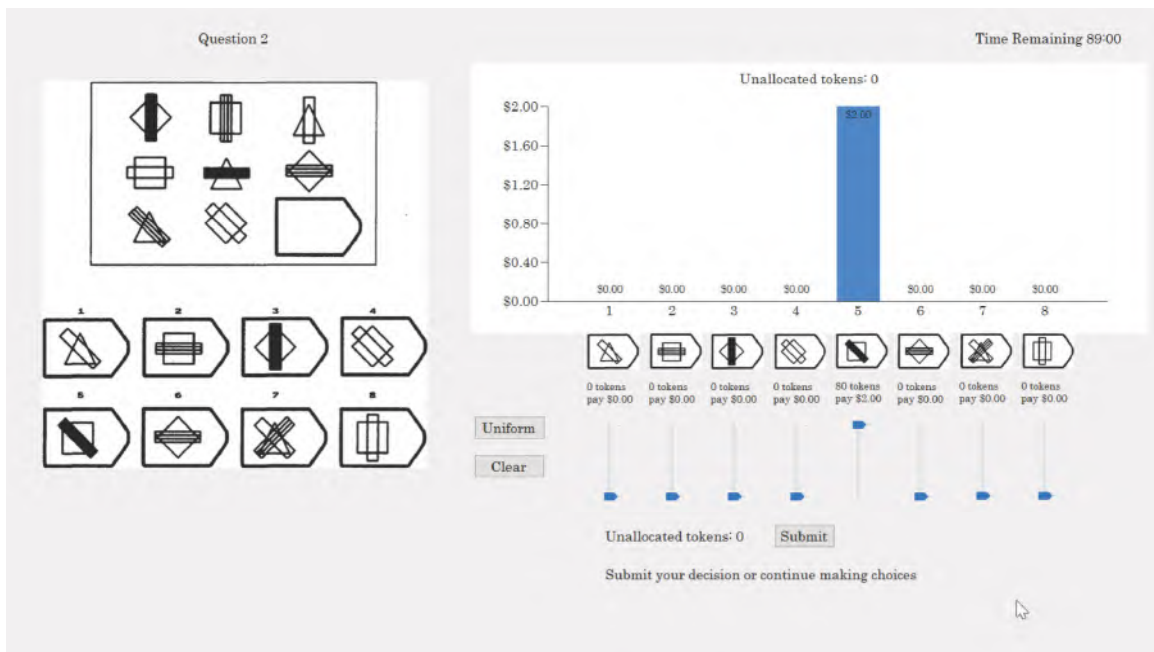
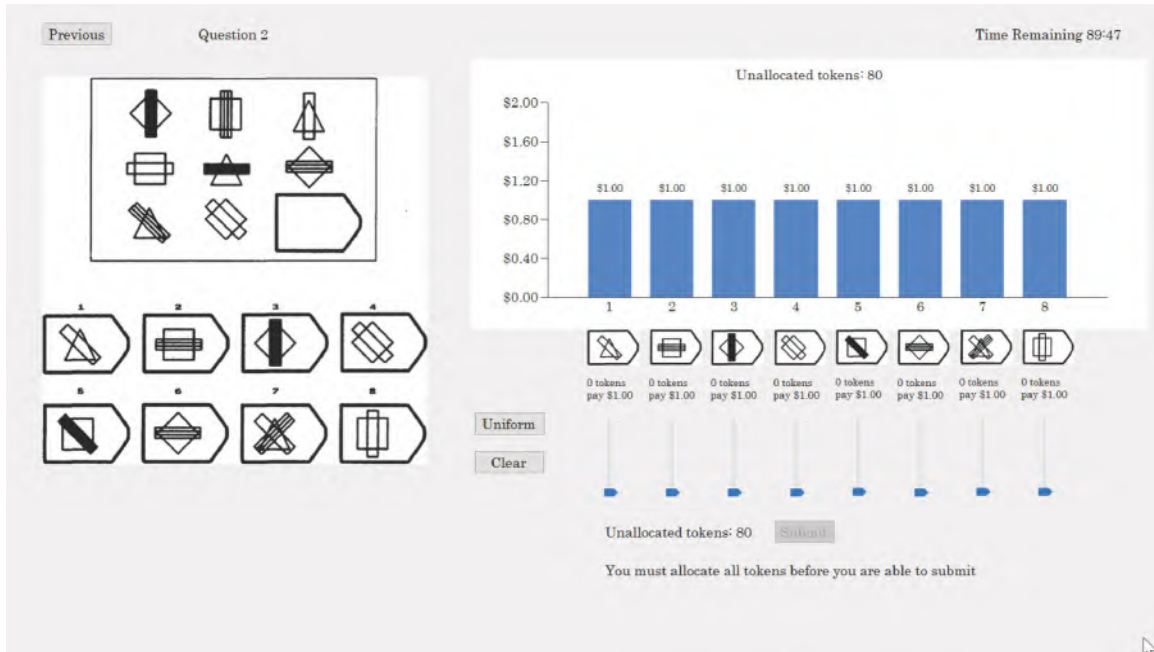


Figure 4: Belief Elicitation Interface: Two Possible Displays of Confidence

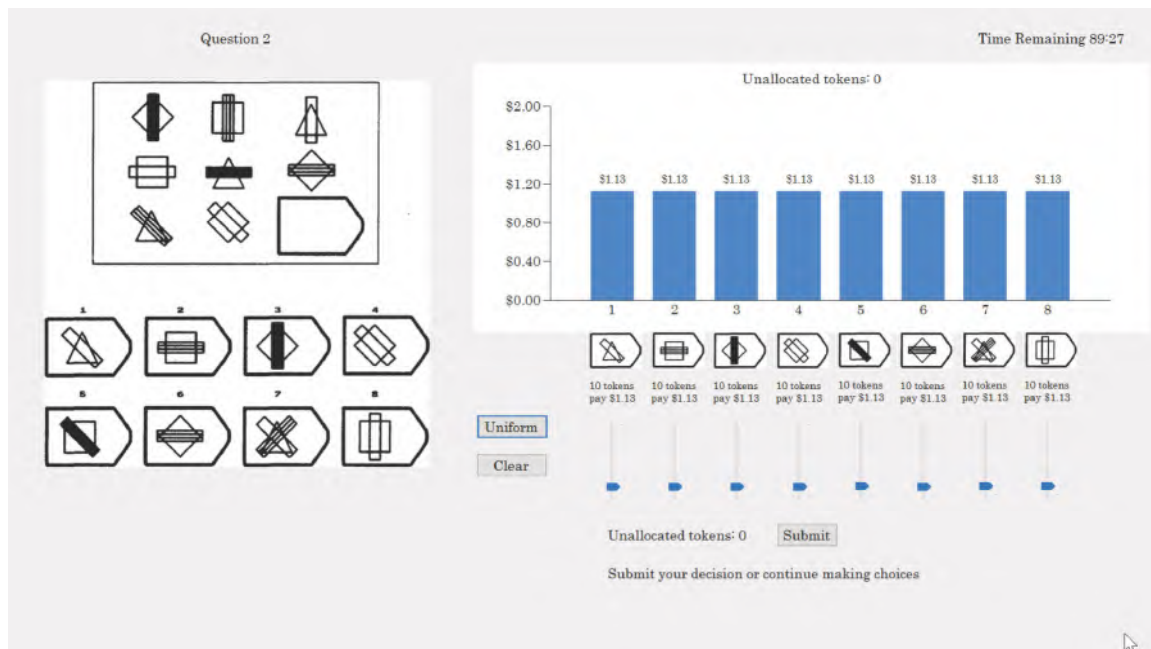
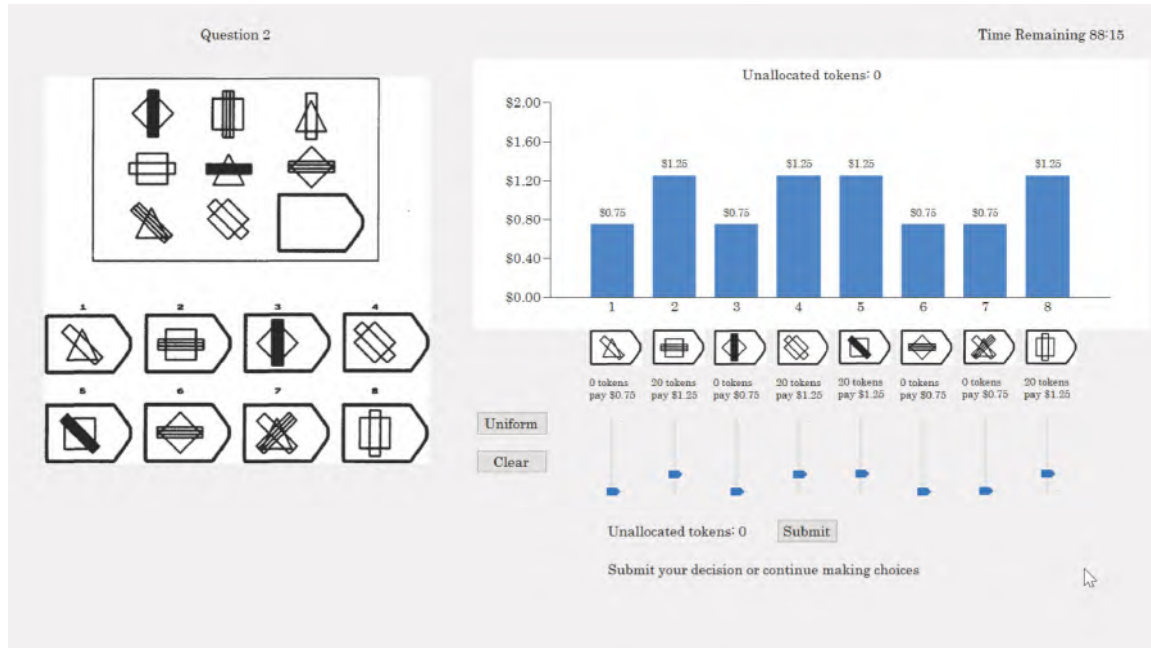


Table 1: Experimental Treatments

Treatment	Presentation of Questions	Sample Size	Response format	Maximum Reward per Question
A. Fluid Intelligence, Confidence and Scaffolds (§2 and §3)				
Baseline (Traditional)	Progressive	55	No QSR; pen & paper	\$0
One Token Progressive	Progressive	95	QSR; software	\$2
Eighty Tokens Progressive	Progressive	82	QSR; software	\$2
One Token Scrambled	Scrambled	67	QSR; software	\$2
Eighty Tokens Scrambled	Scrambled	61	QSR; software	\$2
B. Competitiveness, Confidence and Welfare (§4.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
Eighty Tokens Piece Rate	Scrambled	45	QSR; software	\$2
Eighty Tokens Tournament	Scrambled	45	QSR; software	\$8
Eighty Tokens Select	Scrambled	45	QSR; software	\$2 or \$8

Figure 5: Average Accuracy and Incentives with Progressive Raven Problems

Accuracy measured by percent of tokens allocated to the correct answer
Beliefs assumed to be the same as elicited reports

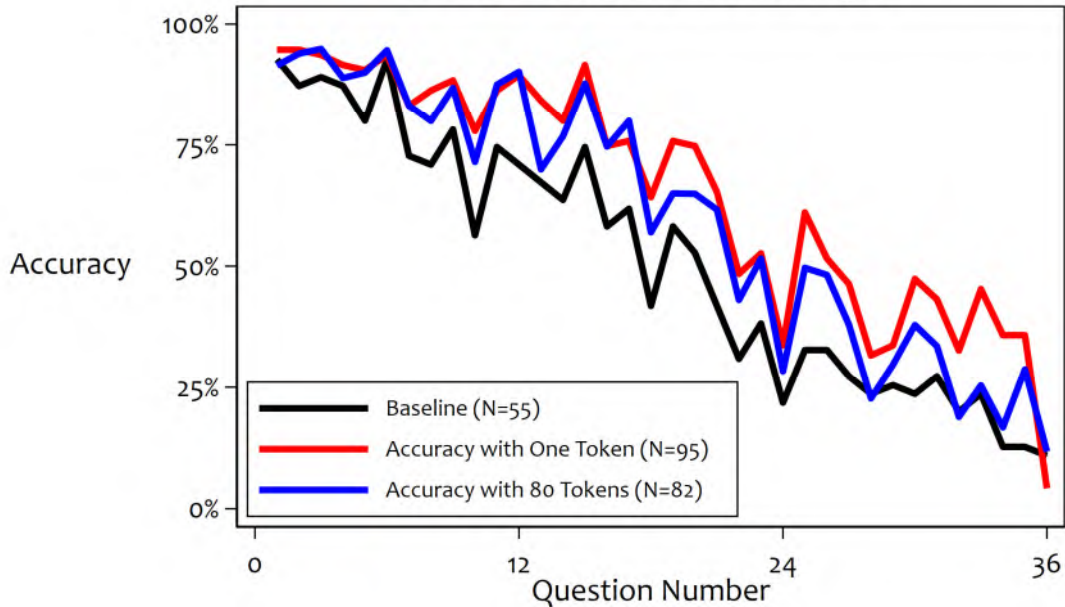


Figure 6: Average Efficiency with Progressive Raven Problems

Efficiency measured by percent of realized earnings compared to maximum earnings
Uniform token allocation ensures earning \$1.13 and efficiency of 56.5%
Beliefs assumed to be the same as elicited reports

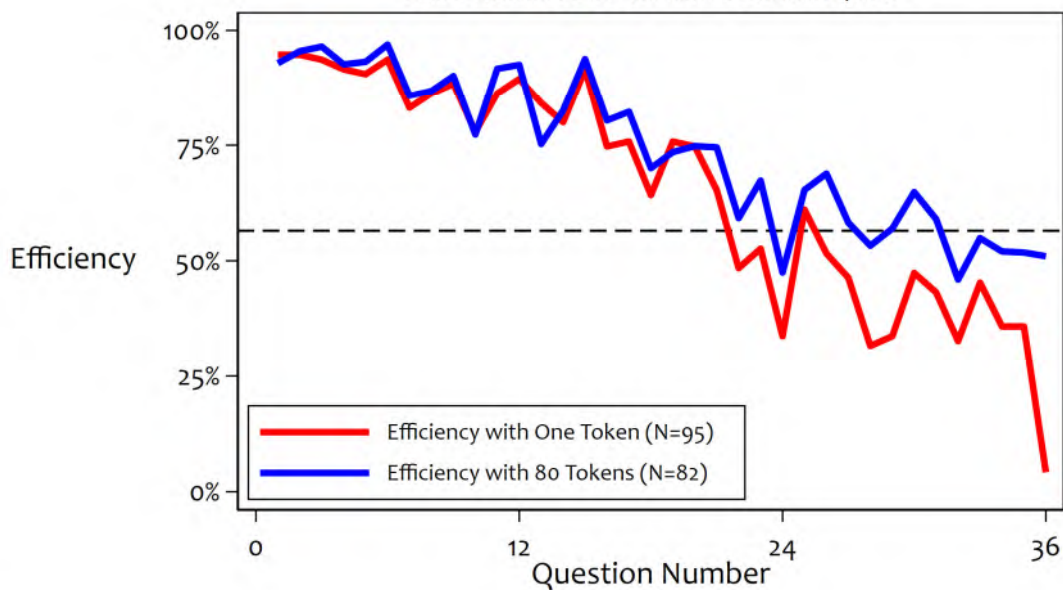


Figure 7: Effects of Demographics with Progressive Raven Problems

All results used the original progressive order
Average marginal effects from Fractional Regression

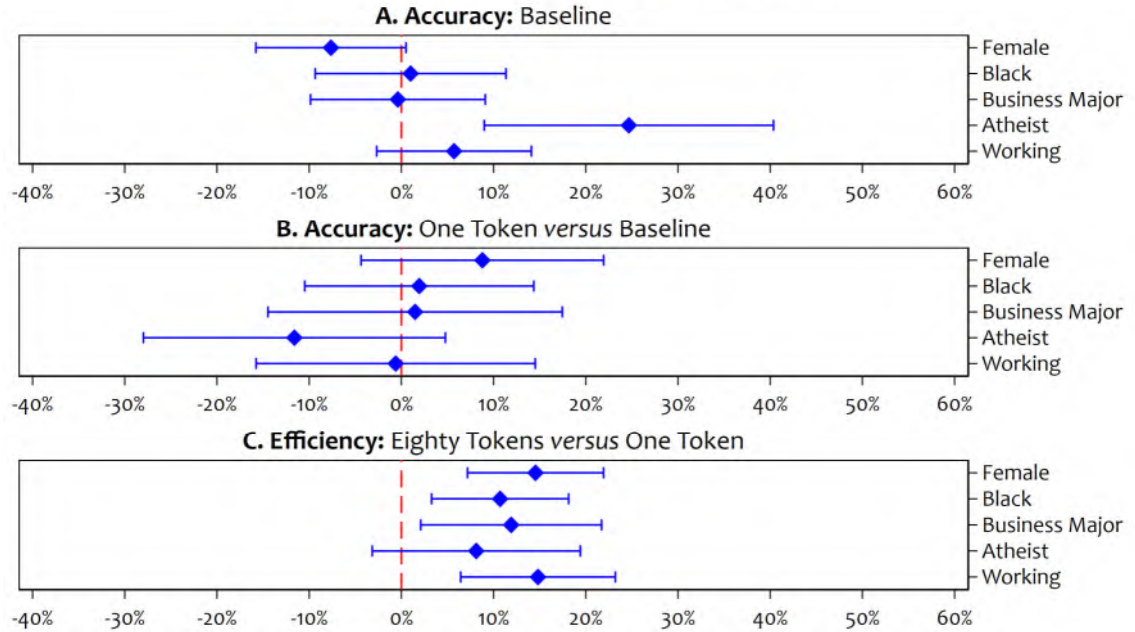


Figure 8: Allocation of Tokens by Raven Question

Pooled over incentivized responses in the Eighty Tokens Progressive condition

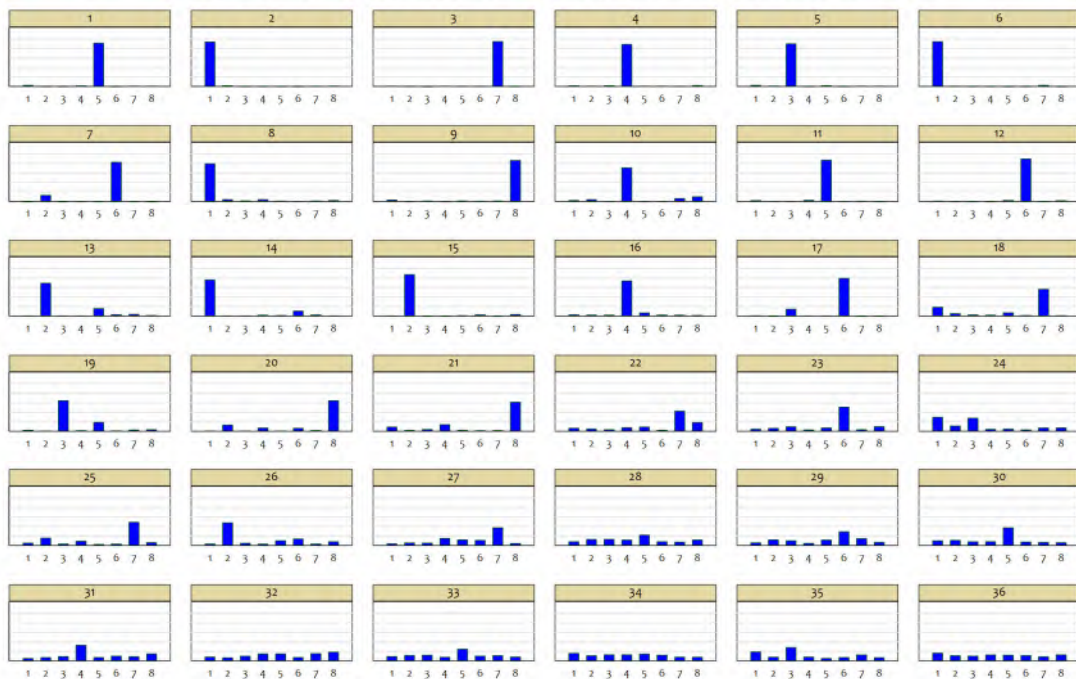


Figure 9: Allocation of Tokens in Raven Question #20

Incentivized responses in the Eighty Tokens Progressive condition. The correct answer is 8

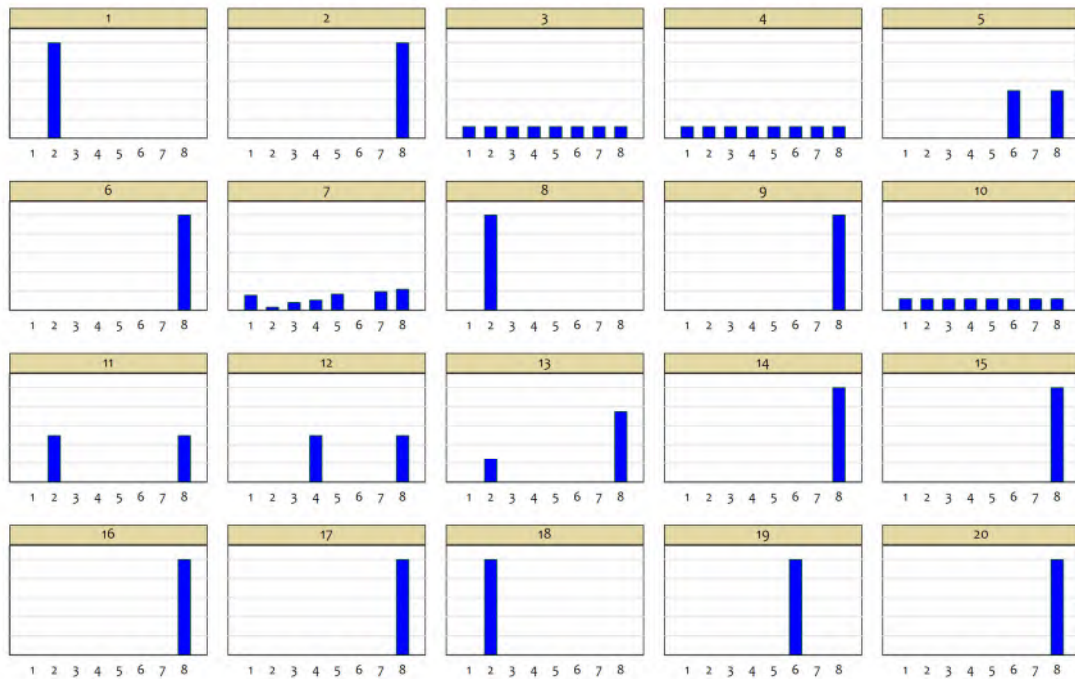


Figure 10: Time Trends of Time for Response and for Type of Response

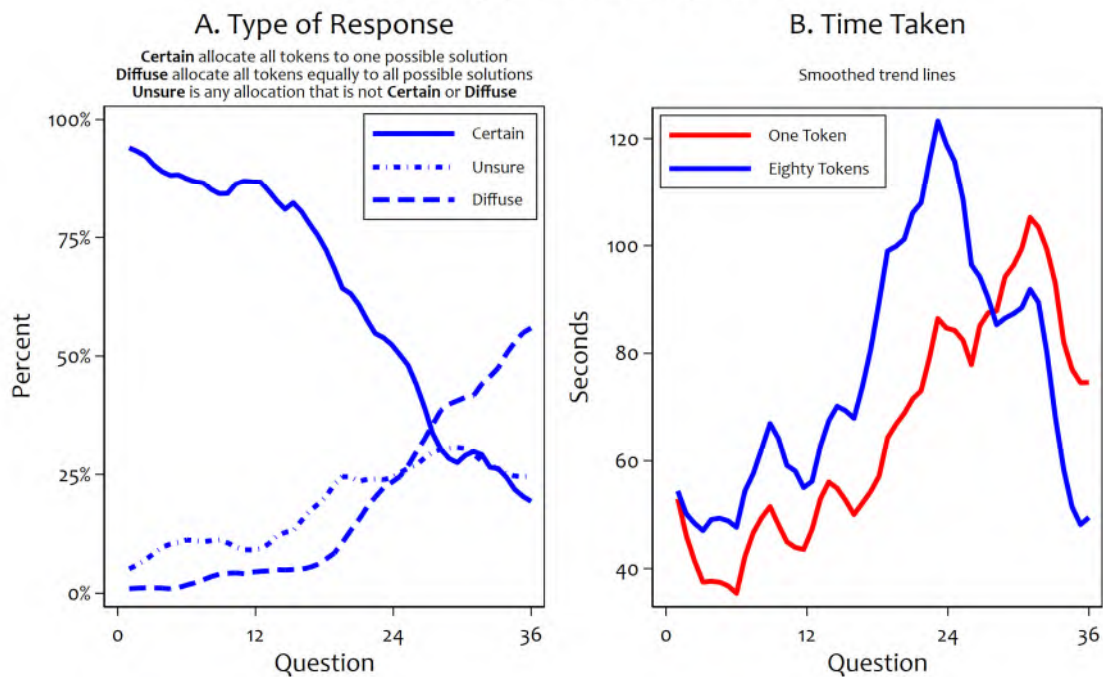


Figure 11: Average Efficiency with Scrambled Raven Problems

Efficiency measured by percent of realized earnings compared to maximum earnings
 Uniform token allocation ensures earning \$1.13 and efficiency of 56.5%
 Question number from original progressive order

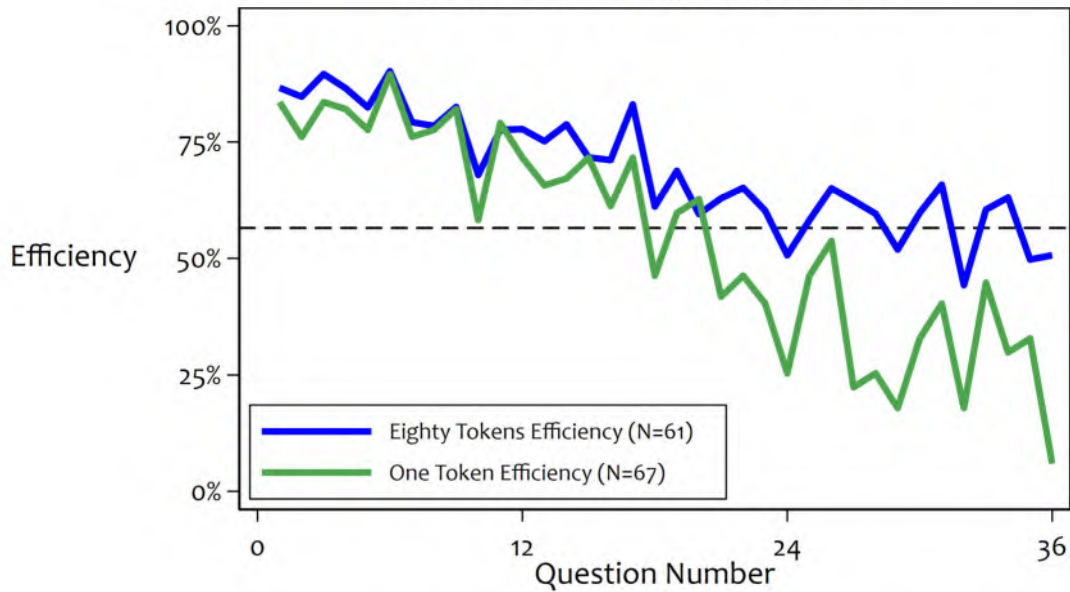


Figure 12: Effects of Eighty Tokens on Efficiency with Scrambled Raven Problems

Average marginal effects of 80 Tokens from Fractional Regression
 Pooling 80 token and 1 token results
 Solely comparing results within the Scrambled Treatment

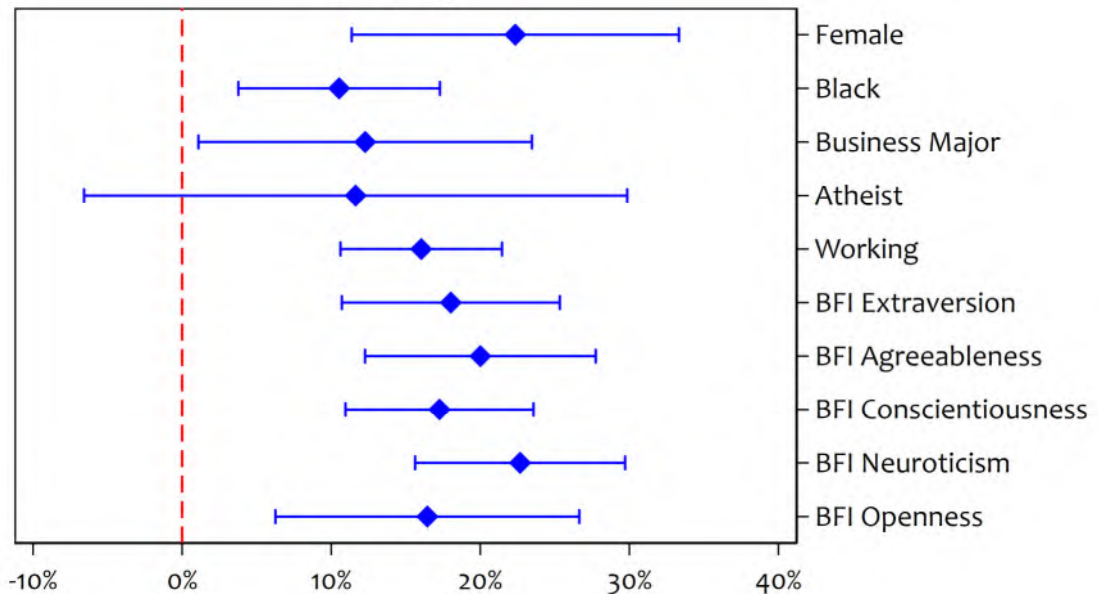


Figure 13: Expected Welfare Cost to Respondent of Being Required to Only Report Modal Belief

Evaluated using RDU risk preferences of each subject
Using recovered beliefs from **Eighty Tokens** task

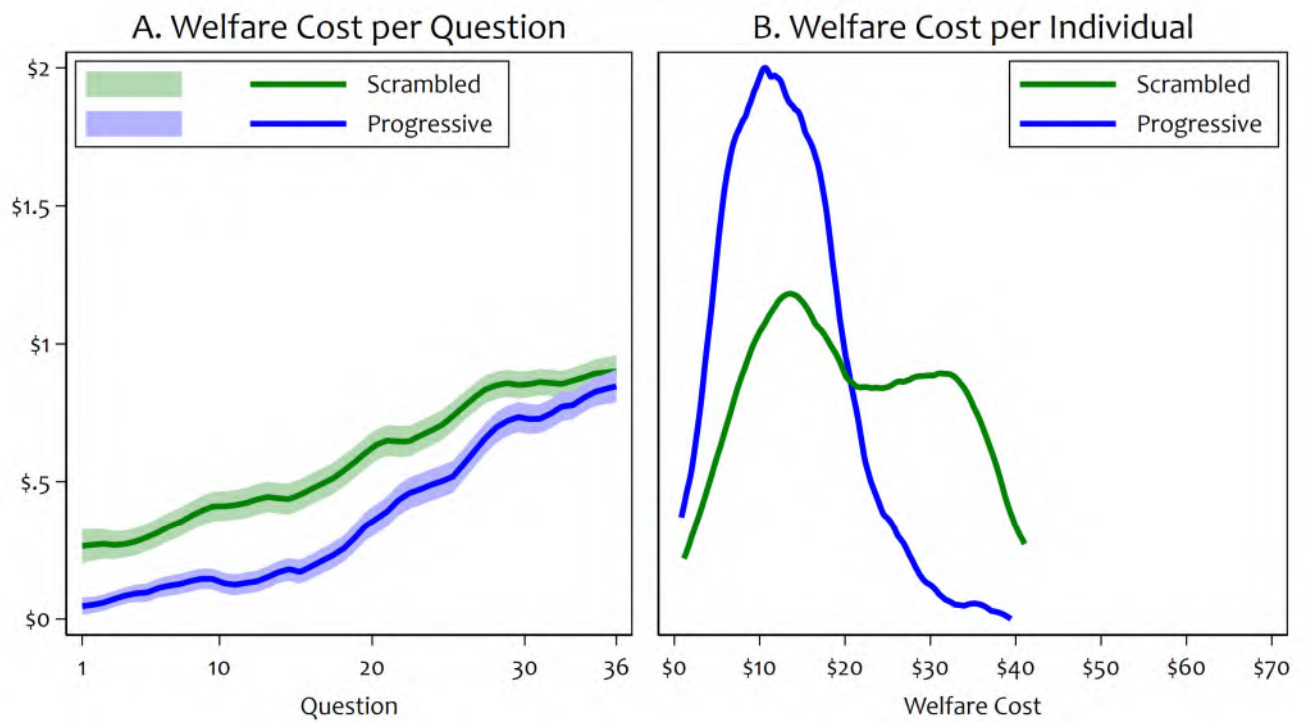


Table 2: Welfare Evaluation of Willingness to Compete Decisions

Model of Risk Preference	Risk Neutral		EUT		RDU	
Gender	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	0.26	0.28	0.26	0.28	0.26	0.28
Tournament Lose	0.74	0.72	0.74	0.72	0.74	0.72
C. Risk Preference Parameters						
Utility parameter α	0	0	0.6	0.56	0.71	0.7
PWF parameter η	1	1	1	1	0.97	0.9
PWF parameter φ	1	1	1	1	1	0.92
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	\$0.87	\$0.98	\$0.09	\$0.18
E. Expected Welfare Gain for 80 Tokens						
Piece Rate Compensation	-\$0.10	-\$0.24	+\$0.50	+\$0.38	+\$1.28	+\$1.18
Tournament Compensation	+\$0.10	+\$0.24	-\$0.50	-\$0.38	-\$1.28	-\$1.18
F. Expected Welfare Gains for 1 Token						
Piece Rate Compensation	+\$0.11	+\$0.20	+\$0.51	+\$0.62	+\$0.99	+\$1.09
Tournament Compensation	-\$0.11	-\$0.20	-\$0.51	-\$0.62	-\$0.99	-\$1.09

Note: parameter values for payoffs, subjective probabilities and risk preferences are documented in Appendix C (online).

Figure 14: Ex Ante Welfare Effects of the Decision to Compete

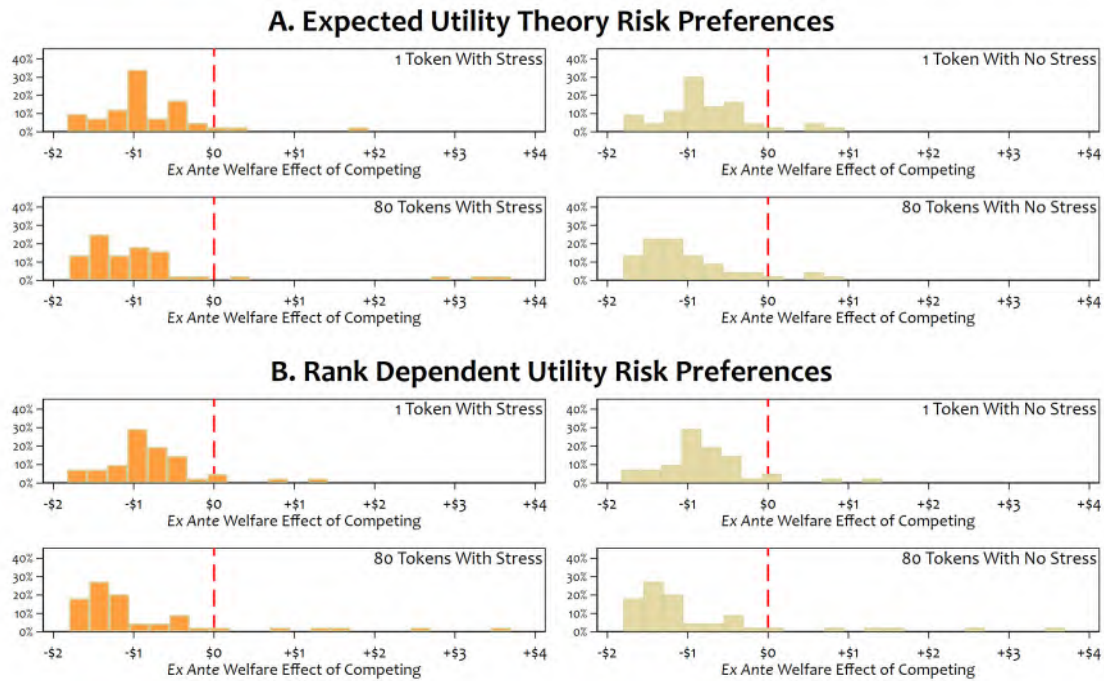


Figure 15: Ex Post Welfare Effects of the Observed Decision to Compete (or Not)

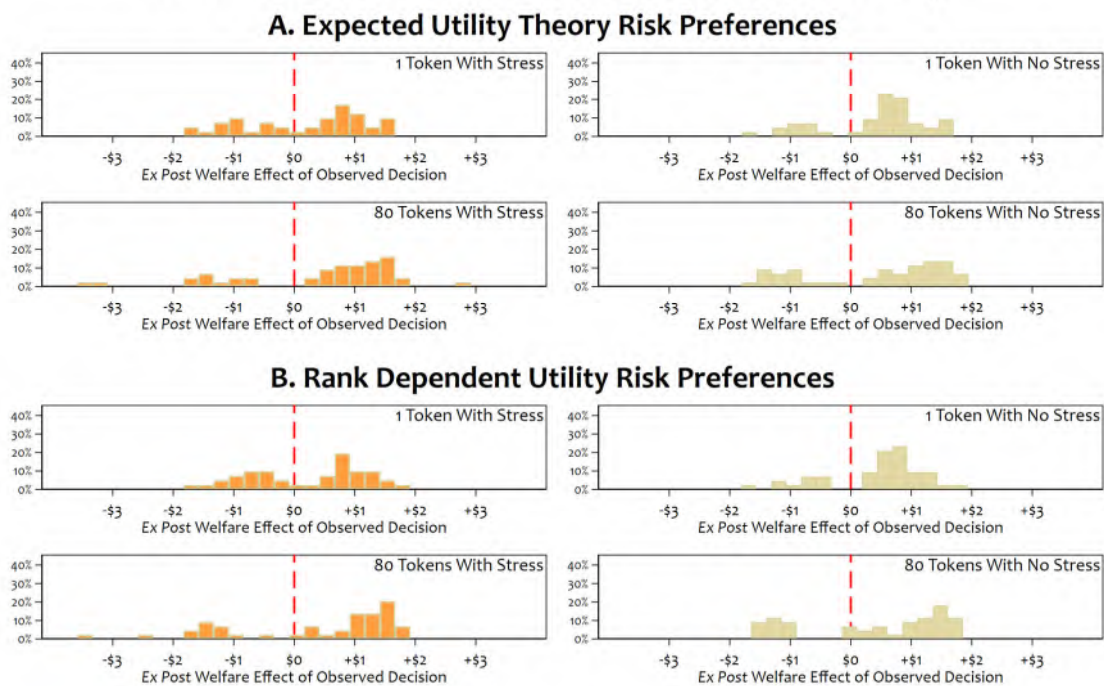
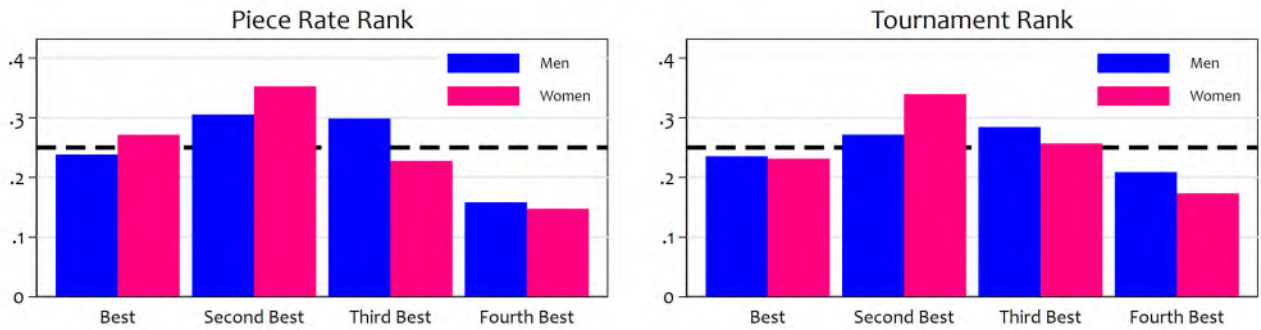


Figure 16: Recovered Beliefs on the Probability of Personal Performance Ranks by Men and Women

A. One Token



B. Eighty Tokens

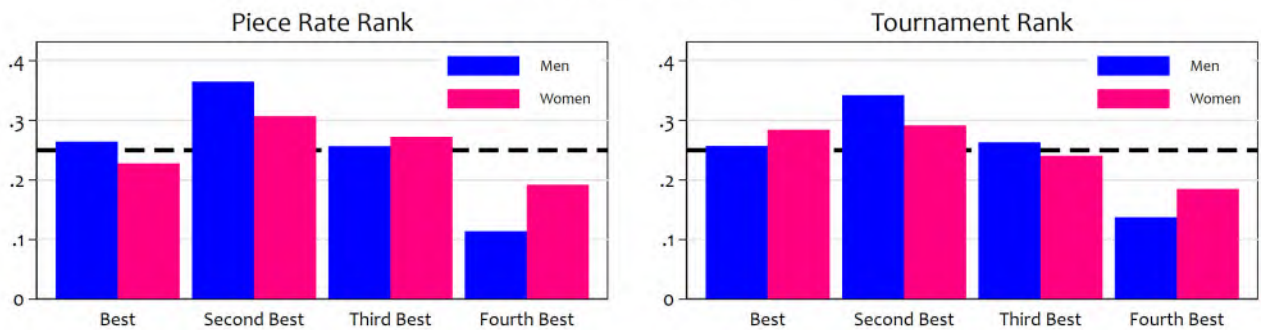


Figure 17: Bias and Confidence in the Literacy of Men and Women on the Inflation Question

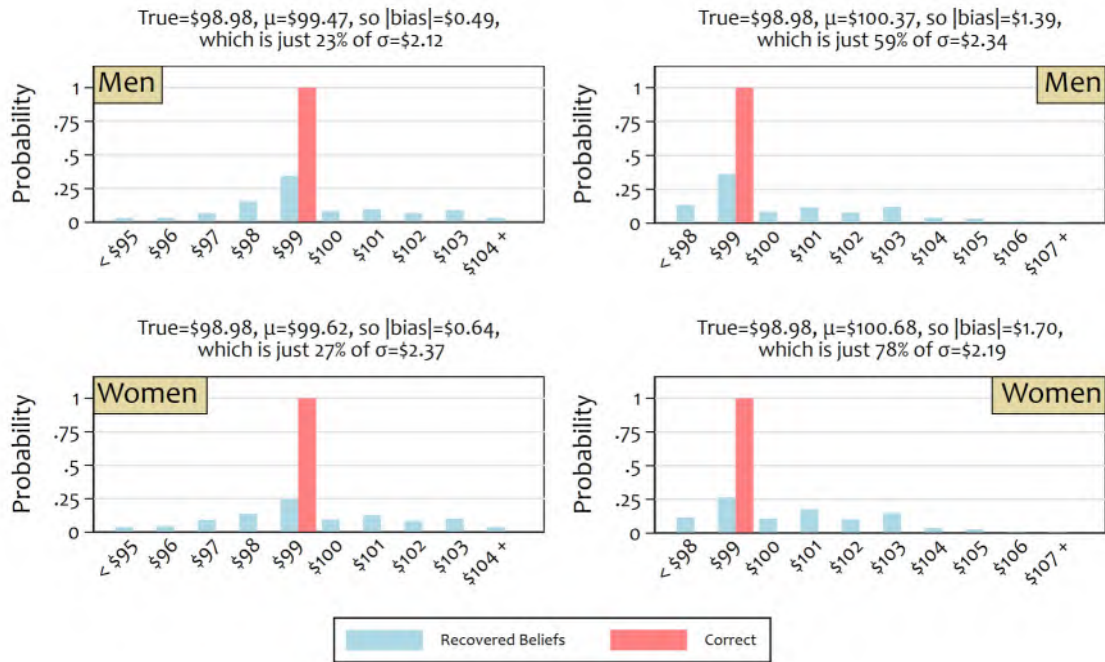
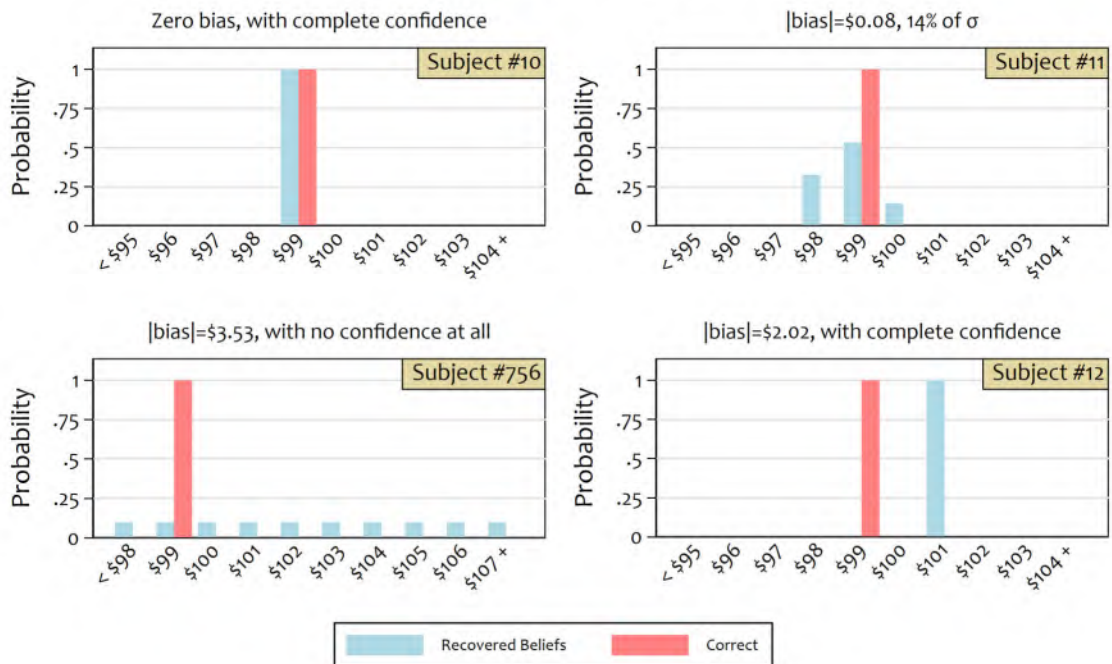


Figure 18: Literacy and Confidence: The Good, the Bad, and the Ugly



References

- Alan, Sule, and Ertac, Seda, “Mitigating the Gender Gap in the Willingness to Compete: Evidence from a Randomized Field Experiment,” *Journal of the European Economic Association*, 17(4), 2019, 1147–1185.
- Allen, Franklin, “Discovering Personal Probabilities When Utility Functions Are Unknown,” *Management Science*, 33(4), 1987, 452-454.
- Alpert, Marc, and Raiffa, Howard, “A Progress Report on the Training of Probability Assessors,” in D. Kahneman, P. Slovic & A. Tversky (eds.) *Judgment under Uncertainty: Heuristics and Biases* (New York: Cambridge University Press, 1982).
- Andersen, Steffen; Fountain, John; Harrison, Glenn W., and Rutström, E. Elisabet, “Estimating Subjective Probabilities,” *Journal of Risk & Uncertainty*, 48, 2014a, 207-229.
- Andersen, Steffen; Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, “Discounting Behavior: A Reconsideration,” *European Economic Review*, 71, 2014b, 15–33.
- Arthur, Winfred Jr., and Day, David V., “Development of a Short Form for the Raven Advanced Progressive Matrices Test,” *Educational and Psychological Measurement*, 54(2), Summer 1994, 394-403.
- Balafoutas, Loukas and Sutter, Matthias, “Affirmative Action Policies Promote Women and Do Not Harm Efficiency in the Laboratory,” *Science*, 335(6068), February 3, 2012, 579-582.
- Ben-Simon, Anat; Budescu, David V., and Nevo, Baruch, “A Comparative Study of Measures of Partial Knowledge in Multiple-Choice Tests,” *Applied Psychological Measurement*, 21(1), 1977, 65-88.
- Bereby-Meyer, Yoella; Meyer, Joachim, and Budescu, David V., “Decision Making Under Internal Uncertainty: the Case of Multiple-Choice Tests with Different Scoring Rules,” *Acta Psychologica*, 112, 2003, 207-220.
- Berg, Joyce E.; Daley, Lane A.; Dickhaut, John W.; and O’Brien, John R., “Controlling Preferences for Lotteries on Units of Experimental Exchange,” *Quarterly Journal of Economics*, 101, May 1986, 281-306.
- Bernardo, José, M. and Smith, Adrian F.M., *Bayesian Theory* (Chichester, UK: Wiley, 2000).
- Bickel, J. Eric, “Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules,” *Decision Analysis*, 4(2), 2007, 49-65.
- Bickel, J. Eric, “Scoring Rules and Decision Analysis Education,” *Decision Analysis*, 7(4), 2010, 346-357.
- Bickerton, Derek, *More Than Nature Needs: Language, Mind and Evolution* (Cambridge, MA: Harvard University Press, 2014).
- Borghans, Lex; Duckworth, Angela Lee; Heckman, James J., and ter Weel, Bas, “The Economics and Psychology of Personality Traits,” *Journal of Human Resources*, 46(4), January 2008, 972-1059.

- Borghans, Lex; Golsteyn, Bart H.H.; Heckman, James J., and Meijers, Huub, "Gender Differences in Risk Aversion and Ambiguity Aversion," *Journal of the European Economic Association*, 7(2-3), 2009, 649-658.
- Borghans, Lex; Meijers, Huub, and ter Weel, Bas, "The Role of Noncognitive Skills in Explaining Cognitive Test Scores," *Economic Inquiry*, 46(1), January 2008, 2-12.
- Borghans, Lex; ter Weel, Bas, and Weinberg, Bruce A., "Interpersonal Styles and Labor Market Outcomes," *Journal of Human Resources*, 43(4), 2008, 815-858.
- Bruner, Jerome, "Processes of Cognitive Growth: Infancy," *Clark University Press Heinz Werner Lectures*, January 1968; available at <https://commons.clarku.edu/heinz-werner-lectures/20>.
- Bucher-Koenen, Tabea; Alessie, Rob J.; Lusardi, Annamaria, and van Rooij, Maarten, "Fearless Woman: Financial Literacy and Stock Market Participation," *NBER Working Paper 28723*, National Bureau of Economics, April 2021.
- Bucher-Koenen, Tabea; Lusardi, Annamaria; Alessie, Rob J., and van Rooij, Maarten, "How Financially Literate are Women? An Overview and New Insights," *Journal of Consumer Affairs*, 51(2), 2017, 255-283.
- Budescu, David V. and Johnson, Timothy R., "A Model-Based Approach for the Analysis of the Calibration of Probability Judgments," *Judgment and Decision Making*, 6(8), 2011, 857-869.
- Camerer, Colin F. and Hogarth, Robin M., "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk & Uncertainty*, 19, 1999, 7-42.
- Carpenter, Patricia A.; Just, Marcel, and Shell, Peter, "What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test," *Psychological Review*, 97(3), 1990, 404-431.
- Carroll, John B., *Human Cognitive Abilities: A Survey of Factor-Analytic Studies* (New York: Cambridge University Press, 1993).
- Carroll, John B., "The Three-Stratum Theory of Cognitive Abilities," in D. P. Flanagan and L. Harrison (eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (New York: Guilford Press, 1997).
- Cawley, John; Conneely, Karen; Heckman, James, and Vytlačil, Edward, "Cognitive Ability, Wages, and Meritocracy," in B. Devlin, S.E. Fienberg, D.P. Resnick, K. Roeder (eds.), *Intelligence, Genes, and Success* (New York: Springer, 1997).
- Cawley, John; Heckman, James, and Vytlačil, Edward, "Three Observations on Wages and Measured Cognitive Ability," *Labour Economics*, 8(4), 2001, 419-442.
- Chater, Nick, *The Mind is Flat* (New York: Penguin, 2018).
- Chen, Yuanyuan; Feng, Shuaizhang; Heckman, James J., and Kautz, Tim, "Sensitivity of Self-reported Noncognitive Skills to Survey Administration Conditions," *Proceedings of the National Academy of*

- Sciences*, 117(2), 2020, 931–935.
- Civelli, Andrea, and Deck, Cary, “A Flexible and Customizable Method for Assessing Cognitive Abilities,” *Review of Behavioral Economics*, 5, 2018, 123-147.
- Clark, Andy, *Being There: Putting Brain, Body, and World Together Again* (Cambridge, MA: MIT Press, 1997).
- Clark, Andy, *Natural-Born Cyborgs* (New York: Oxford University Press, 2003).
- Clark, Andy, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (New York: Oxford University Press, 2011).
- Court, J. H., “Sex Differences in Performance on Raven’s Progressive Matrices: A Review,” *Alberta Journal of Educational Research*, 29, 1983, 54-74.
- Cunho, Flavio, and Heckman, James J., “The Technology of Skill Formation,” *American Economic Review (Papers & Proceedings)*, 97(2), 2007, 31-47.
- Cunho, Flavio, and Heckman, James J., “Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Journal of Human Resources*, 43(4), 2008, 738-782.
- Cunho, Flavio; Heckman, James J., and Schennach, Susanne M., “Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Econometrica*, 78(3), 2010, 883-931.
- Danz, David; Vesterlund, Lise, and Wilson, Alistair J., “Belief Elicitation and Behavioral Incentive Compatibility,” *American Economic Review*, 112(9), 2022, 2851-2883.
- Daston, Lorraine, “The Naturalized Female Intellect,” *Science in Context*, 5(2), 1992, 209-235.
- Dennett, Daniel, *Kinds of Minds* (New York: Basic Books, 1996).
- Dennett, Daniel, *Consciousness Explained* (New York: Little, Brown & Company, 1991).
- Dennett, Daniel, *From Bacteria to Bach and Back* (New York: Norton, 2017).
- de Finetti, Bruno, “Methods for Discriminating Levels of Partial Knowledge Concerning a Test Item,” *British Journal of Mathematical and Statistical Psychology*, 18, May 1965, 87-123.
- Di Girolamo, Amalia; Harrison, Glenn W.; Lau, Morten I., and Swarthout, J. Todd, “Subjective Belief Distributions and the Characterization of Economic Literacy,” *Journal of Behavioral and Experimental Economics*, 59, 2015, 1-12.
- Dreyfus, Hubert L., *Alchemy and Artificial Intelligence* (Santa Monica, CA: RAND Corporation, 1965, Document #P-3244); available at <https://www.rand.org/pubs/papers/P3244.html>.
- Dreyfus, Hubert L., *What Computers Can’t Do: The Limits of Artificial Intelligence* (New York: Harper & Row, 1972).
- Evans, Dylan, *Risk Intelligence: How to Live with Uncertainty* (New York: The Free Press, 2012).

- Flynn, James R., *What is Intelligence? Beyond the Flynn Effect* (New York: Cambridge University Press, 2007).
- Gao, Xiaoxue Sherry; Harrison, Glenn W., and Tchernis, Rusty, “Behavioral Welfare Economics and Risk Preferences: A Bayesian Approach,” *Experimental Economics*, 26, 2023, 273-303.
- Gigerenzer, Gerd, and Hoffrage, Ulrich, “How to Improve Bayesian Reasoning Without Instruction: Frequency Formats,” *Psychological Review*, 102(4), 1995, 684-704.
- Gignac, Gilles E., “A Moderate Financial Incentive Can Increase Effort, But Not Intelligence Test Performance in Adult Volunteers,” *British Journal of Psychology*, 109(3), 2018, 500-516.
- Gneezy Uri; Niederle Muriel, and Rustichini Aldo, “Performance in Competitive Environments: Gender Differences,” *Quarterly Journal of Economics*, 118, 2003, 1049–1074.
- Gudykunst, William B., and Nishida, Tsukasa, “Attributional Confidence in Low- and High-Context Cultures,” *Human Communication Research*, 12, 1986, 525-549.
- Hanushek, Eric A., and Woessmann, Ludger, “The Role of Cognitive Skills in Economic Development,” *Journal of Economic Literature*, 46(3), 2008, 607-668.
- Hardcastle, Joseph; Hermann-Abell, Cari F., and DeBoer, George E., “Comparing Student Performance on Paper-and-Pencil and Computer-Based Tests,” *Working Paper*, Presented at the American Educational Research Association Annual Meeting, April 2018; available at <https://files.eric.ed.gov/fulltext/ED574099.pdf>.
- Harrison, Glenn W.; Hofmeyr, Andre; Kincaid, Harold; Monroe, Brian; Ross, Don; Schneider, Mark, and Swarthout, J. Todd, “Subjective Beliefs and Economic Preferences During the COVID-19 Pandemic,” *Experimental Economics*, 25, 2022a, 795-823.
- Harrison, Glenn W.; Martínez-Correa, Jimmy, and Swarthout, J. Todd, “Eliciting Subjective Probabilities with Binary Lotteries,” *Journal of Economic Behavior and Organization*, 101, May 2014a, 128-140.
- Harrison, Glenn W.; Martínez-Correa, Jimmy; Swarthout, J. Todd, and Ulm, Eric R., “Eliciting Subjective Probability Distributions with Binary Lotteries,” *Economics Letters*, 127, 2015, 68-71.
- Harrison, Glenn W.; Martínez-Correa, Jimmy; Swarthout, J. Todd, and Ulm, Eric “Scoring Rules for Subjective Probability Distributions,” *Journal of Economic Behavior & Organization*, 134, 2017, 430-448.
- Harrison, Glenn W.; Monroe, Brian, and Ulm, Eric, “Recovering Subjective Probability Distributions: A Bayesian Approach,” *CEAR Working Paper 2022-03*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2022b.
- Harrison, Glenn W.; Morsink, Karlijn, and Schneider, Mark, “Literacy and the Quality of Index Insurance Decisions,” *Geneva Risk & Insurance Review*, 2022c, 47, 66-97.
- Harrison, Glenn W., and Phillips, Richard D., “Subjective Beliefs and the Statistical Forecasts of Financial Risks: the Chief Risk Officer Project,” in T.J. Andersen (ed.) *Contemporary Challenges in*

- Risk Management* (New York, Palgrave Macmillan, 2014b).
- Harrison, Glenn W., and Ross, Don, “Behavioral Welfare Economics and the Quantitative Intentional Stance,” in G.W. Harrison and D. Ross (eds.), *Models of Risk Preferences: Descriptive and Normative Challenges* (Bingley, UK: Emerald, Research in Experimental Economics, 2023).
- Harrison, Glenn W., and Swarthout, J. Todd, “Belief Distributions, Bayes Rule and Bayesian Overconfidence,” *CEAR Working Paper 2020-11*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2022.
- Harrison, Glenn W., and Swarthout, J. Todd, “Cumulative Prospect Theory in the Laboratory: A Reconsideration,” in G.W. Harrison and D. Ross (eds.), *Models of Risk Preferences: Descriptive and Normative Challenges* (Bingley, UK: Emerald, Research in Experimental Economics, 2023).
- Heckman, James J., “Lessons from the Bell Curve,” *Journal of Political Economy*, 103(5), 1995, 1091-1120.
- Heckman, James J., “Schools, Skills, and Synapses,” *Economic Inquiry*, 46(3), 2008, 289-324.
- Heckman, James J., *Giving Kids a Fair Chance* (Cambridge, MA: MIT Press, 2013).
- Heckman, James J.; Heinrich, Carolyn, and Smith, Jeffrey, “The Performance of Performance Standards,” *Journal of Human Resources*, 37(4), 2002, 778-811.
- Heckman, James J.; Humphries, John Eric, and Kautz, Tim (eds.), *The Myth of Achievement Tests: The GED and the Role of Character in American Life* (Chicago: University of Chicago Press, 2014).
- Heckman, James J., and Kautz, Tim, “Hard Evidence on Soft Skills,” *Labour Economics*, 19, 2012, 451-464.
- Heckman, James J., and Kautz, Tim, “Fostering and Measuring Skills: Interventions that Improve Character and Cognition,” in Heckman, J.J.; Humphries, J.E., and Kautz, T. (eds.), *The Myth of Achievement Tests: The GED and the Role of Character in American Life* (Chicago: University of Chicago Press, 2014).
- Heckman, James; Pinto, Rodrigo and Savelyev, Peter, “Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes,” *American Economic Review*, 103(6), 2013, 2052–2086.
- Hossain, Tanjim and Okui, Ryo, “The Binarized Scoring Rule,” *Review of Economic Studies*, 2013, 80, 984-991.
- Hutchins, Edwin, *Cognition in the Wild* (Cambridge, MA: MIT Press, 1995).
- John, Oliver P., and Srivastava, Sanjay, “The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives,” In L.A. Pervin and O.P. John (eds.), *Handbook of Personality: Theory and Research* (New York: Guilford Press, Volume 2, 1999).
- Kadane, Joseph B. and Fischhoff, Baruch, “A Cautionary Note on Global Recalibration,” *Judgment and Decision Making*, 8(1), 2013, 25–27.

- Kansup, Wanlop, and Hakistan, A. Ralph, "A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-Choice Tests: I. Scoring Procedures," *Journal of Educational Measurement*, 12, 1975, 219-230.
- Karay, Yassin; Schaubert, Stefan K.; Stosch, Christoph, and Schüttpeitz-Brauns, Katrin, "Computer Versus Paper – Does It Make Any Difference in Test Performance?" *Teaching and Learning in Medicine*, 27(1), 2015, 57-62.
- Kirsh, David, and Maglio, Paul, "On Distinguishing Epistemic From Pragmatic Action," *Cognitive Science*, 18(4), 1994, 513-549.
- Klibanoff, Peter; Marinacci, Massimo, and Mukerji, Sujoy, "A Smooth Model of Decision Making Under Ambiguity," *Econometrica*, 73(6), November 2005, 1849-1892.
- Koehler, Roger A., "A Comparison of the Validities of Conventional Choice Testing and Various Confidence Marking Procedures," *Journal of Educational Measurement*, 8, 1971, 297-303.
- Koehler, Roger A., "Overconfidence on Probabilistic Tests," *Journal of Educational Measurement*, 11, 1974, 101-108.
- Levitt, Steven D.; List, John A.; Neckermann, Susanne, and Sadoff, Sally, "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," *American Economic Journal: Economic Policy*, 8(4), 2016, 183-219.
- Lichtenstein, Sarah; Fischhoff, Baruch, and Phillips, Lawrence D., "Calibration of Probabilities: The State of the Art to 1980," in D. Kahneman, P. Slovic & A. Tversky (eds.) *Judgment under Uncertainty: Heuristics and Biases* (New York: Cambridge University Press, 1982).
- Lynn, Richard, and Irwing, Paul, "Sex Differences On the Progressive Matrices: A Meta-analysis," *Intelligence*, 32, 2004, 481-498.
- Lusardi, Annamaria, and Mitchell, Olivia S., "Planning and Financial Literacy: How Do Women Fare?" *American Economic Review (Papers & Proceedings)*, 98(2), 2008, 413-417.
- Matzen, Laura E.; Benz, Zachary O.; Dixon, Kevin R.; Posey, Jamie; Kroger, James K., and Speed, Ann E., "Recreating Raven's: Software for Systematically Generating Large Numbers of Raven-Like Matrix Problems with Normed Properties," *Behavior Research Methods*, 42(2), 2010, 525-541.
- McDaniel, Tanga M., and Rutström, E. Elisabet, "Decision Making Costs and Problem Solving Performance," *Experimental Economics*, 4, 2001, 145-161.
- McKelvey, Richard D., and Page, Talbot, "Public and Private Information: An Experimental Study of Information Pooling," *Econometrica*, 58(6), November 1990, 1321-1339.
- Moore, Don A., and Healy, Paul J., "The Trouble With Overconfidence," *Psychological Review*, 115(2), 2008, 502-517.
- Murphy, Alan, "A Note on the Utility of Probabilistic Predictions and the Probability Score in the Cost-Loss Ratio Decision Situation," *Journal of Applied Meteorology*, 5, 1966, 534-537.

- Niederle, Muriel, "Gender," in J.Kagel and A.E. Roth (eds.), *Handbook of Experimental Economics: Volume 2* (Princeton: Princeton University Press, 2015).
- Niederle, Muriel; Segal, Carmit, and Vesterlund, Lise, "How Costly Is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness," *Management Science*, 59(1), 2013, 1-16.
- Niederle, Muriel, and Vesterlund, Lise, "Do Women Shy Away From Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics*, 122, 2007, 1067-1101.
- Niederle, Muriel, and Vesterlund, Lise, "Explaining the Gender Gap in Math Test Scores: The Role of Competition," *Journal of Economic Perspectives*, 24(2), 2010, 129-144.
- Pearl, Judea, "An Economic Basis for Certain Methods of Evaluating Probabilistic Forecasts," *International Journal of Man-Machine Studies*, 10, 1978, 175-183.
- Planer, Ronald J, and Sterelny, Kim, *From Signal to Symbol: The Evolution of Language* (Cambridge, MA: MIT Press 2021).
- Prelec, Drazen, "The Probability Weighting Function," *Econometrica*, 66, 1998, 497-527.
- Raven, J.; Raven, J. C., and Court, J. H., *Coloured Progressive Matrices*. London: HK Lewis, 1962.
- Raven, J.; Raven, J. C., and Court, J. H., *Raven Manual: Section 1. General Overview*. Oxford: Oxford Psychologist Press, 1993.
- Raven, J.; Raven, J. C., and Court, J. H., *Raven Manual: Section 4. Advanced Progressive Matrices*. Oxford: Oxford Psychologist Press, 1998.
- Raven, J.; Raven, J. C., and Court, J. H., *Raven Manual: 3. Standard Progressive Matrices*. Oxford: Oxford Psychologist Press, 2000.
- Rippey, Robert M, "Probabilistic Testing," *Journal of Educational Measurement*, 5, 1968, 211-215.
- Rippey, Robert M, "A Comparison of Five Different Scoring Functions for Confidence Tests," *Journal of Educational Measurement*, 7, 1970, 165-170.
- Roth, Alvin E., and Malouf, Michael W. K., "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, 86, 1979, 574-594.
- Rutström, E. Elisabet, and Wilcox, Nathaniel T., "Stated Beliefs Versus Empirical Beliefs: A Methodological Inquiry and Experimental Test," *Games and Economic Behavior*, 67, 2009, 616-632.
- Savage, Leonard J., "Elicitation of Personal Probabilities and Expectations," *Journal of American Statistical Association*, 66, December 1971, 783-801.
- Savage, Leonard J., *The Foundations of Statistics* (New York: Dover Publications, 1972; Second Edition).
- Schlag, Karl H., and van der Weele, Joël, "Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk-Neutrality," *Theoretical Economics Letters*, 3, 2013, 38-42.

- Shapiro, Lawrence (ed.), *The Routledge Handbook of Embodied Cognition* (London, Routledge, 2014).
- Shuford, Emir H.; Arthur, Albert, and Massengill, H. Edward, "Admissible Probability Measurement Procedures," *Psychometrika*, 31, June 1966, 124-145.
- Shurchkov, Olga, "Under Pressure: Gender Differences In Output Quality And Quantity Under Competition And Time Constraints," *Journal of the European Economic Association*, 10(5), 2012, 1189-1213.
- Segal, Carmit, "Working When No One Is Watching: Motivation, Test Scores, and Economic Success," *Management Science*, 58(8), 2012, 1438-1457.
- Seidenfeld, Teddy, "Calibration, Coherence and Scoring Rules," *Philosophy of Science*, 52, 1985, 274-294.
- Sloman, Steven, and Fernbach, Philip, *The Knowledge Illusion: Why We Never Think Alone* (New York: Riverhead, 2017).
- Smith, Cedric A.B., "Consistency in Statistical Inference and Decision," *Journal of the Royal Statistical Society*, 23, 1961, 1-25.
- Smith, Vernon L., "Microeconomic Systems as an Experimental Science," *American Economic Review*, 72(5), December 1982, 923-955.
- ter Weel, Bas, "The Noncognitive Determinants of Labor Market and Behavioral Outcomes," *Journal of Human Resources*, 43(4), 2008, 729-737.
- van Rooij, Maarten; Lusardi, Annamaria, and Alessie, Rob J., "Financial Literacy and Stock Market Participation," *Journal of Financial Economics*, 101(2), 2011, 449-472.
- Vygotskij, Lev S., *Thought and Language* (Cambridge, MA: MIT Press, 1962).
- Vygotskij, Lev S., "Thinking and Speech," in R.W. Rieber & A.S. Carton (eds.), *The Collected Works of L.S. Vygotsky, Volume 1: Problems of General Psychology* (New York: Plenum Press, 1987); original work published in 1934.
- Wainer, Howard; Bradlow, Eric T., and Wang, Xiaohui, *Testlet Response Theory and Its Applications* (New York: Cambridge University Press, 2007).
- Wang, Ke, and Su, Zhendong, "Automatic Generation of Raven's Progressive Matrices," *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015, 903-909.

Appendix A: Instructions (Online Working Paper)

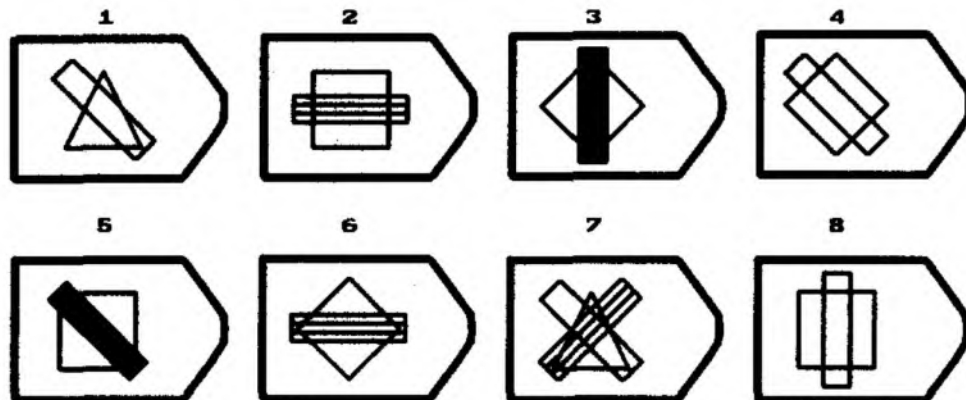
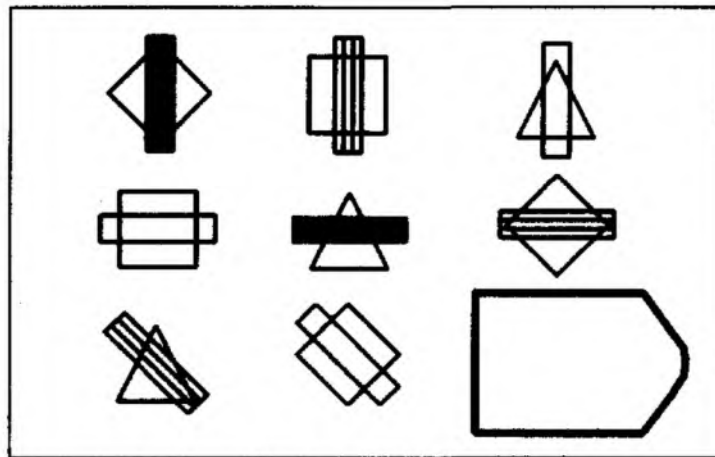
1. Non-Salient "Pen and Paper" Treatment

C

Your Instructions

This task is a test of perception and clear-thinking. You have already completed the first part of the task, in a previous session, when you were given 12 similar problems. We will now consider a fresh set of 36 problems.

Consider a similar problem, shown here.



The top part of this problem is a pattern with a bit cut out of it. Look at the pattern, and think what the piece needed to complete the pattern correctly both across and down must be like. Then find the right piece out of the eight bits shown, and numbered, in the bottom part of this problem.

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. In this case, if you were certain that piece #5 is the single best answer, you should write the number 5 on the answer sheet you have. You should write your answers in this part of the answer sheet:

Set II							
1			13			25	
2			14			26	
3			15			27	
4			16			28	
5			17			29	
6			18			30	
7			19			31	
8			20			32	
9			21			33	
10			22			34	
11			23			35	
12			24			36	

So you write the answer for the problem marked 1 in the box to the right of the number 1, and so on for all 36 problems.

This illustrative problem is relatively easy, but some of the later problems will not be so easy, and it might be harder to identify the one correct answer. You would still benefit by ruling out certain pieces so that you are more likely to select the correct answer from the pieces you have not ruled out.

You will find that the problems soon get difficult. Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. The important thing is to notice how the problems develop and to learn the method of solving them. In each problem you look across each row, or down each column, and decide what the missing figure is like. Then look for the answer that is right both ways, across **and** down, from the options you have to select from.

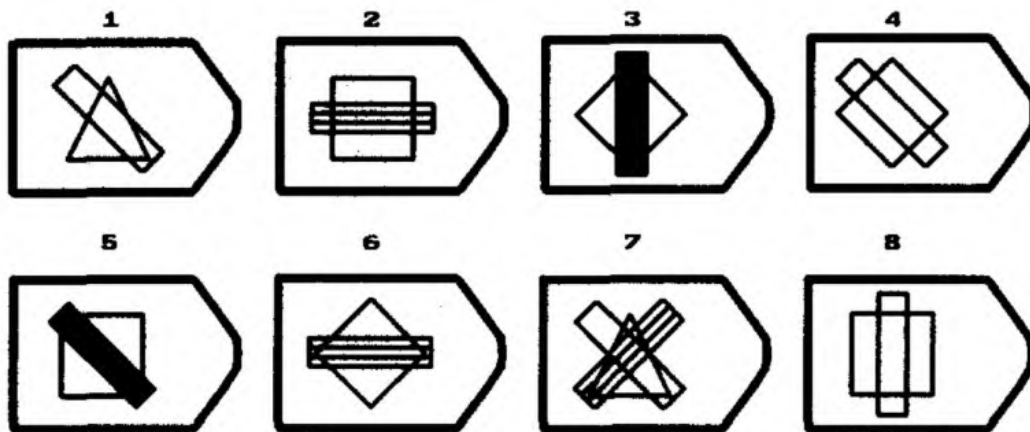
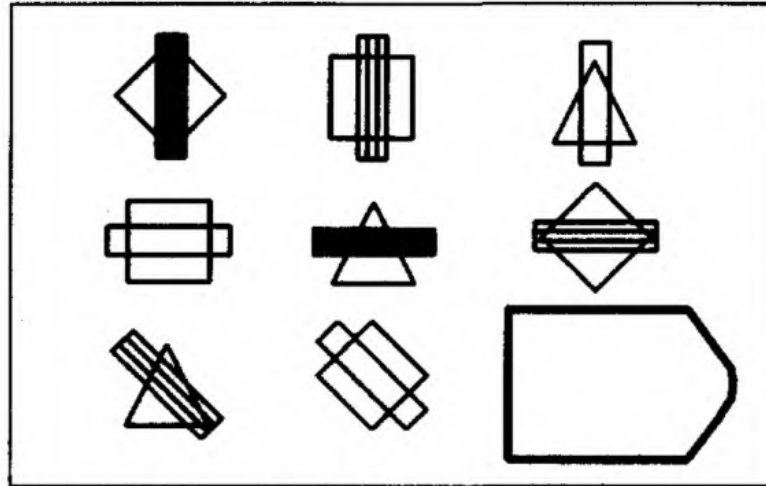
You can have as much time as you want, although we have to be out of the room in 90 minutes. Remember it is accurate work that counts. Attempt each problem in turn. Do your best to eliminate the incorrect pieces and find the correct piece to complete it before going on to the next problem.

You will receive \$5 for completing this task.

Your Instructions

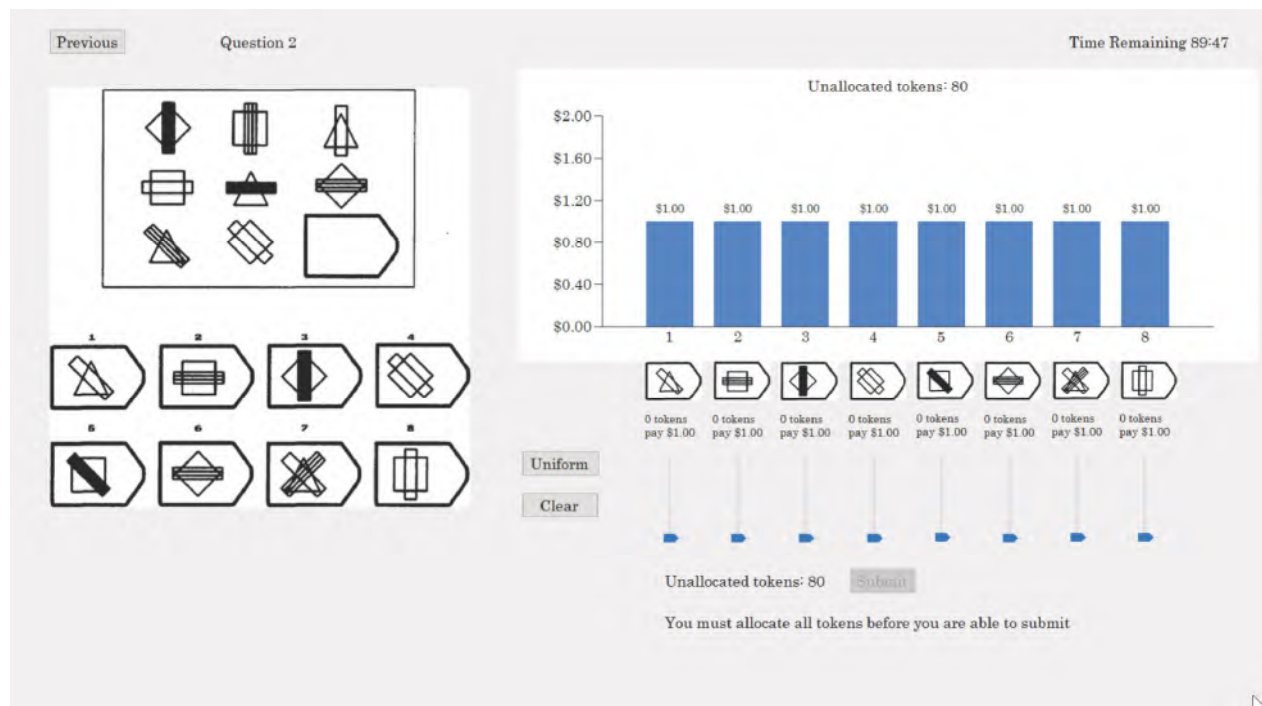
This task is a test of perception and clear-thinking. You have already completed the first part of the task, in a previous session, when you were given 12 similar problems. We will now consider a fresh set of 36 problems.

Consider a similar problem, shown here.

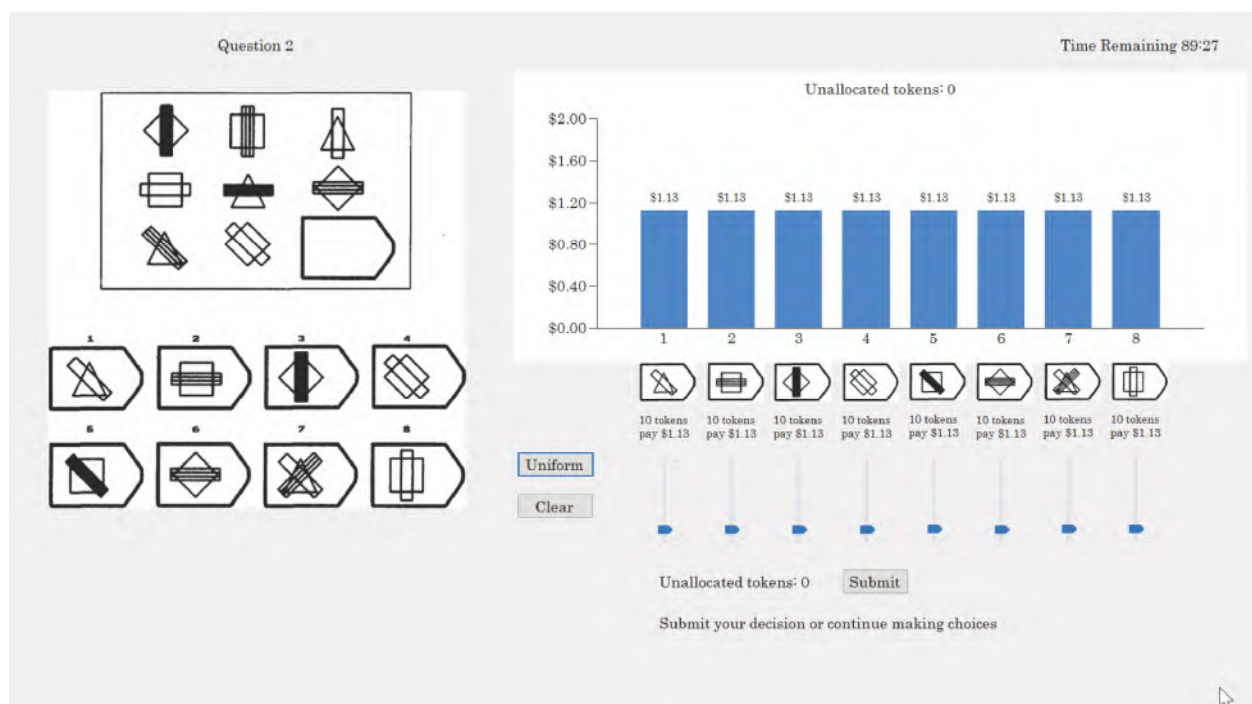


The top part of this problem is a pattern with a bit cut out of it. Look at the pattern, and think what the piece needed to complete the pattern correctly both across and down must be like. Then find the right piece out of the eight bits shown, and numbered, in the bottom part of this problem.

You will be asked to report your beliefs about the correct answer using an interface like this one, which is also generally familiar to you from a previous session.



The version you will see on the computer will be larger and easier to read. The problem and possible solutions are shown on the left of the screen, in the usual manner. On the right of the screen you have 80 tokens to allocate across the 8 possible answers. We start off with 0 tokens allocated to each of the possible answers. If you wanted to change this initial allocation so that there were 10 tokens allocated to each possible answer, just click on the **Uniform** button, and you will see this display:

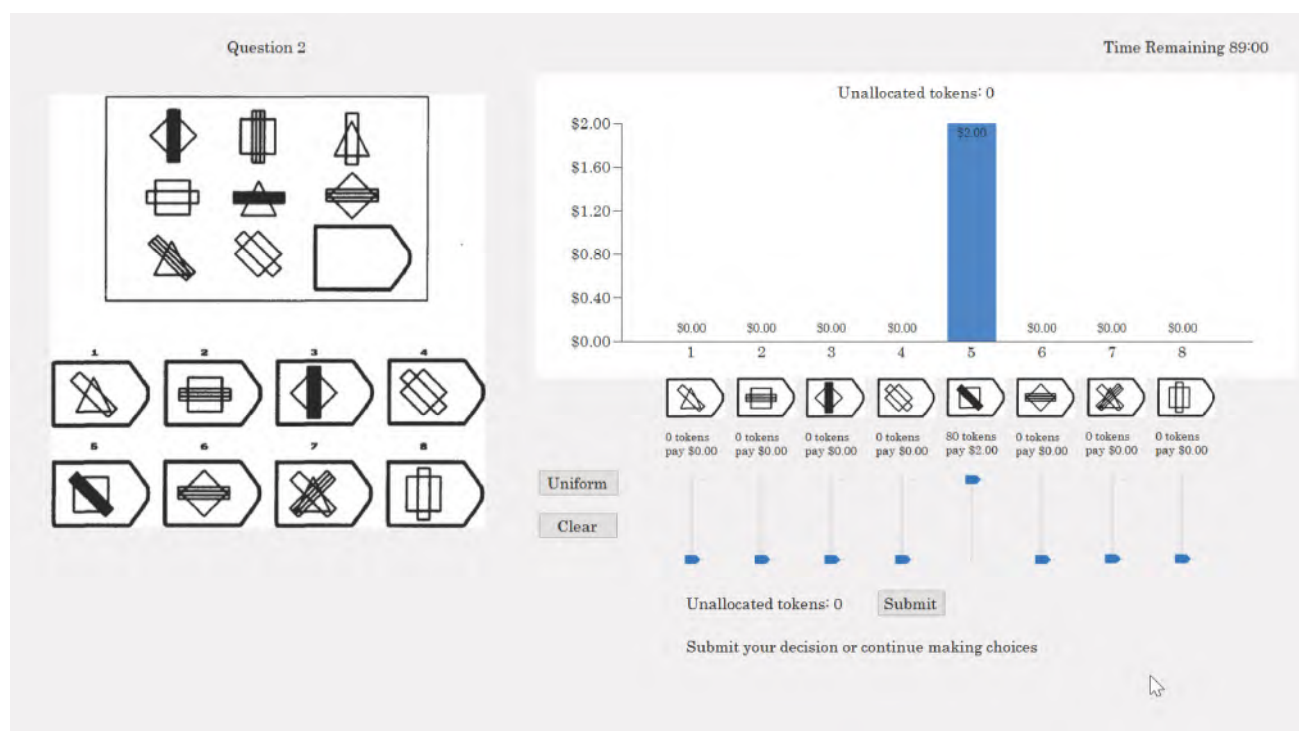


Here you have allocated 10 tokens to each possible answer, and you would earn \$1.13 if you reported this allocation of tokens, since only one of the 8 possible answers is correct. You can return to the initial allocation of 0 tokens for each possible answer by clicking on the **Clear** button.

As you allocate tokens, by moving the sliders up or down, the earnings will change above each bar. These are the earnings that you will receive for this problem if that bar refers to the correct answer to the problem. **You will be paid for all 36 problems, and each problem will pay between \$0 and \$2 depending on your answer.**

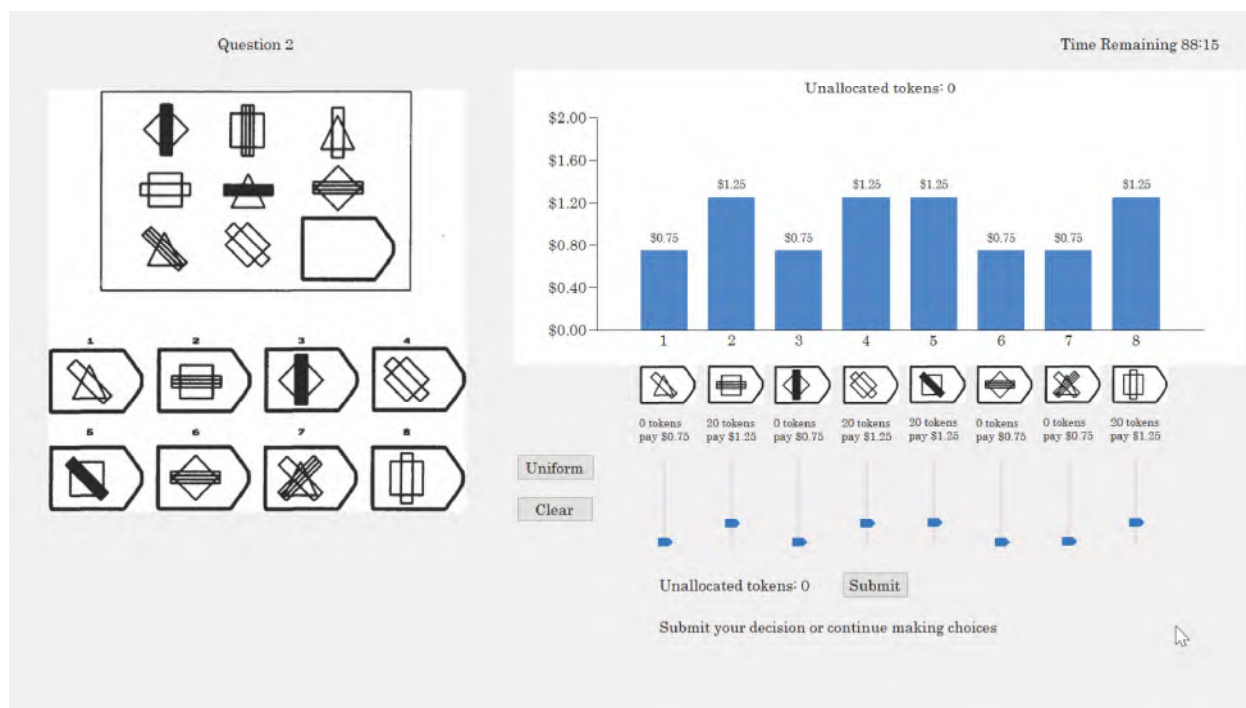
Return now to the problem itself. Only one of these pieces is perfectly correct. Pieces #2, #5 and #8 complete the problem correctly going down, in the sense that they have a square as the larger piece. Pieces #1, #4 and #5 are correct going across, in the sense that they have the right slope for the 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. In this case, if you were certain that piece #5 is the single best answer, you should allocate all 80 tokens to the bin representing piece #5 as in this display:



So in this case you would earn \$2.00 if indeed the correct answer was #5. Of course, if any of the other pieces turned out to be the correct answer you would, in this case, earn \$0.

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:



You can see that even if you eliminate some pieces, such as #1, #3, #6 and #7, that are clearly wrong, you give yourself a 1-in-4 chance of earning more money than if you guessed across all 8 pieces. In this case you would expect to earn \$1.25 if indeed one of pieces #2, #4, #5 or #8 had been correct. Recall that if you had allocated the tokens roughly equally across all 8 bars, thinking that any of the 8 pieces might be correct, you would only earn \$1.13 for this problem.

This illustrative problem is relatively easy, but some of the later problems will not be so easy, and it might be harder to identify the one correct answer. You would still benefit by ruling out certain pieces so that you can allocate more tokens to the ones that you have not ruled out.

You will find that the problems soon get difficult. Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. The important thing is to notice how the problems develop and to learn the method of solving them. In each problem you look across each row, or down each column, and decide what the missing figure is like. Then look for the answer that is right both ways, across **and** down, from the options you have to select from.

As you recall from allocating tokens in a similar task in a previous session, you can change your token allocation as many times as you like before you hit the **Submit** button and **Confirm** your token allocation. You can also review and modify your prior decisions. You will have navigation buttons in the top left corner of the screen. These navigation buttons are only available before you start moving the sliders for a problem. Once you move any sliders for a problem, the navigation buttons disappear and you must submit your answer and move to the next problem in order to see the navigation buttons again.

You can work at your own speed, although we have to be out of the room in 90 minutes. Your screen displays how many minutes (and seconds) are remaining in the top, right corner. Remember it is accurate work that counts. Attempt each problem in turn. Do your best to eliminate the incorrect pieces and find the correct piece to complete it before going on to the next problem. You will have the option to go back to previous problems if you want to. After you have completed the last question you must confirm if you are finished. When you are finished you will be told your total earnings, and will not have the option to go back to any of the problems. If you run out of time, you will automatically be treated as having finished, and told your total earnings.

Where you position each slider depends on your beliefs about the correct answer to the question. Again, each bar shows the amount of money you could earn if the true outcome corresponds to the possible solution shown under the bar.

You will be **paid for each of the 36 problems**, so you should think carefully about each problem. Since you can earn up to \$2 for each problem, you could earn up to \$72 over all 36 problems. You will not earn anything on any problems for which you have not confirmed an allocation of tokens.

It is up to you to balance the strength of your personal beliefs with the possibility of them being wrong. There are several important points for you to keep in mind when making your decisions:

- First, your belief about the correct answer to each problem is your personal judgment.
- Second, you have up to 90 minutes to complete this task.
- Third, you will not earn anything on any problems for which you have not confirmed an allocation of tokens.
- Fourth, depending on your choices and the correct answer you can earn up to \$2 for each problem, or up to \$72 over all 36 problems.
- Finally, your choices might also depend on your willingness to take risks or to gamble.

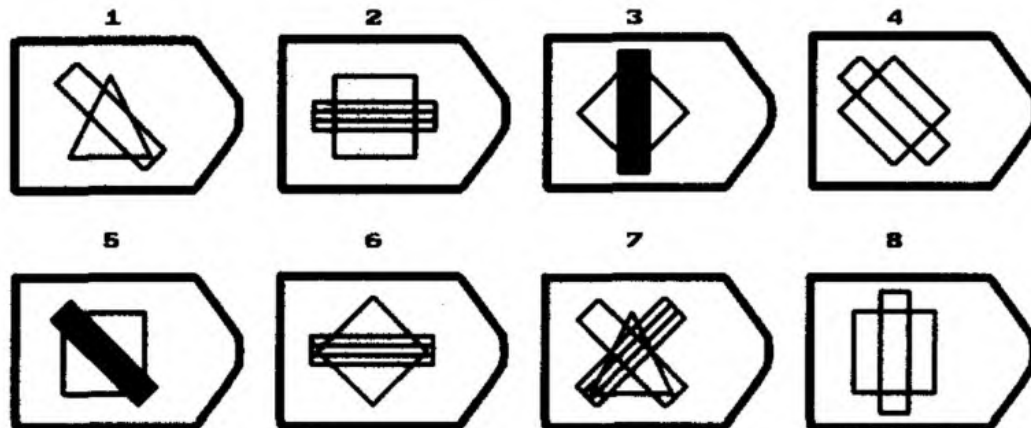
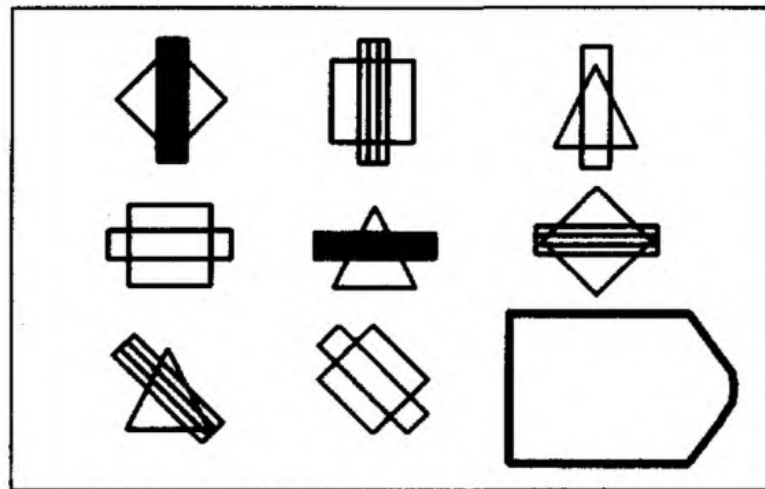
The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are finished you will see a summary of your responses and your total earnings from this task. Your earnings are in addition to the show-up payment you receive for participating.

Your Instructions

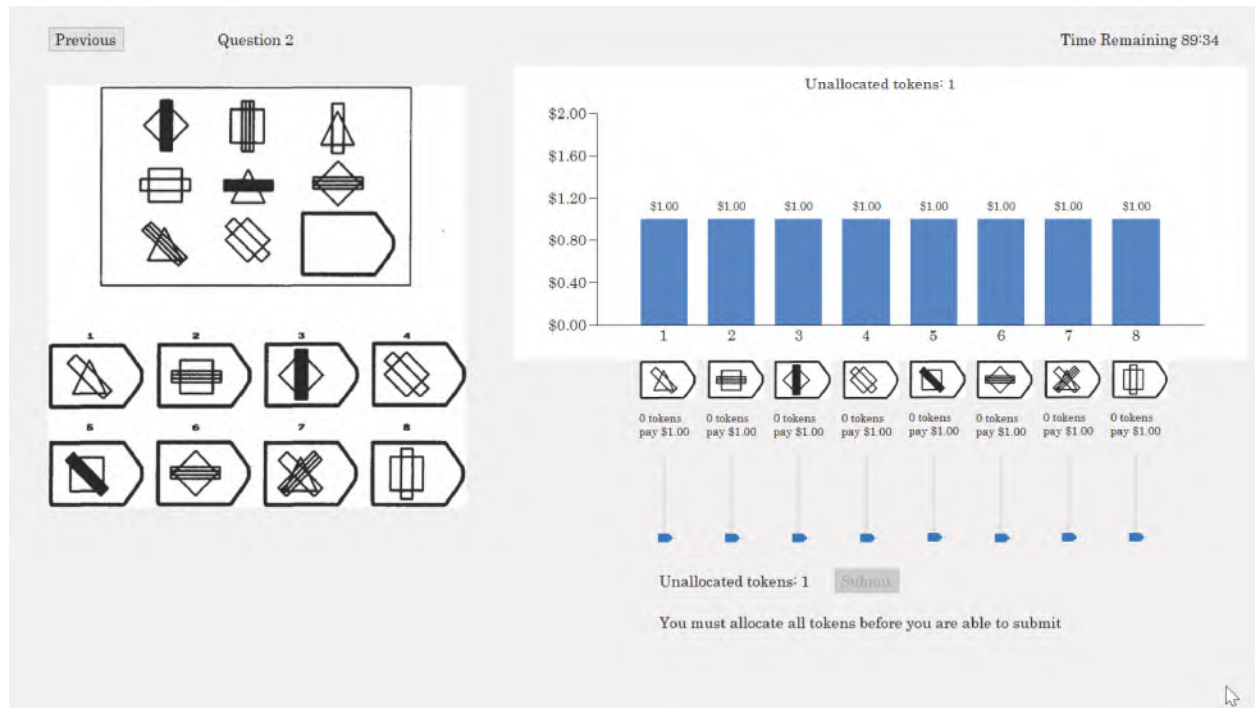
This task is a test of perception and clear-thinking. You have already completed the first part of the task, in a previous session, when you were given 12 similar problems. We will now consider a fresh set of 36 problems.

Consider a similar problem, shown here.



The top part of this problem is a pattern with a bit cut out of it. Look at the pattern, and think what the piece needed to complete the pattern correctly both across and down must be like. Then find the right piece out of the eight bits shown, and numbered, in the bottom part of this problem.

You will be asked to report your beliefs about the correct answer using an interface like this one, which is also generally familiar to you from a previous session. The version you will see on the computer will be larger and easier to read.

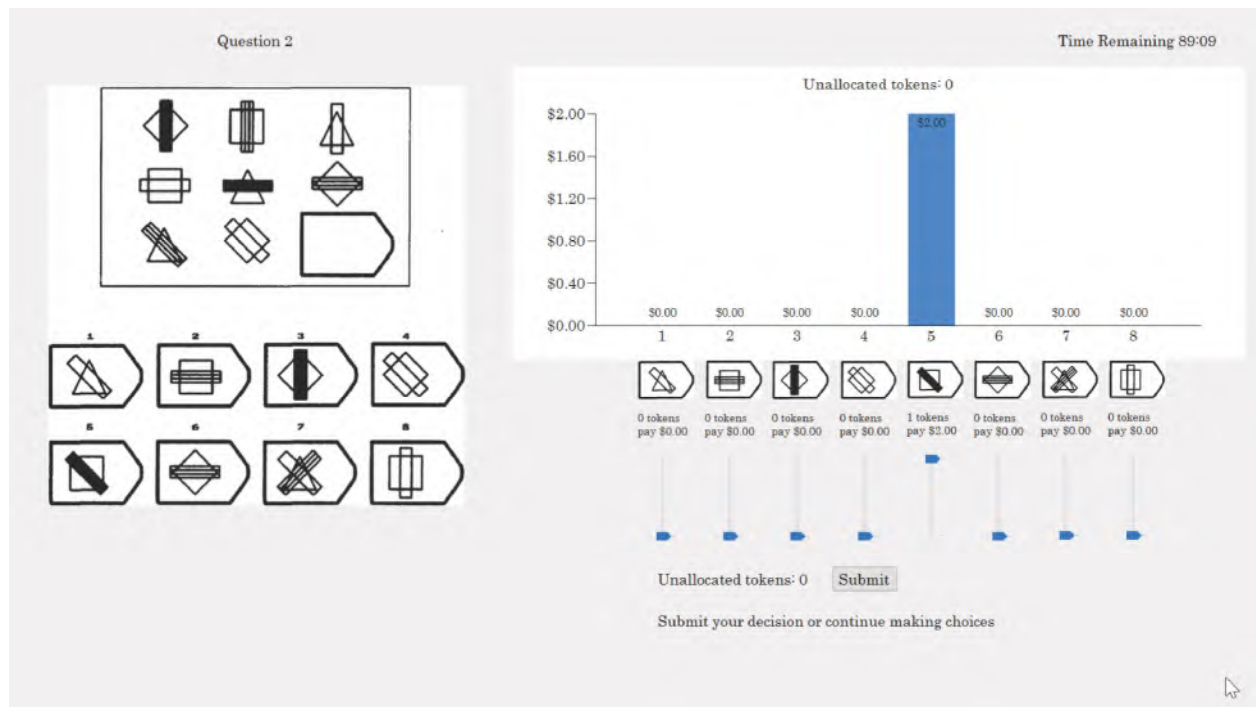


The problem and possible solutions are shown on the left of the screen, in the usual manner. On the right of the screen you have 1 token to allocate across the 8 possible answers. We start off with 0 tokens allocated to each of the possible answers.

As you allocate the token, by moving the sliders up or down, the earnings will change above each bar. These are the earnings that you will receive for this problem if that bar refers to the correct answer to the problem. **You will be paid for all 36 problems, and each problem will pay between \$0 and \$2 depending on your answer.**

Return now to the problem itself. 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. In this case, if you were certain that piece #5 is the single best answer, you should allocate the token to the bin representing piece #5 as in this display:



So in this case you would earn \$2.00 if indeed the correct answer was #5. Of course, if any of the other pieces turned out to be the correct answer you would, in this case, earn \$0.

This illustrative problem is relatively easy, but some of the later problems will not be so easy, and it might be harder to identify the one correct answer. You would still benefit by ruling out certain pieces so that you can allocate your token to one of the possible answers that you have not ruled out.

You will find that the problems soon get difficult. Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. The important thing is to notice how the problems develop and to learn the method of solving them. In each problem you look across each row, or down each column, and decide what the missing figure is like. Then look for the answer that is right both ways, across **and** down, from the options you have to select from.

As you recall from allocating tokens in a similar task in a previous session, you can change your token allocation as many times as you like before you hit the **Submit** button and **Confirm** your token allocation. You can also review and modify your prior decisions. You will have navigation buttons in the top left corner of the screen. These navigation buttons are only available before you start moving the sliders for a problem. Once you move any sliders for a problem, the navigation buttons disappear and you must submit your answer and move to the next problem in order to see the navigation buttons again.

You can work at your own speed, although we have to be out of the room in 90 minutes. Your screen displays how many minutes (and seconds) are remaining in the top, right corner. Remember it is accurate work that counts. Attempt each problem in turn. Do your best to eliminate the incorrect pieces and find the correct piece to complete it before going on to the next problem. You will have the option to go back to previous problems if you want to. After you have completed the last question you must confirm if you are finished. When you are finished you will be told your total earnings, and will not have the option to go back to any of the problems. If you run out of time, you will automatically be treated as having finished, and told your total earnings.

Where you allocate the token depends on your beliefs about the correct answer to the question. Again, each bar shows the amount of money you could earn if the true outcome corresponds to the possible solution shown under the bar.

You will be **paid for each of the 36 problems**, so you should think carefully about each problem. Since you can earn up to \$2 for each problem, you could earn up to \$72 over all 36 problems. You will not earn anything on any problems for which you have not confirmed an allocation of your token.

It is up to you to balance the strength of your personal beliefs with the possibility of them being wrong. There are several important points for you to keep in mind when making your decisions:

- First, your belief about the correct answer to each problem is your personal judgment.
- Second, you have up to 90 minutes to complete this task.
- Third, you will not earn anything on any problems for which you have not confirmed an allocation of your token.
- Fourth, depending on your choices and the correct answer you can earn up to \$2 for each problem, or up to \$72 over all 36 problems.
- Finally, your choices might also depend on your willingness to take risks or to gamble.

The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are finished you will see a summary of your responses and your total earnings from this task. Your earnings are in addition to the show-up payment you receive for participating.

4. *Scrambled Treatment*

The instructions for the Scrambled treatments were the same as their one-token and eight-token counterparts, with this text changed:

This illustrative problem is relatively easy, but some of the later problems will not be so easy, and it might be harder to identify the one correct answer. You would still benefit by ruling out certain pieces so that you can allocate more tokens to the ones that you have not ruled out.

You will find that the problems soon get difficult. Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. The important thing is to notice how the problems develop and to learn the method of solving them. In each problem you look across each row, or down each column, and decide what the missing figure is like. Then look for the answer that is right both ways, across **and** down, from the options you have to select from.

As you recall from allocating tokens in a similar task in a previous session, you can change your token allocation as many times as you like before you hit the **Submit** button and **Confirm** your token allocation. You can also review and modify your prior decisions. You will have navigation buttons in the top left corner of the screen. These navigation buttons are only available before you start moving the sliders for a problem. Once you move any sliders for a problem, the navigation buttons disappear and you must submit your answer and move to the next problem in order to see the navigation buttons again.

The new text completely replaces the second paragraph, deletes the word “later” from the first line of the first paragraph, and replaces “problems develop” with “problems can be solved” in the third paragraph:

This illustrative problem is relatively easy, but some of the problems will not be so easy, and it might be harder to identify the one correct answer. You would still benefit by ruling out certain pieces so that you can allocate more tokens to the ones that you have not ruled out.

The order in which you see the easier and harder problems, as well as the ones in between, will be selected at random for each person. So the problems do not generally become easier or harder as you work through them.

Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. The important thing is to notice how the problems can be solved, and to learn the method of solving them. In each problem you look across each row, or down each column, and decide what the missing figure is like. Then look for the answer that is right both ways, across **and** down, from the options you have to select from.

Appendix B: Subjective Beliefs and Risk Preferences (Online Working Paper)

In the main text all results use reported beliefs. In this appendix we present the logic behind the calculation of recovered beliefs that take into account estimated risk preferences of individual subjects. We only reproduce the displays for which the distinction between reports and recovered beliefs could matter. To ensure that the exposition here can be read alone, we repeat some text that is included in the main text.

There are three general approaches to eliciting beliefs, which we review in §5.D. The approach we employ is to directly pay subjects for their reports using financial incentives, and then recover beliefs if needed using estimates of the risk preferences of individuals. In order to infer subjective beliefs we therefore need two experimental tasks: one in which we elicit risk preferences defined over objective lotteries, and one in which we elicit subjective beliefs using the QSR defined over monetary payoffs. We ensure that the scale of payoffs in each task is comparable, to avoid extrapolation. We have each subject undertake both tasks to allow estimation of risk preferences, and recovery of subjective beliefs at the level of the individual.

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]. We estimate risk preferences for each subject using a Bayesian Hierarchical Model that allows us to have informative priors for inferences about the risk preferences of each individual, following Gao, Harrison and Tchernis [2023]. We estimate risk preferences for each individual assuming EUT or assuming RDU. Although RDU nests EUT, for some normative applications it is useful to have EUT estimates. A major advantage of the Bayesian approach to estimating risk preferences is that each of the simulated values of the posterior distribution of parameters can be *directly* used to infer subjective beliefs as a matter of the theory developed in Harrison, Monroe and Ulm [2022].

It is a straightforward matter to evaluate the expected welfare cost to a respondent of being forced to report their subjective beliefs in a way that requires that they only report their modal belief. In effect, this is the welfare cost to respondents of forcing them to use our **One Token** instrument instead of the **Eighty Tokens** 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. 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.

In the main text we used the RDU risk preferences of the subject to calculate the welfare cost, and reported them in Figure 13. Some find EUT a more attractive normative metric for such calculations, so we report in Figure B13 the same calculations when one uses EUT risk preferences for the subject. The qualitative pattern is virtually the same as if one used RDU risk preferences (this is not, in general, the case).

In the main text we used recovered beliefs using the RDU risk preferences of the subjects in Figure 16, for the belief of winning the competitions. Figure B16 show the same results for these subjects, but using their reported beliefs. We draw the same conclusions, apart from women being more optimistic of winning the tournament in the 80-token case; they were already modestly optimistic in terms of their recovered beliefs.

In the main text we used recovered beliefs using the RDU risk preferences of the subjects in Figures 17 and 18, for the core financial literacy questions. Figures B17 and B18 show the same results for these subjects, but using their reported beliefs. We draw the same conclusions.

Figure B13: Expected Welfare Cost to Respondent of Being Required to Only Report Modal Belief

Evaluated using EUT risk preferences of each subject
Using recovered beliefs from **Eighty Tokens** task

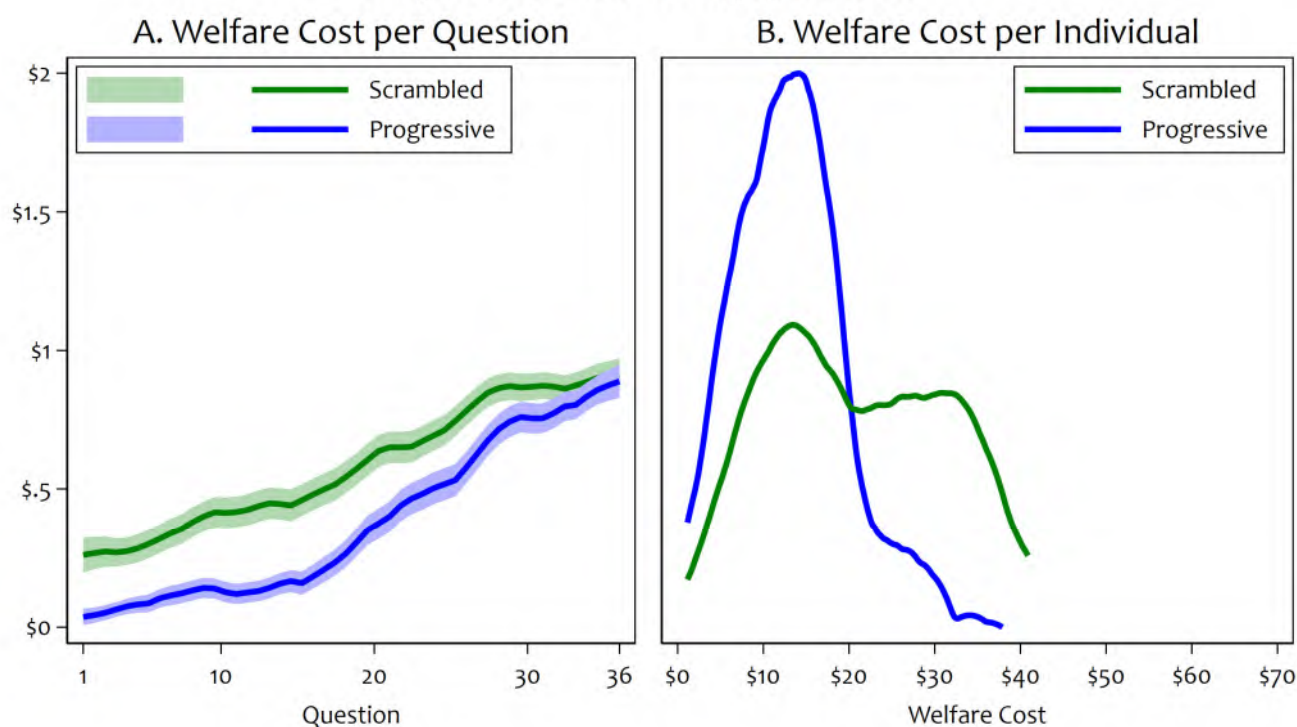
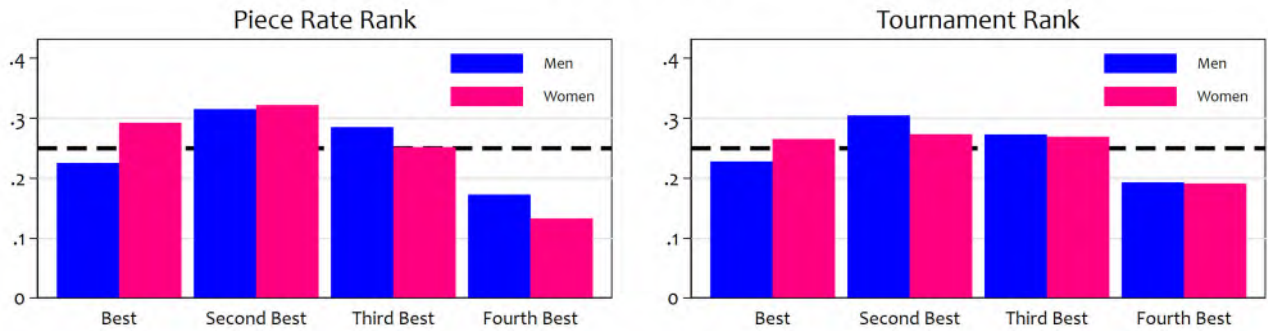


Figure B16: Reported Beliefs on the Probability of Personal Performance Ranks by Men and Women

A. One Token



B. Eighty Tokens

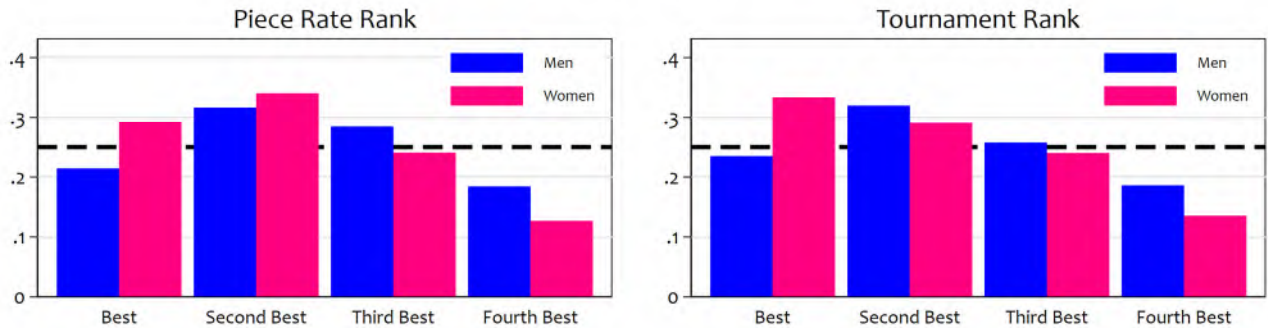


Figure B17: Bias and Confidence in the Literacy of Men and Women on the Inflation Question

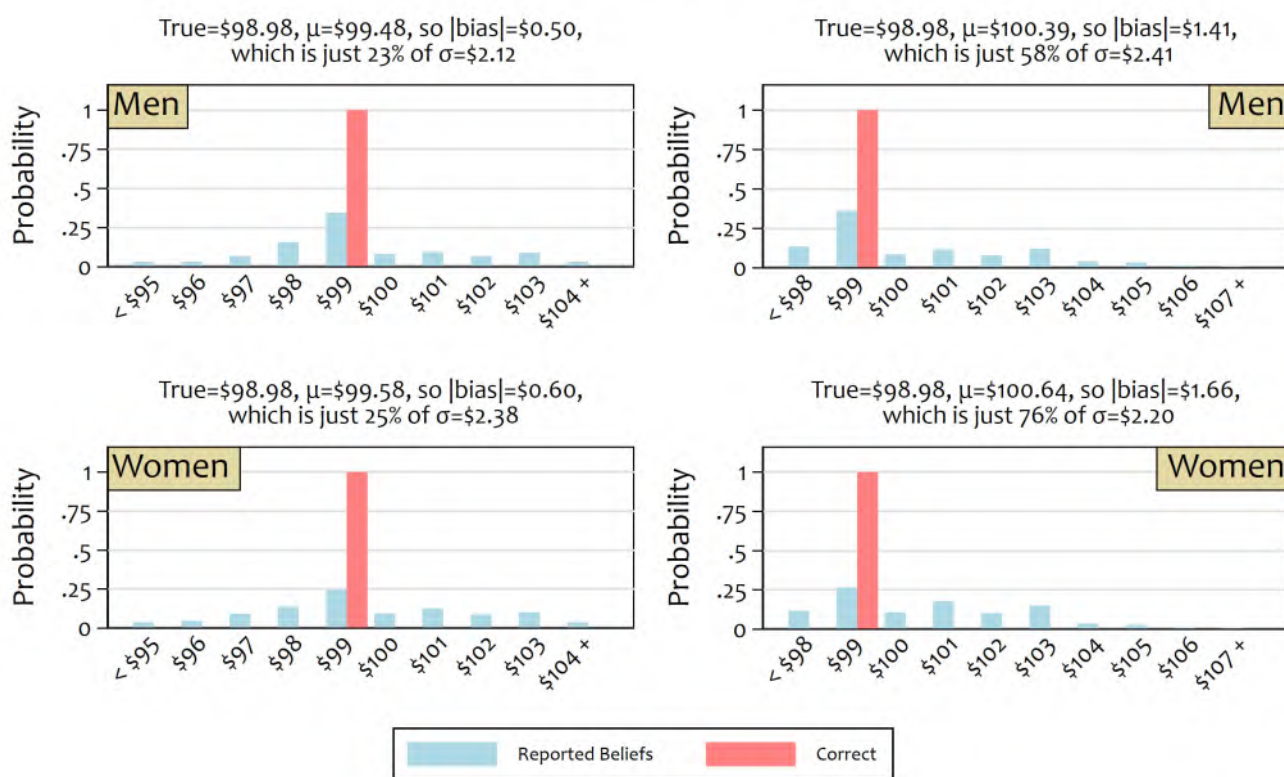
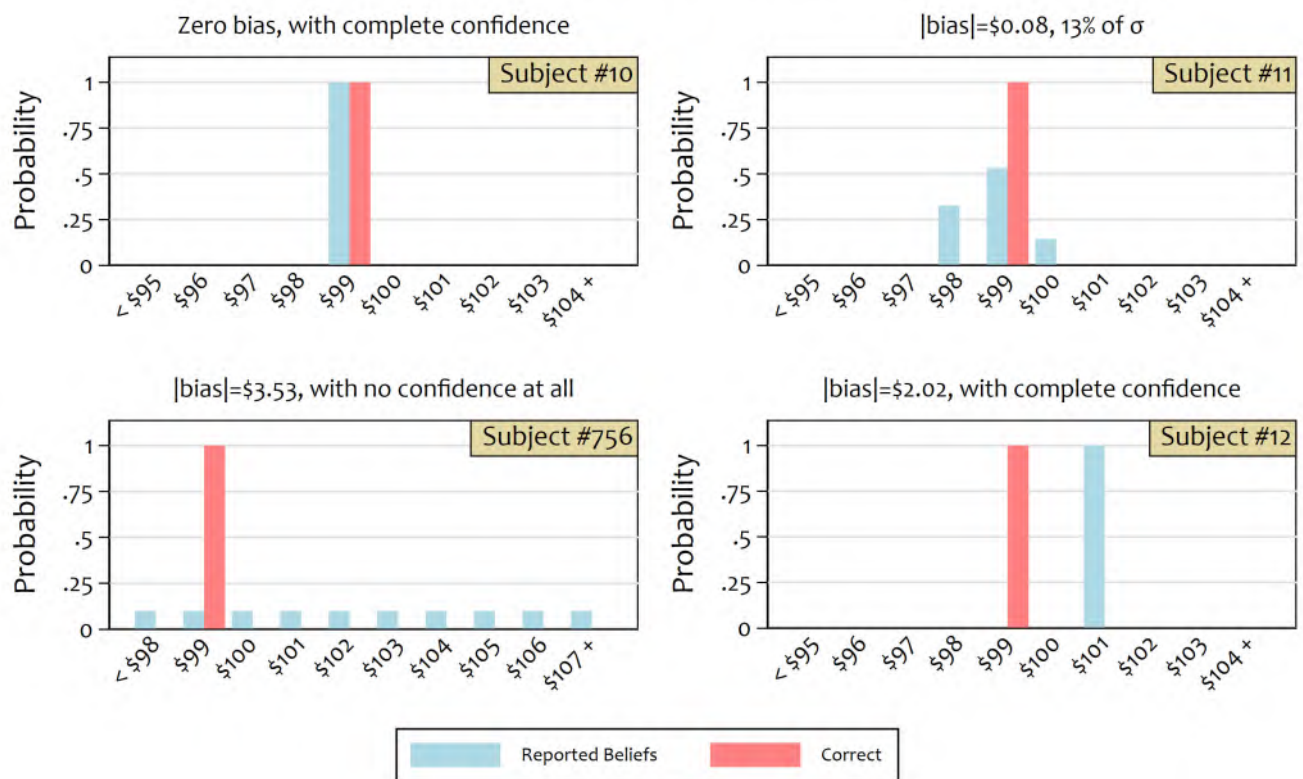


Figure B18: Literacy and Confidence:
The Good, the Bad, and the Ugly



Appendix C: Gender and Competitiveness (Online Working Paper)

In Section 1 we provide a numerical example of the welfare calculations of behavior using data drawn from the experiment of Niederle and Vesterlund [2007]. We refer to the core results in the text, and use the same logic to evaluate behavior in our own experiment. In Section 2 we provide the instructions used in our extension of their design. These instructions refer to the 80 token condition, and were different in obvious ways for the 1 token condition. In Section 3 we show additional results from our experiments, referred to in the text, as well as used in Table 2.

1. Numerical Examples

The *Excel* display in Table C1 uses data on average solutions from Niederle and Vesterlund [2007; Table I, p. 1081]. These values are in columns C and D. The payoffs in cells E, F and G then reflect the payoffs described in the experiment. Cell E4 is the average female production under a piece rate scheme, multiplied by \$0.50 per unit. Similarly, cell F5 assumes that the average woman produced 11.77 units in their Task 2 with a tournament would win, and hence receives \$2 times 11.77 units. Sadly, if she loses, she receives the \$0 in cell G4, no matter how many units she produces. Cells H4, I4 and J4 just fill in the probabilities explained in the text, under the null hypothesis that subjective probabilities of winning equal empirical probabilities of winning.

We assume EUT for the moment. Nothing hinges on that, other than ease of exposition of the basic point. Further assuming a CRRA utility function $u(x) = x^{1-r}/(1-r)$, in column K three possible values of the CRRA parameter r are considered. Risk neutrality is when $r = 0$, very modest risk aversion is when $r = 0.1$, and a more realistic risk aversion is considered when $r = 0.53$ for women and $r = 0.43$ for men. These estimates come from a sample of 614 GSU undergraduates, in experiments reported in Harrison, Morsink and Schneider [2022], with roughly 60% of the sample being female.

It is then a simple matter to calculate the utility of each possible payoff, the expected utility (EU) of the safe piece rate lottery and the EU of the risky tournament lottery. These calculations are in columns S, T and U and W and X, respectively, of the Excel (not shown in Table D1). Armed with the EU of each compensation scheme, and the CRRA parameter, we can calculate the Certainty Equivalent (CE) of selecting each compensation scheme, since $u(CE) = EU$ solves for $CE = [EU \times (1-r)]^{1/(1-r)}$ for this CRRA utility function. The CE of each scheme is displayed in columns L and M. The difference in the CE, assuming the tournament scheme is chosen, is then displayed in column N. This is the Expected Consumer Surplus (ECS) of selecting the tournament scheme.

We observe confirmation of the intuition stated in the text. If subjects are all risk neutral, and $r = 0$, then it does make sense to select the tournament scheme. Even for women that actually selected the piece rate scheme, if they had been risk neutral they would have gained in welfare terms by the equivalent of receiving \$0.71 for certain. For risk-neutral women that actually selected the tournament scheme, the welfare gain would have been \$1.07. For men, in each case, the welfare gains would have been \$0.59 and \$0.77, respectively.

The rub comes when we consider risk aversion. In all but one case, women that chose the tournament, being modestly risk averse with $r = 0.1$ means a welfare loss from selecting the tournament. These welfare losses are only around \$0.13, and \$0.20, but it is the negative sign that matters.

If we consider more realistic risk aversion estimates of the CRRA parameter r , the welfare losses from selecting the tournament become noticeable. In fact, we can state the welfare effects precisely:

- The women that chose the piece rate scheme *gained* in welfare terms by +\$3.94.
- The women that chose the tournament scheme *lost* in welfare terms by \$3.65.
- The men that chose the piece rate scheme *gained* in welfare terms by +\$3.01.
- The men that 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, in their Task 3 and assuming that EUT describes their risk preferences, by selected the tournament less than 50% of the time: only 35% of the time. But men tended to do the wrong thing, by selecting the tournament more than 50% of the time: 73% of the time, in fact.

The *Excel* example in Table C2 corresponds to the RDU example summarized in Table 2 in the text, using data from our experiments.

Table C1: Some EUT Finger-Math Applied to the Niederle and Vesterlund [2007] Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Selected	Average Performance		Payoffs				Probabilities		CRRAR	Certainty Equivalents		ECS
2	Gender	Scheme	Piece Rate	Tournament	Piece Rate	Tournament if Win	Tournament if Lose	Piece Rate	Tournament Win	Tournament Lose		Piece Rate	Tournament	Tournament
3														
4	Female	Piece Rate	10.35	11.77	\$5.18	\$23.54	\$0.00	1	0.25	0.75	0	\$5.18	\$5.89	\$0.71
5			10.35	11.77	\$5.18	\$23.54	\$0.00	1	0.25	0.75	0.1	\$5.18	\$5.05	-\$0.13
6			10.35	11.77	\$5.18	\$23.54	\$0.00	1	0.25	0.75	0.53	\$5.18	\$1.26	-\$3.92
7														
8	Female	Tournament	9.79	11.93	\$4.90	\$23.86	\$0.00	1	0.25	0.75	0	\$4.90	\$5.97	\$1.07
9			9.79	11.93	\$4.90	\$23.86	\$0.00	1	0.25	0.75	0.1	\$4.90	\$5.11	\$0.22
10			9.79	11.93	\$4.90	\$23.86	\$0.00	1	0.25	0.75	0.53	\$4.90	\$1.27	-\$3.62
11														
12	Male	Piece Rate	9.91	11.09	\$4.96	\$22.18	\$0.00	1	0.25	0.75	0	\$4.96	\$5.55	\$0.59
13			9.91	11.09	\$4.96	\$22.18	\$0.00	1	0.25	0.75	0.1	\$4.96	\$4.75	-\$0.20
14			9.91	11.09	\$4.96	\$22.18	\$0.00	1	0.25	0.75	0.43	\$4.96	\$1.96	-\$3.00
15														
16	Male	Tournament	10.97	12.52	\$5.49	\$25.04	\$0.00	1	0.25	0.75	0	\$5.49	\$6.26	\$0.78
17			10.97	12.52	\$5.49	\$25.04	\$0.00	1	0.25	0.75	0.1	\$5.49	\$5.37	-\$0.12
18			10.97	12.52	\$5.49	\$25.04	\$0.00	1	0.25	0.75	0.43	\$5.49	\$2.21	-\$3.28

Table C2: Some RDU Finger-Math Applied to Our Experimental Data

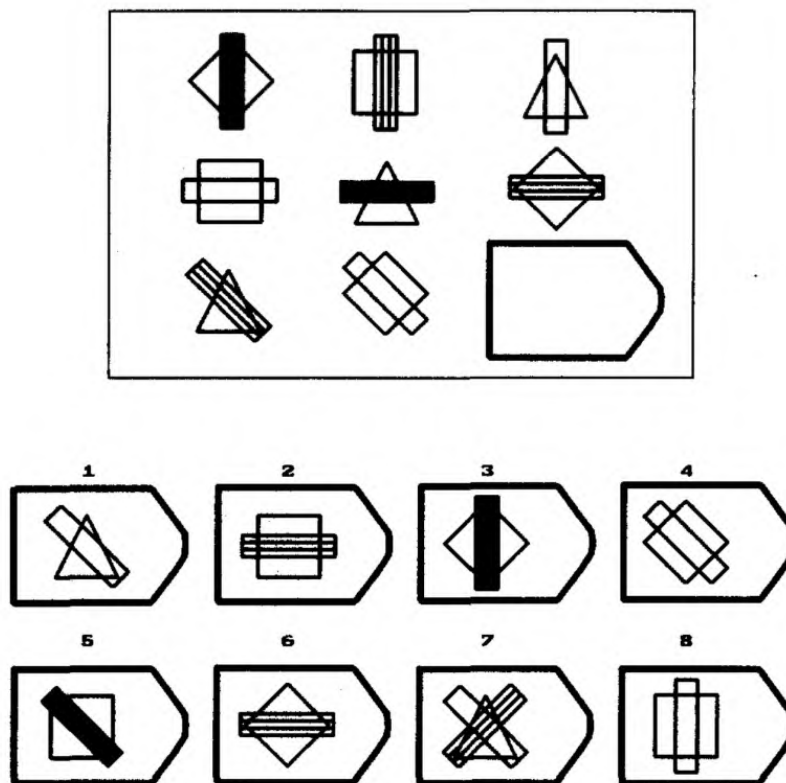
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1			Payoffs			Probabilities			CRRAR	PWF		Certainty Equivalents		ECS
2	Gender		Piece Rate	Tournament if Win	Tournament if Lose	Piece Rate	Tournament Win	Tournament Lose	r	η	φ	Piece Rate	Tournament	Tournament
3														
4	Female		\$1.36	\$5.72	\$0.00	1	0.33	0.67	0	1	1	\$1.36	\$1.89	\$0.53
5			\$1.36	\$5.72	\$0.00	1	0.33	0.67	0.1	1.25	1	\$1.36	\$1.23	-\$0.13
6			\$1.36	\$5.72	\$0.00	1	0.33	0.67	0.70	0.90	0.92	\$1.36	\$0.26	-\$1.10
7														
8	Male		\$1.37	\$5.66	\$0.00	1	0.23	0.77	0	1	1	\$1.37	\$1.30	-\$0.07
9			\$1.37	\$5.66	\$0.00	1	0.23	0.77	0.1	1.25	1	\$1.37	\$0.74	-\$0.63
10			\$1.37	\$5.66	\$0.00	1	0.23	0.77	0.71	0.97	1.00	\$1.37	\$0.07	-\$1.30

WELCOME!

The final task today consists of six parts. The first part will take a bit longer than the rest, since it provides an introduction that will apply to all parts. We will randomly select one of the six parts of this task and pay you based on your performance in that part. Once you have completed the six parts we determine which part counts for payment by drawing a number between 1 and 6. The method we use to determine your earnings varies across parts. Before each part we will describe in detail how your payment is determined.

Part 1 – Piece Rate

This part of the task is a test of perception and clear thinking. In a previous session you were given 12 similar problems. We will now consider a fresh set of 12 problems, like this one:



The top part of this problem is a pattern with a piece cut out of it. Look at the pattern, and think what the piece needed to complete the pattern correctly both across and down must be like. Then find the right piece out of the eight pieces shown, and numbered, in the bottom part of this problem.

For **Part 1** you will be asked to report your beliefs about 12 problems like this one, and you will have up to 10 minutes to do so. If **Part 1** is the part randomly selected for payment, then you get up to 200 points per problem you solve correctly in the 10 minutes. If **Part 1** is the part randomly selected for payment, then each point you earn gets converted to money, and 100 points is converted to \$1. So if you earn 200 points per problem you solve correctly, in **Part 1** you will earn \$2 per problem you solve correctly. Your payment does not decrease if you provide an incorrect answer to a problem. We refer to this payment as the **Piece Rate** payment.

You will be asked to report your beliefs about the correct answer using an interface like this one, which is also generally familiar to you from a previous task. The version you will see on the computer will be larger and easier to read.



The problem and possible solutions are shown on the left of the screen, in the usual manner. On the right of the screen you have 80 tokens to allocate across the 8 possible answers. We start off with 0 tokens allocated to each of the possible answers. If you wanted to change this initial allocation so that there were 10 tokens allocated to each possible answer, just click on the **Uniform** button, and you will see this display:

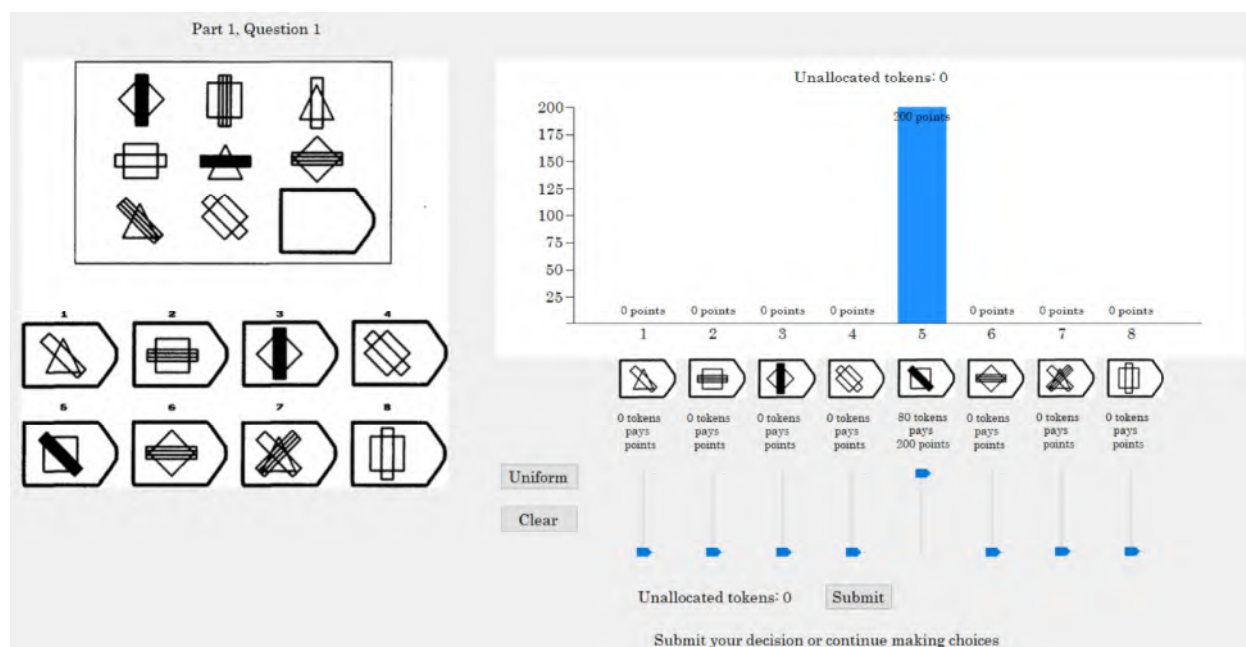


Here you have allocated 10 tokens to each possible answer, and you would earn 112 points if you reported this allocation of tokens, since only one of the 8 possible answers is correct. You can return to the initial allocation of 0 tokens for each possible answer by clicking on the **Clear** button.

As you allocate tokens, by moving the sliders up or down, the points will change above each bar. These are the points that you will receive for this problem if that bar corresponds to the correct answer to the problem. **You will be paid for all 12 problems, and each problem will pay between 0 points and 200 points depending on your answer.**

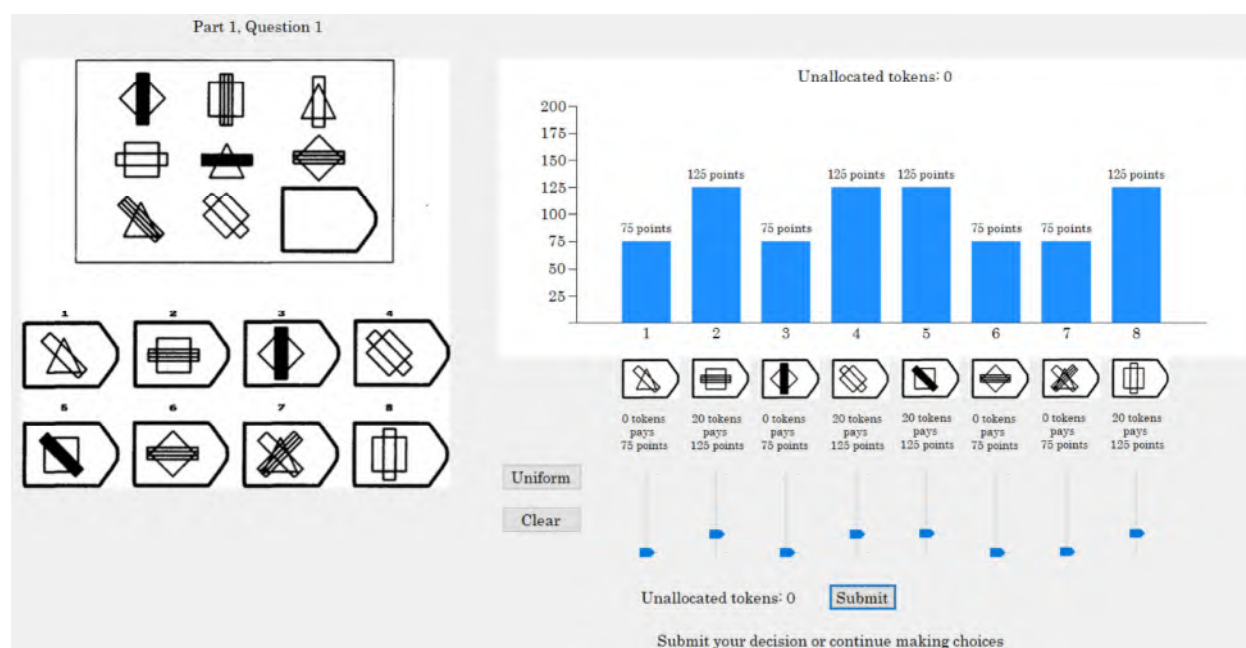
Return now to the problem itself. Only one of these pieces is perfectly correct. Pieces #2, #5 and #8 complete the problem correctly going down, in the sense that they have a square as the larger piece. Pieces #1, #4 and #5 are correct going across, in the sense that they have the correct slope for the rectangle.

Piece #5 is the correct piece both ways, isn't it? It is correct going down as well as going across. So the answer is #5. In this case, if you were certain that piece #5 is the single best answer, you should allocate all 80 tokens to the bin representing piece #5 as in this display:



So in this case you would earn 200 points if indeed the correct answer was #5. Of course, if any of the other pieces turned out to be the correct answer you would, in this case, earn 0 points.

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:



You can see that even if you eliminate some pieces, such as #1, #3, #6 and #7, that are clearly wrong, you give yourself a 1-in-4 chance of earning more money than if you guessed across all 8 pieces. In this case you would expect to earn 125 points if indeed one of pieces #2, #4, #5 or #8 had been correct. Recall that if you had allocated the tokens roughly equally across all 8 bars, thinking that any of the 8 pieces might be correct, you would only earn 112 points for this problem.

This illustrative problem is relatively easy, but some of the problems will not be so easy, and it might be harder to identify the one correct answer. You would still benefit by ruling out certain pieces so that you can allocate more tokens to the ones that you have not ruled out.

The order in which you see the easier and harder problems, as well as the ones in between, will be selected at random. So the problems do not generally become easier or harder as you work through them.

Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. The important thing is to notice how the problems can be solved, and to learn the method of solving them. In each problem you look across each row, or down each column, and decide what the missing piece should be like. Then look for the answer that is correct both ways, across **and** down, from the options you have to select from.

As you recall from allocating tokens in a similar part in a previous session, you can change your token allocation as many times as you like before you hit the **Submit** button and **Confirm** your token allocation. You can also review and modify your prior decisions. You will have navigation buttons to the left of the sliders. These navigation buttons are only available before you start moving the sliders for a problem.

You will be allowed 10 minutes to complete this part. Your screen displays how many minutes (and seconds) are remaining in the top, right corner. Remember it is accurate work that counts. Attempt each problem in turn. Do your best to eliminate the incorrect pieces and find the correct piece to complete it before going on to the next problem. You will have the option to go back to previous problems if you want to. After you have completed the last question you must confirm if you are finished. When you are finished you will be told your total points **if Part 1 is selected for payment**, and will not have the option to go back to any of the problems. If you run out of time, you will automatically be treated as having finished, and told your total points if Part 1 is selected for payment.

Where you position each slider depends on your beliefs about the correct answer to the question. Again, each bar shows the number of points you could earn if the true outcome corresponds to the possible solution shown under the bar.

You will be **paid for each of the 12 problems**, so you should think carefully about each problem. Since you can earn up to 200 points for each problem, you could earn up to 2,400 points over all 12 problems. And **if Part 1 is selected for payment**, these 2,400 points would be converted to \$24. You will not earn anything on any problems for which you have not confirmed an allocation of tokens.

It is up to you to balance the strength of your personal beliefs with the possibility of them being wrong. There are several important points for you to keep in mind when making your decisions:

- First, your belief about the correct answer to each problem is your personal judgment.
- Second, you have up to 10 minutes to complete this part.
- Third, you will not earn anything on any problems for which you have not confirmed an allocation of tokens.
- Fourth, depending on your choices and the correct answer you can earn up to 200 points for each problem, or up to 2,400 points over all 12 problems.
- Finally, your choices might also depend on your willingness to take risks or to gamble.

The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are finished you will see a summary of your responses and your total earnings from this part of the task if it is selected for payment.

Part 2 – Tournament

As in **Part 1** you will be given 10 minutes to report your beliefs about the solution to up to 12 problems like the ones in **Part 1**. However, for this part your payment depends on the points earned from your reports compared to the points of three other participants in your group. Each group consists of four people, selected at random from other participants in **Part 2**.

If **Part 2** is the part of the task randomly selected for payment, then your cash earnings depend on the points you get from your reports compared to the points of the three other people in your group. The individual with the most points over all 12 problems will receive up to \$8 per problem, while the other participants receive no cash payment for **Part 2**. In other words, every 100 points you get from your reports gets converted to \$4 in cash, and every 200 points you get from your reports gets converted to \$8 in cash. We refer to this as the **Tournament** payment. You will not be told how you did in the tournament until all six parts have been completed, although you will know your own cash earnings in **Part 2** if it is selected for payment and you were the winner. If there are ties the winner will be randomly selected.

Part 3 – Choose Payment Scheme

As in the previous two parts, you will be given 10 minutes to report your beliefs about the solution to up to 12 problems like the ones in **Part 1** and **Part 2**. However, in **Part 3** you will now get to choose which of the two previous payment schemes you prefer to apply to your points. Here is the screen you will face to make your choice.

Part 3 - Choice

Previously:

- In **Part 1** your potential earnings were calculated as a **Piece Rate** payment.
- In **Part 2** your potential earnings were calculated as a **Tournament** payment.

Now you must select which of these payment schemes you prefer for the next 12 questions:

Piece rate	You receive up to \$2 for each question you correctly solve in this Part.
Tournament	You receive up to \$8 for each question you correctly solve in this Part if your points in Part 3 are larger than the points in Part 2 of the other 3 members of your group.

If **Part 3** is the part of the task randomly selected for payment, then your cash earnings for this task are determined as follows.

- If you choose the **Piece Rate** payment you receive up to \$2 per problem you solve correctly. So, as in Part 1, every 100 points you get from your reports converts to \$1 in cash, and every 200 points you get from your reports converts to \$2 in cash.
- If you choose the **Tournament** payment your points will be compared to the points of the other three participants in your group in **Part 2 – Tournament**. The **Part 2 – Tournament** is the one you just completed. If your points in **Part 3** are larger than their reports in **Part 2**, then you receive four times the payment from the piece rate: you receive up to \$8 per problem you solve correctly. So, as in **Part 2**, every 100 points you get from your reports converts to \$4 in cash, and every 200 points you get from your reports converts to \$8 in cash. You will receive no earnings for this part if you choose the **Tournament** and points from your reports in **Part 3** are smaller than the points from the reports of the three others in your group in **Part 2**. You will not be informed of how you did in the **Tournament** until all six parts have been completed. If there are ties the winner will be randomly determined.

You are asked to choose whether you want the **Piece Rate** or the **Tournament** applied to your performance. You will then be given 10 minutes to report your beliefs on the solution to up to 12 problems.

Part 4 – Submit Your Piece Rate Performance

You do not have to report your beliefs on any more problems for **Part 4**. Instead, you may be paid one more time for your earnings in **Part 1 – Piece Rate** of the task. However, you now must choose which payment scheme you want applied to your points from **Part 1**. You can either choose to be paid according to the **Piece Rate**, or according to the **Tournament**. Here is the screen you will face to make your choice.

Part 4

Previously:

- In **Part 1** your potential earnings were calculated as a **Piece Rate** payment.

If **Part 4** is selected for payment, you will be paid according to your **Part 1** performance.

Now you must select which payment scheme you prefer for **Part 4**:

Piece rate	You receive up to \$2 for each question you correctly solve in Part 1 .
Tournament	You receive up to \$8 for each question you correctly solve in Part 1 if your points in Part 1 are larger than the points in Part 1 of the other 3 members of your group.

If **Part 4** is the part of the task selected for payment, then your earnings for this task are determined as follows.

- If you choose the **Piece Rate** payment you receive up to \$2 per problem you correctly solved in **Part 1**, so 100 points again converts to \$1 and 200 points again converts to \$2.
- If you choose the **Tournament** payment your performance will be compared to the performance of the other three participants in your group in **Part 1 – Piece Rate**. If the points from your reports in **Part 1** are larger than the points from their reports in **Part 1**, then you receive four times the payment compared to the piece rate: you receive up to \$8 per problem you solve correctly, since in this case 100 points converts to \$4 and 200 points converts to \$8. You will receive no earnings for this part if you choose the **Tournament** and your points from **Part 1** are smaller than the points of the three others in your group in **Part 1**.


Part 5 – Beliefs About Your Part 1 Rank

In **Part 5** you are asked to report your beliefs about the rank that *your* points in the **Part 1 – Piece Rate** task had, when compared to the points in **Part 1** of the other three participants of your group. Here is the screen you will face to make your choice.

Part 5

In **Part 1 - Piece Rate**, how did your points compare to the points of the other 3 participants in your group?

Unallocated tokens: 100



Rank	Payoff (\$)
Best	\$15.00
Second Best	\$15.00
Third Best	\$15.00
Fourth Best	\$15.00

0 tokens 0 tokens 0 tokens 0 tokens

Pays \$15.00 Pays \$15.00 Pays \$15.00 Pays \$15.00

Uniform

Clear

Unallocated tokens: 100 Submit

You must allocate all tokens before you are able to submit

You have 100 tokens to allocate. And you must allocate them over the four possible outcomes for the rank of your points in the **Part 1 – Piece Rate** task compared to the rank of the points of the other three participants in your group. If **Part 5** is selected for payment you will be paid the cash earnings indicated on this screen for the actual rank of your points, which is based on the number of tokens you allocate to that points rank outcome.

Part 6 – Beliefs About Your Part 2 Rank

In **Part 6** you are asked to report your beliefs about the rank that *your* points in the **Part 2 – Tournament** task had, when compared to the points in **Part 2** of the other three participants of your group. Here is the screen you will face to make your choice.

Part 6

In **Part 2 - Tournament**, how did your points compare to the points of the other 3 participants in your group?

Unallocated tokens: 100



Rank Outcome	Payoff (\$)
Best	15.00
Second Best	15.00
Third Best	15.00
Fourth Best	15.00

0 tokens 0 tokens 0 tokens 0 tokens

Pays \$15.00 Pays \$15.00 Pays \$15.00 Pays \$15.00

Uniform

Clear

Unallocated tokens: 100 Submit

You must allocate all tokens before you are able to submit

You again have 100 tokens to allocate. And you must allocate them over the four possible outcomes for the rank of your points in the **Part 2 – Tournament** task compared to the rank of the points of the other three participants in your group. If **Part 6** is selected for payment you will be paid the cash earnings indicated on this screen for the actual rank of your points, which is based on the number of tokens you allocate to that points rank outcome.

3. Additional Results

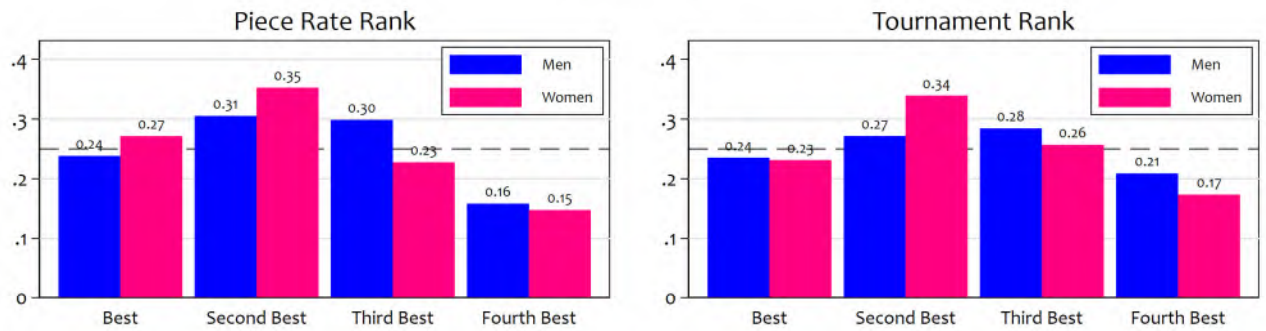
Figures C1 and C2 provide detailed results on the reported beliefs of men and women, or Blacks and non-Blacks, referred to in the text.

Figures C3 and C4 provide displays of the reported beliefs and recovered beliefs of subject #2. This subject had a concave utility function and probability weighting implying very slight, global probability pessimism, so relatively close to EUT risk preferences. Her reports for the QSR are therefore slightly smoother than her beliefs. She also exhibits strikingly different beliefs over her piece rate ranks and her tournament ranks, with considerable confidence of winning the tournament.

Figures C5 and C6 provide displays of the reported beliefs and recovered beliefs of subject #3. This subject also had a concave utility function, but his probability weighting implying significant, global probability pessimism, so clearly not close to EUT risk preferences. His reports for the QSR are again smoother than his beliefs, but the difference between beliefs and reports is much more substantial than for subject #2. He reported the same beliefs over his piece rate ranks and his tournament ranks, with considerable confidence of being ranked second or third in both.

Figure C1: Recovered Beliefs on the Probability of Personal Performance Ranks by Men and Women

A. One Token



B. Eighty Tokens

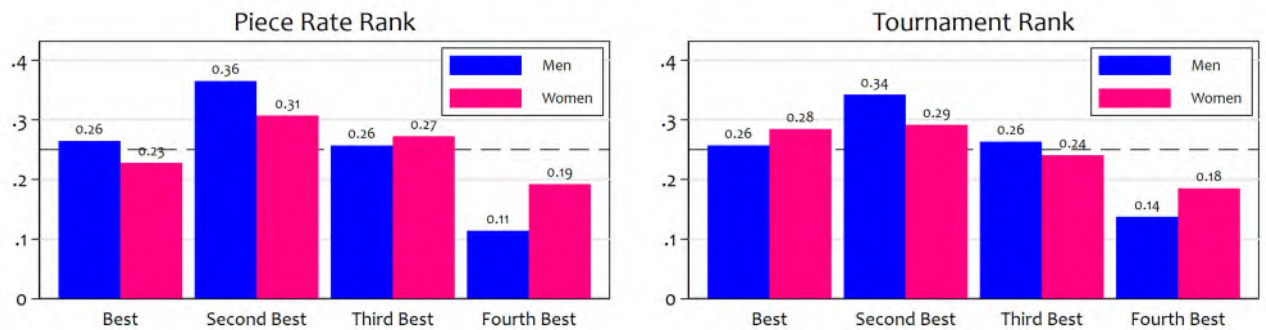
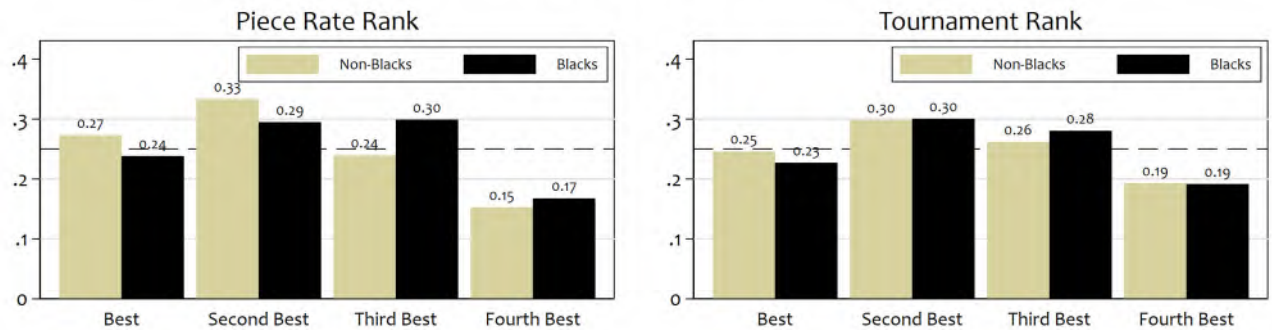


Figure C2: Recovered Beliefs on the Probability of Personal Performance Ranks by Blacks and Non-Blacks

A. One Token



B. Eighty Tokens

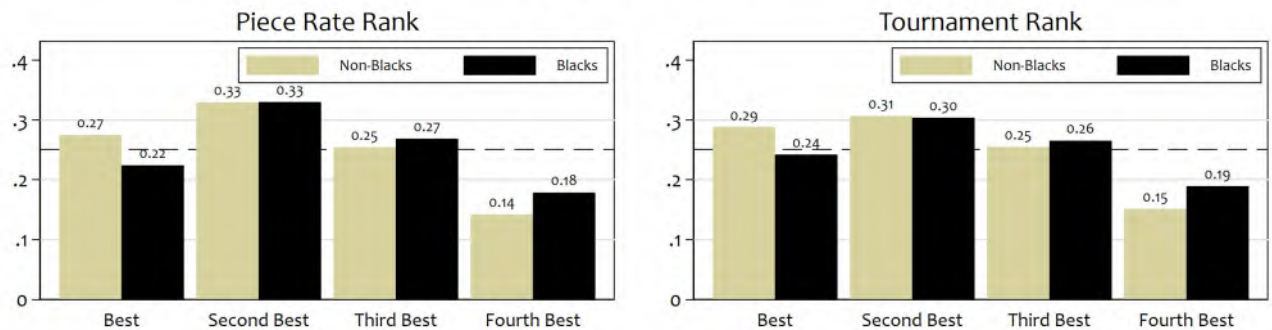


Figure C3: Beliefs of Subject #2 for the Piece Rate Rank Question

Posterior mean RDU risk preferences are $r = 0.68$, $\eta = 1.22$ and $\varphi = 1.02$

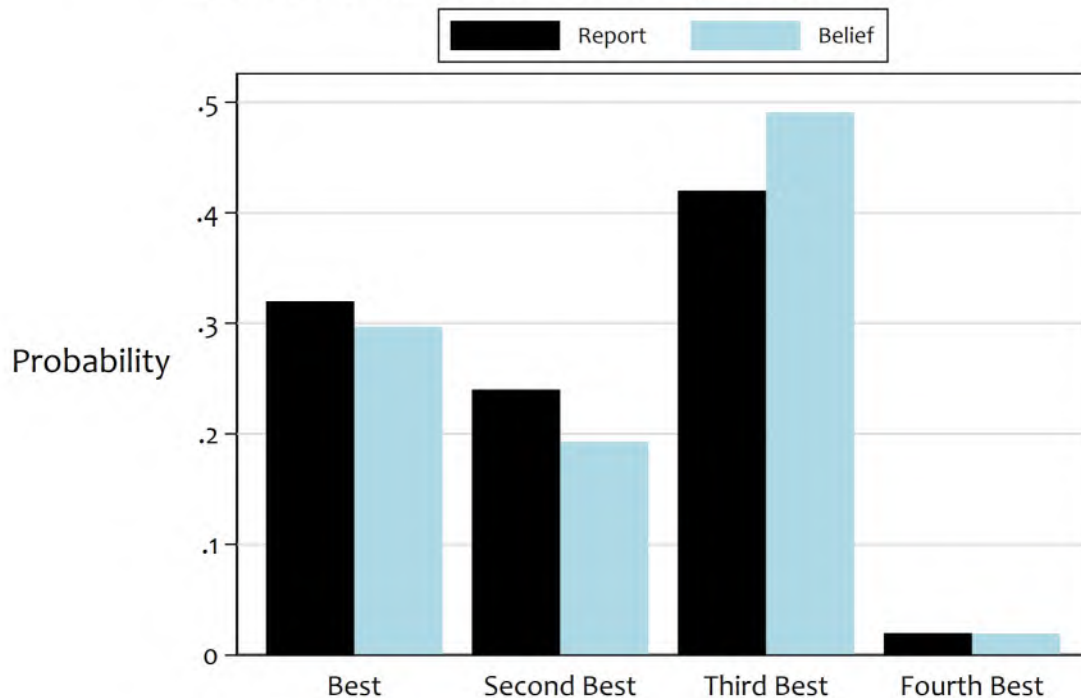


Figure C4: Beliefs of Subject #2 for the Tournament Rank Question

Posterior mean RDU risk preferences are $r = 0.68$, $\eta = 1.22$ and $\varphi = 1.02$

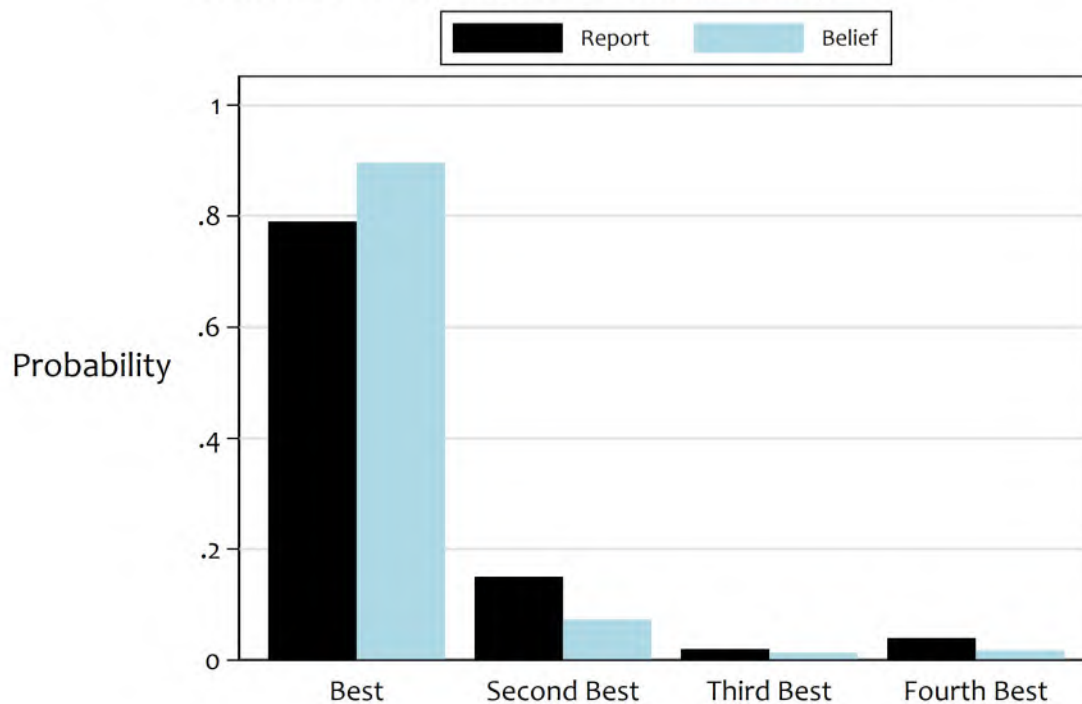


Figure C5: Beliefs of Subject #3 for the Piece Rate Rank Question

Posterior mean RDU risk preferences are $r = 0.70$, $\eta = 1.53$ and $\varphi = 0.68$

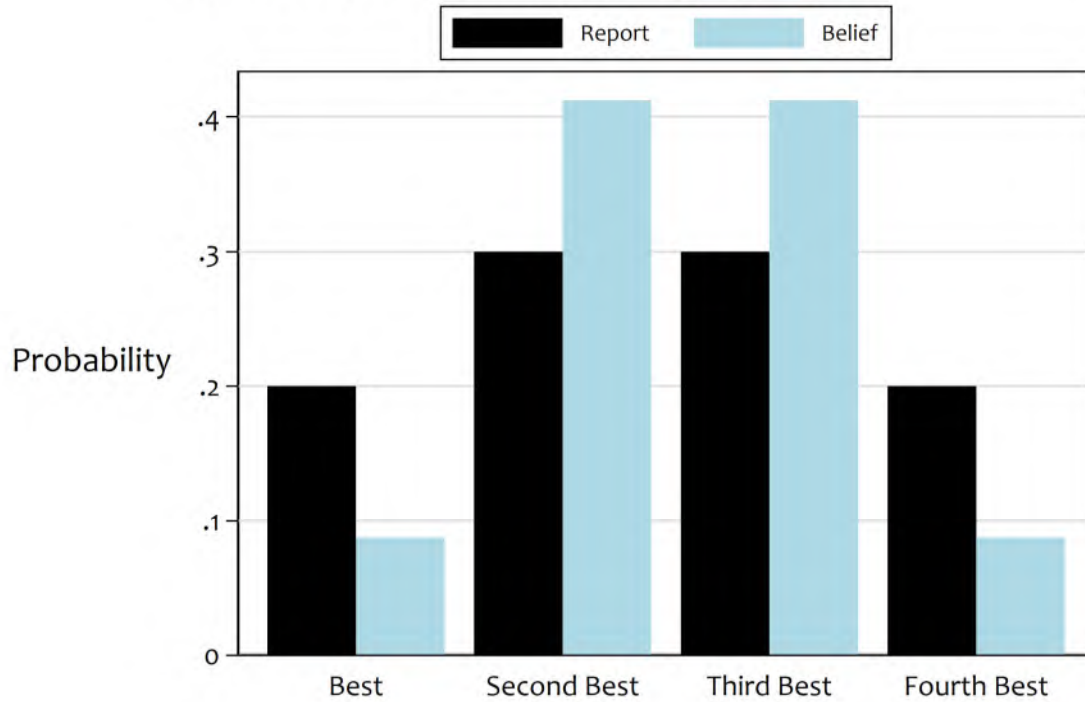
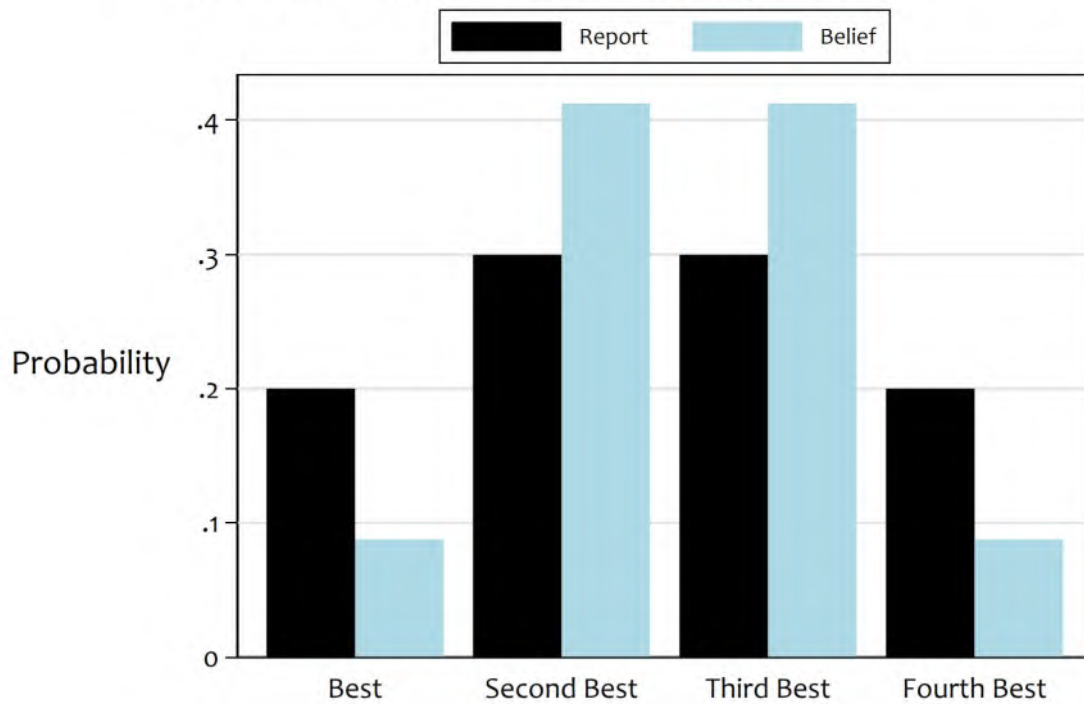


Figure C6: Beliefs of Subject #3 for the Tournament Rank Question

Posterior mean RDU risk preferences are $r = 0.70$, $\eta = 1.53$ and $\varphi = 0.68$



The listing below documents the conclusions drawn from the statistical analysis of qualitative welfare effects, and reported in the text. All beliefs reported are recovered from the observed reports, using the RDU risk preferences of the individual. Similarly, all welfare results reflect those recovered beliefs and the RDU risk preferences of the individual. Comparable results using the EUT risk preferences of the individual are also listed. The parameters used in Table 2 are also documented.

Variable name	Storage type	Display format	Value label	Variable label
welfare_gain	float	%9.0g		Actual welfare gain (=1) or loss (=0)
female	byte	%8.0g	flab	Female
age	byte	%10.0g		Age
black	byte	%8.0g		Black
business	byte	%8.0g		Business major
noreligion	byte	%8.0g		Atheist, Nonreligious, or Agnostic
bfi_extra	byte	%8.0g		BFI Extraversion
bfi_agree	byte	%8.0g		BFI Agreeableness
bfi_cons	byte	%8.0g		BFI Conscientiousness
bfi_neur	byte	%8.0g		BFI Neuroticism
bfi_open	byte	%8.0g		BFI Openness
optimistT	float	%9.0g		Subjective probability of winning greater than 0.25 with tournament performance
optimistP	float	%9.0g		Subjective probability of winning greater than 0.25 with piece-rate performance
earnT	float	%9.0g		Piece-Rate earnings in top 25%
earnP	float	%9.0g		

```

. * result 1: overall welfare gain RDU

. summ welfare_gain if rdu==1

  Variable |      Obs      Mean   Std. dev.   Min   Max
-----+-----
welfare_gain |      178   .6910112   .46338      0      1

. * comparison to EUT
. summ welfare_gain if rdu==0

  Variable |      Obs      Mean   Std. dev.   Min   Max
-----+-----
welfare_gain |      178   .7022472   .4585601      0      1

. * result 2: welfare gain RDU with stress or no stress

. summ welfare_gain if rdu==1 & submit==0

  Variable |      Obs      Mean   Std. dev.   Min   Max
-----+-----
welfare_gain |       89   .6516854   .4791357      0      1

. summ welfare_gain if rdu==1 & submit==1

  Variable |      Obs      Mean   Std. dev.   Min   Max
-----+-----
welfare_gain |       89   .7303371   .446299      0      1

. * comparison to EUT
. summ welfare_gain if rdu==0 & submit==0

  Variable |      Obs      Mean   Std. dev.   Min   Max
-----+-----
welfare_gain |       89   .6853933   .46699      0      1

. summ welfare_gain if rdu==0 & submit==1

  Variable |      Obs      Mean   Std. dev.   Min   Max
-----+-----
welfare_gain |       89   .7191011   .4519846      0      1

. * result 3: unconditional welfare gain RDU for women

. tab welfare_gain female if rdu==1, col nofreq exact

  Actual |
welfare |
gain (=1) |
or loss |
  (=0) |      Female
      Male   Female |      Total
-----+-----
0 |      30.00   32.89 |      31.25
1 |      70.00   67.11 |      68.75
-----+-----
Total |     100.00   100.00 |     100.00

      Fisher's exact =           0.744
1-sided Fisher's exact =           0.402

. * comparison to EUT
. tab welfare_gain female if rdu==0, col nofreq exact

  Actual |
welfare |
gain (=1) |
or loss |
  (=0) |      Female
      Male   Female |      Total
-----+-----
0 |      30.00   30.26 |      30.11
1 |      70.00   69.74 |      69.89
-----+-----
Total |     100.00   100.00 |     100.00

      Fisher's exact =           1.000
1-sided Fisher's exact =           0.550

```

. * result 4: conditioning on demographics, BFI and optimism and earnings

```
. probit welfare_gain `demog_i' `bfi' i.optimistT i.earnT i.optimistT#i.earnT
i.female#i.eighty_tokens if rdu==1, vce(cluster sid)
```

Probit regression

Number of obs = 176
Wald chi2(15) = 22.59
Prob > chi2 = 0.0932
Pseudo R2 = 0.0963

Log pseudolikelihood = -98.784529

(Std. err. adjusted for 88 clusters in sid)

		Robust				
welfare_gain	Coefficient	std. err.	z	P> z	[95% conf. interval]	
female						
Female	.3165435	.3648816	0.87	0.386	-.3986112	1.031698
age	-.0154132	.0820441	-0.19	0.851	-.1762167	.1453904
1.black	.3335434	.2242917	1.49	0.137	-.1060602	.773147
1.business	.1071957	.2740318	0.39	0.696	-.4298968	.6442883
1.noreligion	.1896377	.2511163	0.76	0.450	-.3025412	.6818165
bfi_extra	-.1215833	.165013	-0.74	0.461	-.4450028	.2018361
bfi_agree	.2718446	.1945039	1.40	0.162	-.1093761	.6530652
bfi_cons	-.2898139	.1871696	-1.55	0.122	-.6566596	.0770318
bfi_neur	-.0455818	.1930262	-0.24	0.813	-.4239062	.3327426
bfi_open	.2556879	.2162158	1.18	0.237	-.1680872	.679463
1.optimistT	-.788307	.5783565	-1.36	0.173	-1.921865	.345251
1.earnT	-.9688624	.3484725	-2.78	0.005	-1.651856	-.2858689
optimistT#earnT						
1 1	.7409653	.6366913	1.16	0.245	-.5069268	1.988857
female#eighty_tokens						
Male#1	.4930705	.316505	1.56	0.119	-.127268	1.113409
Female#1	-.1092502	.3331375	-0.33	0.743	-.7621877	.5436873
_cons	.5339297	1.550506	0.34	0.731	-2.505005	3.572865

Average marginal effects
Model VCE: Robust

Number of obs = 176

Expression: Pr(welfare_gain), predict()
dy/dx wrt: 1.female age 1.black 1.business 1.noreligion bfi_extra bfi_agree bfi_cons bfi_neur
bfi_open 1.optimistT 1.earnT 1.eighty_tokens

		Delta-method				
	dy/dx	std. err.	z	P> z	[95% conf. interval]	
female						
Female	-.0005873	.0830183	-0.01	0.994	-.1633001	.1621255
age	-.0048969	.0260977	-0.19	0.851	-.0560474	.0462536
1.black	.1066111	.0719475	1.48	0.138	-.0344034	.2476257
1.business	.0337188	.0850712	0.40	0.692	-.1330178	.2004554
1.noreligion	.0590086	.0766894	0.77	0.442	-.0912998	.2093171
bfi_extra	-.0386283	.052335	-0.74	0.460	-.141203	.0639465
bfi_agree	.0863678	.060927	1.42	0.156	-.0330468	.2057824
bfi_cons	-.0920768	.0575288	-1.60	0.109	-.2048312	.0206776
bfi_neur	-.0144818	.0612178	-0.24	0.813	-.1344664	.1055028
bfi_open	.0812346	.0685575	1.18	0.236	-.0531355	.2156048
1.optimistT	-.0694125	.0895648	-0.77	0.438	-.2449563	.1061314
1.earnT	-.2058842	.0736567	-2.80	0.005	-.3502487	-.0615197
1.eighty_tokens	.0713186	.073475	0.97	0.332	-.0726897	.2153269

Note: dy/dx for factor levels is the discrete change from the base level.

. * comparison to EUT

```
. probit welfare_gain `demog_i' `bfi' i.optimistT i.earnT i.optimistT#i.earnT
i.female#i.eighty_tokens if rdu==0, vce(cluster sid)
```

Probit regression

Number of obs = 176
Wald chi2(15) = 23.28
Prob > chi2 = 0.0783
Pseudo R2 = 0.0970

Log pseudolikelihood = -97.239548

(Std. err. adjusted for 88 clusters in sid)

		Robust				
welfare_gain	Coefficient	std. err.	z	P> z	[95% conf. interval]	
female						
Female	.2437876	.3744595	0.65	0.515	-.4901395	.9777148
age	-.0337342	.0800274	-0.42	0.673	-.1905851	.1231166
1.black	.2994134	.2185508	1.37	0.171	-.1289383	.7277651
1.business	.3843431	.268255	1.43	0.152	-.1414271	.9101133
1.noreligion	.1969226	.2326391	0.85	0.397	-.2590418	.6528869
bfi_extra	-.0036687	.1547591	-0.02	0.981	-.3069909	.2996535
bfi_agree	.3142669	.1899836	1.65	0.098	-.0580941	.6866279
bfi_cons	-.3144162	.1761357	-1.79	0.074	-.6596359	.0308034
bfi_neur	-.0313468	.1989629	-0.16	0.875	-.4213069	.3586133
bfi_open	.3098938	.1987146	1.56	0.119	-.0795796	.6993672
1.optimistT	-.6743993	.5726012	-1.18	0.239	-1.796677	.4478783
1.earnT	-.9263598	.3615647	-2.56	0.010	-1.635014	-.2177061
optimistT#earnT						
1 1	.6714149	.6306384	1.06	0.287	-.5646138	1.907443
female#eighty_tokens						
Male#1	.2731307	.3069343	0.89	0.374	-.3284495	.8747109
Female#1	-.0178173	.3267929	-0.05	0.957	-.6583195	.6226849
_cons	-.004545	1.534636	-0.00	0.998	-3.012376	3.003286

Average marginal effects

Number of obs = 176

Model VCE: Robust

Expression: Pr(welfare_gain), predict()

dy/dx wrt: 1.female age 1.black 1.business 1.noreligion bfi_extra bfi_agree bfi_cons bfi_neur
bfi_open 1.optimistT 1.earnT 1.eighty_tokens

		Delta-method				
	dy/dx	std. err.	z	P> z	[95% conf. interval]	
female						
Female	.0281981	.0825764	0.34	0.733	-.1336487	.190045
age	-.010552	.0250349	-0.42	0.673	-.0596195	.0385154
1.black	.0939512	.0692787	1.36	0.175	-.0418327	.229735
1.business	.1153841	.075268	1.53	0.125	-.0321386	.2629067
1.noreligion	.0602455	.0701918	0.86	0.391	-.0773278	.1978188
bfi_extra	-.0011476	.0484119	-0.02	0.981	-.0960332	.093738
bfi_agree	.0983022	.058512	1.68	0.093	-.0163792	.2129836
bfi_cons	-.0983489	.0522622	-1.88	0.060	-.2007809	.0040831
bfi_neur	-.0098052	.0621397	-0.16	0.875	-.1315968	.1119864
bfi_open	.0969343	.0617667	1.57	0.117	-.0241262	.2179948
1.optimistT	-.048208	.0841208	-0.57	0.567	-.2130817	.1166657
1.earnT	-.1973012	.0714649	-2.76	0.006	-.3373699	-.0572325
1.eighty_tokens	.0460322	.0704389	0.65	0.513	-.0920255	.18409

Note: dy/dx for factor levels is the discrete change from the base level.

. * result(s) 5: no-stress in part 4, conditioning on demographics, BFI, optimism and earnings

```
. probit welfare_gain `demog_i' `bfi' i.optimistP i.earnP i.earnP i.optimistP#i.earnP
i.female#i.eighty_tokens if rdu==1 & submit==1, vce(cluster sid)
```

```
Probit regression                                Number of obs =      88
                                                Wald chi2(15) =   21.16
                                                Prob > chi2     =  0.1317
Log pseudolikelihood = -42.372448                Pseudo R2       =  0.1783
```

(Std. err. adjusted for 88 clusters in sid)

		Robust				
welfare_gain	Coefficient	std. err.	z	P> z	[95% conf. interval]	
female						
Female	.6480445	.5032921	1.29	0.198	-.3383899	1.634479
age	-.0164485	.1112026	-0.15	0.882	-.2344016	.2015046
1.black	.65977	.334209	1.97	0.048	.0047325	1.314808
1.business	-.0054182	.3547768	-0.02	0.988	-.700768	.6899317
1.noreligion	.5764353	.3700241	1.56	0.119	-.1487986	1.301669
bfi_extra	-.0403915	.2041438	-0.20	0.843	-.440506	.3597229
bfi_agree	.1974102	.2838089	0.70	0.487	-.3588449	.7536654
bfi_cons	-.1107276	.2631052	-0.42	0.674	-.6264044	.4049491
bfi_neur	.0655545	.262408	0.25	0.803	-.4487557	.5798648
bfi_open	-.1677185	.3043648	-0.55	0.582	-.7642626	.4288255
1.optimistP	-1.013151	.7683918	-1.32	0.187	-2.519171	.4928694
1.earnP	-1.435437	.5071451	-2.83	0.005	-2.429423	-.4414504
optimistP#earnP						
1 1	.4728323	.8618582	0.55	0.583	-1.216379	2.162043
female#eighty_tokens						
Male#1	.3209009	.4130765	0.78	0.437	-.4887141	1.130516
Female#1	-.3235379	.5332501	-0.61	0.544	-1.368689	.721613
_cons	1.469008	2.033738	0.72	0.470	-2.517045	5.455061

Average marginal effects
Model VCE: Robust

Number of obs = 88

```
Expression: Pr(welfare_gain), predict()
dy/dx wrt: 1.female age 1.black 1.business 1.noreligion bfi_extra bfi_agree bfi_cons bfi_neur
bfi_open 1.optimistP 1.earnP 1.eighty_tokens
```

		Delta-method				
	dy/dx	std. err.	z	P> z	[95% conf. interval]	
female						
Female	.0735548	.0881184	0.83	0.404	-.099154	.2462636
age	-.0044415	.0300776	-0.15	0.883	-.0633926	.0545096
1.black	.178667	.0862419	2.07	0.038	.009636	.3476981
1.business	-.0014641	.0959314	-0.02	0.988	-.1894862	.186558
1.noreligion	.145903	.0867272	1.68	0.093	-.0240791	.3158852
bfi_extra	-.0109068	.0549687	-0.20	0.843	-.1186434	.0968298
bfi_agree	.053306	.0758586	0.70	0.482	-.0953742	.2019862
bfi_cons	-.0298994	.0708303	-0.42	0.673	-.1687243	.1089255
bfi_neur	.0177015	.0706158	0.25	0.802	-.120703	.1561059
bfi_open	-.0452884	.0814813	-0.56	0.578	-.2049888	.1144119
1.optimistP	-.1792824	.0990404	-1.81	0.070	-.3733981	.0148332
1.earnP	-.2868691	.0808513	-3.55	0.000	-.4453348	-.1284035
1.eighty_tokens	.012591	.088551	0.14	0.887	-.1609657	.1861477

Note: dy/dx for factor levels is the discrete change from the base level.

. * comparison to EUT

```
. probit welfare_gain `demog_i' `bfi' i.optimistP i.earnP i.earnP i.optimistP#i.earnP
i.female#i.eighty_tokens if rdu==0 & submit==1, vce(cluster sid)
```

Probit regression

Number of obs = 88

Wald chi2(15) = 23.31

Prob > chi2 = 0.0778

Pseudo R2 = 0.1933

Log pseudolikelihood = -42.367424

(Std. err. adjusted for 88 clusters in sid)

		Robust				
welfare_gain	Coefficient	std. err.	z	P> z	[95% conf. interval]	
female						
Female	.6705979	.5071849	1.32	0.186	-.3234662	1.664662
age	-.0326749	.1073441	-0.30	0.761	-.2430656	.1777157
1.black	.4054507	.3453843	1.17	0.240	-.2714902	1.082392
1.business	.3919002	.3702614	1.06	0.290	-.3337989	1.117599
1.noreligion	.6433324	.3776918	1.70	0.089	-.09693	1.383595
bfi_extra	.0172992	.2023596	0.09	0.932	-.3793182	.4139167
bfi_agree	.1985773	.2844121	0.70	0.485	-.3588601	.7560147
bfi_cons	-.2067891	.2751567	-0.75	0.452	-.7460864	.3325083
bfi_neur	.0944357	.2749979	0.34	0.731	-.4445502	.6334217
bfi_open	.1360013	.3325084	0.41	0.683	-.5157031	.7877058
1.optimistP	-.894054	.7503942	-1.19	0.233	-2.3648	.5766915
1.earnP	-1.505667	.513225	-2.93	0.003	-2.51157	-.4997648
optimistP#earnP						
1 1	.2464838	.8354424	0.30	0.768	-1.390953	1.883921
female#eighty_tokens						
Male#1	.0046846	.4043314	0.01	0.991	-.7877905	.7971597
Female#1	-.1707547	.5486411	-0.31	0.756	-1.246071	.9045621
_cons	.7091177	2.148426	0.33	0.741	-3.501719	4.919955

Average marginal effects

Number of obs = 88

Model VCE: Robust

Expression: Pr(welfare_gain), predict()

dy/dx wrt: 1.female age 1.black 1.business 1.noreligion bfi_extra bfi_agree bfi_cons bfi_neur
bfi_open 1.optimistP 1.earnP 1.eighty_tokens

		Delta-method				
	dy/dx	std. err.	z	P> z	[95% conf. interval]	
female						
Female	.1484242	.0877143	1.69	0.091	-.0234928	.3203412
age	-.0088015	.028891	-0.30	0.761	-.0654268	.0478238
1.black	.1087823	.0903986	1.20	0.229	-.0683957	.2859602
1.business	.1004807	.0891177	1.13	0.260	-.0741869	.2751482
1.noreligion	.1622946	.0872815	1.86	0.063	-.008774	.3333633
bfi_extra	.0046598	.0545352	0.09	0.932	-.1022272	.1115468
bfi_agree	.0534898	.0760134	0.70	0.482	-.0954938	.2024734
bfi_cons	-.0557018	.073797	-0.75	0.450	-.2003413	.0889377
bfi_neur	.0254377	.0737507	0.34	0.730	-.1191111	.1699865
bfi_open	.036634	.0892619	0.41	0.682	-.138316	.2115841
1.optimistP	-.1975527	.0980081	-2.02	0.044	-.389645	-.0054604
1.earnP	-.3203305	.0780357	-4.10	0.000	-.4732778	-.1673833
1.eighty_tokens	-.0170335	.0921194	-0.18	0.853	-.1975842	.1635172

Note: dy/dx for factor levels is the discrete change from the base level.

. * Table 2 risk parameters

. bysort female: summ r_eut r_rdu eta phi if record==1 & female ~= .

-> female = 0

Variable	Obs	Mean	Std. dev.	Min	Max
r_eut	50	.5960055	.1778426	.2659489	.8953205
r_rdu	50	.7075162	.0379635	.6257992	.8038139
eta	50	.9709536	.5589444	.3421361	2.794703
phi	50	.9975626	.4088162	.3699581	2.381267

-> female = 1

Variable	Obs	Mean	Std. dev.	Min	Max
r_eut	38	.562654	.1637901	.1287987	.8315247
r_rdu	38	.6988082	.0329121	.626736	.7685972
eta	38	.9009178	.4462629	.2821337	1.803559
phi	38	.9193542	.3137058	.4147758	1.611156

. * Table 2 payouts and probabilities

For the 80 token case, the reported probability of winning the Tournament under Stress for men was 0.26, and was 0.28 for women.
For the 80 token case, the reported probability of winning the Tournament with no Stress for men was 0.26, and was 0.23 for women.
For the 80 token case, the Piece Rate payout for men was \$1.37, and was \$1.36 for women.
For the 80 token case, the Tournament payout for men, conditional on winning, was \$5.66, and was \$5.72 for women.

. * Used for the Table 2 welfare calculations

For the 1 token case, the reported probability of winning the Tournament under Stress for men was 0.24, and was 0.23 for women.
For the 1 token case, the reported probability of winning the Tournament with no Stress for men was 0.24, and was 0.27 for women.
For the 1 token case, the Piece Rate payout for men was \$1.04, and was \$1.17 for women.
For the 1 token case, the Tournament payout for men, conditional on winning, was \$3.87, and was \$4.23 for women.

4. Evaluating the Costs of Competition Interventions

As noted in the text, Niederle, Segal and Vesterlund [2013] and Balafoutas and Sutter [2012] consider a variety of “affirmative action” policies to reduce the gender gap in 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. Here we review the methods used in each of these studies to evaluate the cost of the behavior they observe.

We start in reverse order, since **Alan and Ertac [2019]** provides the cleanest example. They evaluate Efficiency, as they refer to it in the title of §5.4, as follows:

Based on performance, for some children it is payoff-maximizing to compete, whereas for others it is better to stay out. One can be concerned that interventions such as the one we evaluate in this paper may lead to unintended inferior outcomes for some children, by inducing decisions that turn out to be bad for payoffs ex post. Analyzing how children’s decisions fare in terms of expected material payoffs can

shed light on these issues. To do this, we first simulate the probabilities of winning and tying in competition for any given performance level, using the empirical distribution of performances. For the first stage, we use the actual performance of the whole sample when calculating the empirical win probabilities. That is, we use simulations to form random matches from the whole empirical distribution in the first choice stage, and compute the empirical win, loss and tie probabilities that correspond to each realized performance level. [p.1168]

The need to simulate probabilities is simply due to their design not including belief elicitation stages, understandable in the context of a vast field experiments with children at school. In their second stage subjects were told their own performance, so the efficiency analysis was modified by calculating empirical win probabilities compared to others that had strictly different scores.

The next step was to infer the cost of selecting or not selecting the profit-maximizing choice:

Using these probabilities along with realized performances, we calculate each child's expected payoff from competition, and analyze whether the child's actual choice was payoff maximizing ex post. We then estimate the treatment effects on this outcome to see whether the treatment leads to suboptimal choices from a payoff maximization perspective. It should be noted that this analysis does not make utility comparisons and therefore it is not an optimality analysis per se, as it disregards effort costs, which are unobservable. [p.1169]

The final sentence is problematic. It is correct in stating that the analysis does not make utility comparisons, but implies that the sole purpose of using utility functions is to evaluate the difference between expected benefits of competing and expected costs of competing, where costs are solely due to cognitive effort levels needed. But it ignores the role of the utility function (and probability weighting function, if using RDU) in evaluating the risk premium entailed in making the choice to compete in the tournament. As Table D1 and Table 2 illustrate, one can immediately see the importance of allowing for a risk premium even if there is some reasonable guess at levels of risk aversion. Ideally one would like estimates of actual risk aversion, as we illustrate, but the very design of the task, through the relative payoffs of the piece rate compensation scheme and the tournament compensation scheme, assumes risk neutrality as a benchmark, alerting one to the need to consider aversion to risk.

One attractive feature of the evaluation of Alan and Ertac [2019; Figure 4, p.1171] is attention to the full distributional effects on all girls and all boys. The vertical axes show treatment effects in monetary units, and the horizontal axes display quantiles (presumably from the baseline):

One cannot just evaluate welfare, even if assuming risk neutrality, but just seeing if the “best performing” girls did better. The left panel shows that all girls appear to have done statistically significantly better from the treatment in terms of expected payoffs.

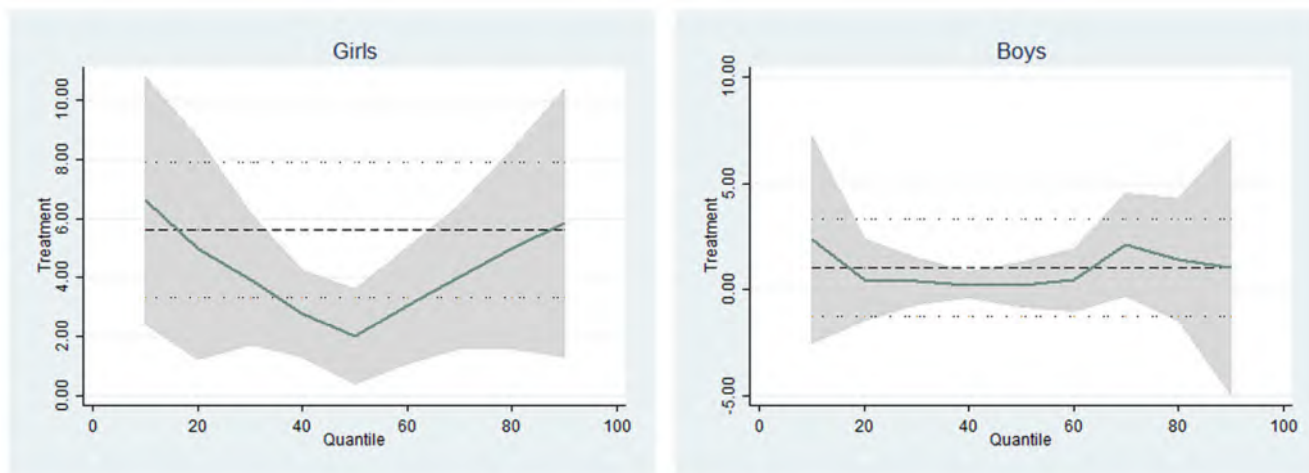


FIGURE 4. Distribution of treatment effects on expected payoff gains. Figures plot the quantile treatment effects on expected payoff gains and 95% confidence bands. The dashed lines indicate the estimated average treatment effects with the dotted lines showing the confidence bands for the average treatment effect estimates.

However, their results for boys need to be evaluated carefully. They make the following claim (p.1171/2) from these displays: “At first glance, the figures show that treatment effects are generally non-negative across quantiles, for both genders, reassuring us that no one was hurt by the treatment in terms of payoffs.” This is not what one should infer from the right panel for boys. Note the 0-gain line for boys on the vertical axis, just below the dashed line showing the average treatment effect. The fact that the 95% confidence band falls *below* the 0-gain line *for all quantiles* says, literally, that some fraction of the sample suffered losses in terms of the expected payoffs effect of the treatment. The confusion arises from an intended inference about the average effect rather than an inference about the sample as a whole. That is, one might infer that there is no evidence of a significant loss at any quantile *for the average boy at each quantile*, but that is not what is being claimed here. With very large samples of 1281 boys across all stages, these confidence intervals paint a clear picture of some losses in for boys in terms of expected payoffs.

We completely set aside a utilitarian social welfare function which would add up expected gains and losses and declare a welfare improvement if the sum (and hence average) is greater than zero. At least initially, the welfare analysis should, as suggested, look at the whole distribution.

The focus of **Balafoutas and Sutter [2012]** is immediately on efficiency, from the title itself, “Affirmative Action Policies Promote Women *and Do Not Harm Efficiency* in the Laboratory” (emphasis added). Their design is a replication of Niederle, Segal and Vesterlund [2013], with extensions to consider alternative versions of Affirmative Action policies. They motivate their concern with efficiency as follows:

Policy interventions to support the promotion of women often face the criticism that they are inefficient in [not] assigning the best available candidate, irrespective of gender, to a particular job when several candidates compete for it [...]. While this is

difficult to measure in the field because it is hard to exactly identify a candidate's qualifications, laboratory-based economic experiments allow for an unambiguous assessment of the efficiency of affirmative action programs in promoting the best candidates, although this is measured in the artificial context of a well-controlled, quantitative task. [p.579].

To make the meaning explicit, we add “[not]” to the opening sentence of this text. Furthermore,

Interventions that promote the entry of women may have two opposing effects on the overall efficiency in selecting the best candidates as winners. On the one hand, any intervention that gives an advantage to women may yield efficiency losses by passing over better-performing men for the sake of promoting women. On the other hand, interventions may induce more high-performing women to choose competition instead of going for the piece rate, leading to efficiency gains. [p.580]

This is defining efficiency solely in terms of production efficiency, measuring whether output is being produced at least cost by the appropriate mix of factors of production. This is one necessary condition for the economy to be efficient, but not sufficient. In particular, it ignores consumption efficiency, measuring whether a given bundle of goods ends up with the individuals that value it the most.

It is easy to see how the outcomes evaluated here could violate consumption efficiency. Imagine that the most productive women were the most risk averse, and place a relatively large premium on avoiding the risk associated with competing in the tournament. *Ceteris paribus*, it could easily be efficient for less productive women, or even less productive men, to take on the tournament competition. Or, more to the point from Table D1 and Table 2, it could easily be efficient *for none of the production to occur with tournament compensation*.

The details of the calculations of production efficiency gains for men and women are not relevant, since they fail to consider consumption efficiency.

Finally, **Niederle, Segal and Vesterlund [2013]** developed the variant of the Niederle and Vesterlund [2007] design to study one implementation of affirmative action policies to encourage women to choose to compete. Again, their title immediately focusses on issues of cost: “How Costly Is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness”. They ask

... whether affirmative action can encourage more high-performing women to enter competitions. We focus not only on determining how the policy changes the decision to compete, but also on how it changes the gender composition of the pool of competitors. Accounting for changes in entry, we ask how costly it is to secure that women be equally represented among those who win competitions. How much lower will the performance be for winners under the policy? How many better performing men will have to be passed by to secure equal representation of those hired? [p.2]

So the focus is again solely on production efficiency. With some simplification, they find that better-performing women tend to displace poorer-performing women in choosing to compete in

the tournament compensation scheme. Hence there tends to be less displacement of better performing men than one might have expected, *a priori*.

An important subtlety is addressed when evaluating “How Costly is Affirmative Action”, the title of §4:

A concern when introducing affirmative action is how costly it will be to achieve a more diverse set of winners. Looking directly at the performance of those who enter and win the tournaments, we do not see deterioration in performance. [...] This difference is not significant [...] and suggests that, in contrast to expectations affirmative action need not decrease performance. Although it is tempting to focus solely on the performance of actual winners, it is important to note that whether an individual is or is not identified as a winner depends on the performance of the group she is randomly assigned to. Thus, examination of actual winners only provides limited insight on the effect of affirmative action. [p.7]

The general way in which they propose to evaluate this point is elegant:

To assess the cost of the policy we view all the participants in the experiment as (potential) candidates and those who enter competitions as applicants for jobs. We ask what the minimum performance requirement would be if a firm wanted to hire a certain number of applicants and wanted to secure that only the best available applicants were hired. We then ask how much lower the requirement has to be if we want to hire the same number of applicants under the equal representation rule. To evaluate the degree of reverse discrimination, we determine how many strictly better performing men will be passed by to secure that women are at least equally represented among those hired. Passing by better performing applicants is inequitable and costly for the firm, as it no longer can hire the best available applicants. Crucial for assessing these two adverse effects is the performance and gender composition of those who decide to enter the competition. [p.7]

The same general point has been referred to in a related setting of gender discrimination, wage determination in a university, as “efficient equity,” the idea that one can search over alternative ways to reach some given equity goal and select the one that does so at least cost: see Coller, Harrison and Rutherford [1998]. It is also used in trade policy evaluation, to facilitate liberalization while ensuring that certain equity goals are met at least cost: see Harrison, Rutherford and Tarr [2003].

Again, as with Balafoutas and Sutter [2012], the details of the calculations of production efficiency gains for men and women are not relevant, since they fail to address consumption efficiency.

Additional References

Coller, Maribeth; Harrison, Glenn W., and Rutherford, Thomas F., "Efficient Equity: Removing Salary Discrimination By Meeting Statistical Legal Constraints at Least Cost," *Economics Letters*, 52, 1996, 81-88.

Harrison, Glenn W.; Rutherford, Thomas F., and Tarr, David G., "Trade Liberalization, Poverty and Efficient Equity," *Journal of Development Economics*, 71, June 2003, 97-128.

Appendix D: Pilot Results with Variants on the Raven Task (Online Working Paper)

1. Time Constraints

One popular variant of the Raven task involves the *imposition of time constraints* on completing the full task. The most common time constraint is 40 minutes, and we modify the Progressive Eighty Tokens condition to examine the effect of that constraint.⁶⁹ Of course, our default experiment formally had a time constraint of 90 minutes, but that was not particularly binding, as illustrated by the solid black line in Figure D1. The average time required was 48 minutes, and most subjects were completed after about an hour. A constraint of 40 minutes happens to be at the 40th percentile of this distribution, so is expected *a priori* to be binding for some individuals.⁷⁰

As mentioned earlier, intelligence measures with binding time constraints are generally viewed as measuring intellectual efficiency rather than some “pure intelligence.”⁷¹ Since we view it as natural to think of cognition and intelligence as contextual, and time is clearly something with an opportunity cost, the manner in which individuals apply their cognitive production functions under varying constraints is of great practical interest to economists.⁷² This pilot experiment also allows us to see the pure effect of a time constraint, which plays a key role in the “gender and competitiveness” designs we examine in §3.

When we impose a 40 minute time constraint in the time-constrained condition we find no evidence of a change in average Accuracy or average Efficiency. The dashed line in Figure D1 shows the re-allocation of time spent on the task, with a sharp peak just immediately prior to the time constraint. Figure D2 displays the effect of the time constraint in terms of Accuracy and Efficiency. The average marginal effect of the time constraint on Accuracy is -4.0 percentage points, but not statistically significant (p -value = 0.20), and the effect on Efficiency is -0.7 percentage points and also not statistically significant (p -value = 0.78). The effect of demographics on the time constraint and Efficiency are shown in Figure D3. The time constraint does significantly lower Accuracy and Efficiency for women, by 8.4 and 4.1 percentage points (p -values

⁶⁹ Raven, Raven and Court [1998] present norms for 30-minute and 40-minute timed versions of the RAPM. Hamel and Schmittmann [2006] evaluated a 20-minute timed version of the test with 397 pseudo-volunteers (viz., psychology students taking the test as a course requirement) against an untimed version with 59 subjects drawn from the same population. They report a correlation of 0.75 between the scores of the untimed and timed versions.

⁷⁰ A 20-minute constraint is at the 3rd percentile of this distribution, and a 30-minute constraint at the 20th percentile.

⁷¹ For example, by Raven, Raven and Court [1993].

⁷² Borghans, Meijers and Ter Weel [2008] find a striking interaction between the effect of financial incentives and time constraints in an experiment in which university students were presented with a battery of cognitive tests. Extra financial incentives led to substantially more *time* being spent on answering those questions. But whether or not that extra time effort generated better scores depended on time constraints, amongst other things. When subjects were time constrained, there was less scope for the extra effort to be rewarded with better performance, and hence less scope for financial incentives to show an effect on performance. In addition, the link between greater effort and greater performance was mediated by non-cognitive factors, such as personality traits, and the specific set of cognitive questions considered. Only two of the ten cognitive questions were “Raven-like,” and it is inappropriate to view this set of 10 as a test of the effect of incentives and time constraints on fluid intelligence (let alone IQ, as they claim). But the general results are, nonetheless, important.

of 0.013 and 0.084) respectively. So even if the average effect of this time constraint are not statistically significant at conventional levels, there are significant distributional effects correlated with one important demographic for policy.

2. Progressively Increasing Incentives

Apart from imposing a binding time constraint to performance, another question for economists to ask is whether Accuracy and/or Efficiency would improve if *financial incentives were increased for the harder questions*.⁷³ We therefore increased the earnings for allocating all 80 tokens to the correct solution from the default \$2 for all questions, to be \$2 for questions 1 through 12, \$3 for questions 13 through 24, and \$5 for questions 25 through 36. This increased potential aggregate earnings from \$72 up to \$120.

We observe a marked *improvement* in Accuracy *and* Efficiency when financial incentives within the RAPM are progressively increased. Figure D4 displays the effect of the increasing financial incentives in terms of Accuracy and Efficiency. The average marginal effect of increasing incentives on average Accuracy is +11.8 percentage points (p -value = 0.002), and the effect on average Efficiency is +7.3 percentage points (p -value = 0.013). Figure D5 shows effects of the increasing financial incentives for certain demographics, There are significant improvements in Efficiency for Blacks (+6.2 pp, p -value = 0.03) and Business majors (+16.9 pp, p -value = 0.01), accompanied by comparably significant improvements in Accuracy. For women there are clear improvements in Accuracy of +7.5 pp (p -value = 0.04), and *some* evidence of improvements in Efficiency of +4.7 pp (p -value = 0.11).

Additional References

Hamel, Ronald, and Schmittmann, Verena, D., “The 20-Minute Version as a Predictor of the Raven Advanced Progressive Matrices Test,” *Educational and Psychological Measurement*, 66(6), December 2006, 1039-1046.

⁷³ McDaniel and Rutström [2001] considered a similar *between-subjects* treatment in an incentivized task in which subjects attempted to solve the Tower of Hanoi puzzle in the least number of moves. They gave 25 subjects “low” rewards for a better score and 30 subjects “high” rewards for a better score, and found (p.153) no statistically significant difference in performance. Consistent with some of the literature on RAPM, such as Gignac [2018], they did report an increase in “effort.” Our treatment is *within-subject*.

Figure D1: Distribution of Time Taken

Time taken to complete the Progressive 80 Tokens task with a 90-minute constraint or a 40-minute constraint

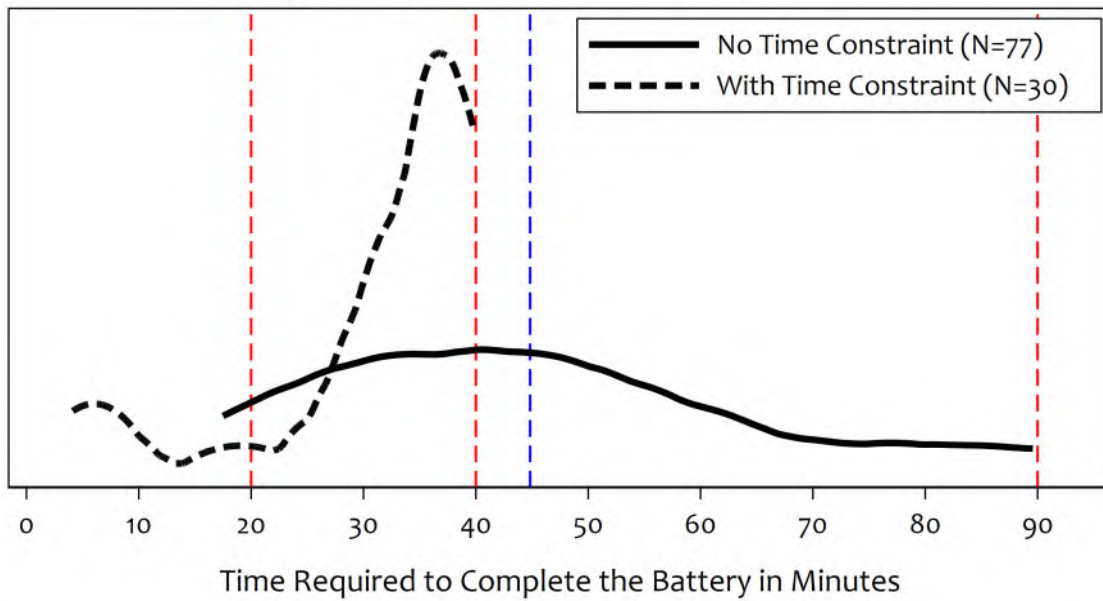


Figure D2: Effect of 40-Minute Time Constraint on Accuracy and Efficiency

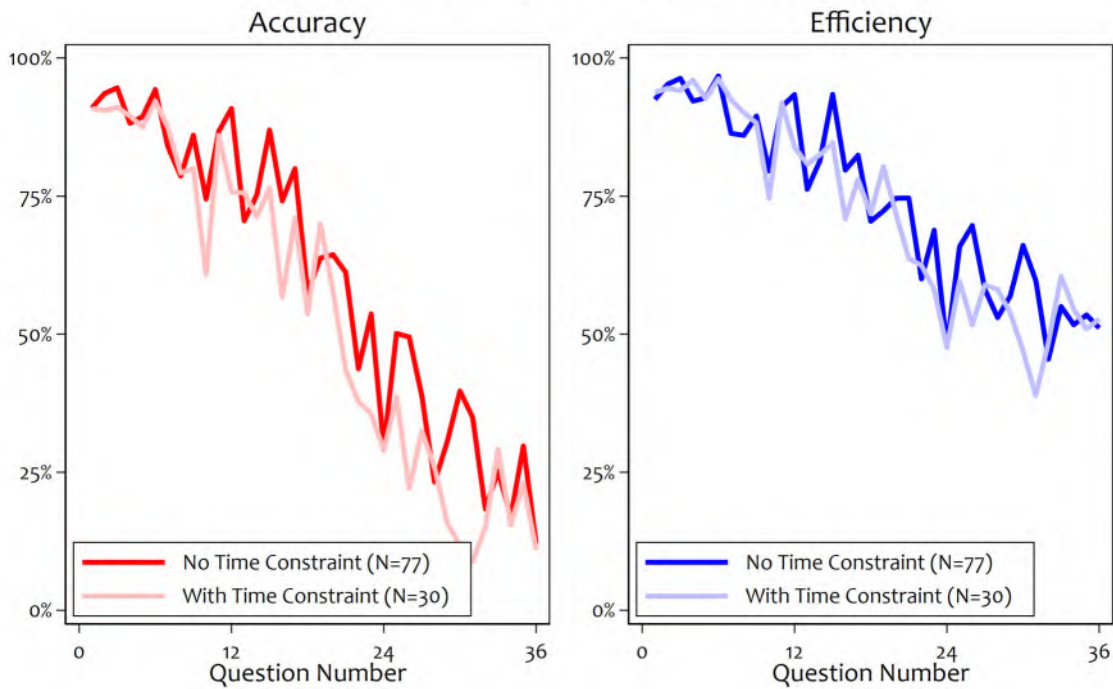


Figure D3: Effects of Time Constraint on Efficiency

Average marginal effects of 40-minute time constraint from Fractional Regression
Solely comparing results within the Progressive Treatment
subjects with a time-constraint, and 77 unconstrained

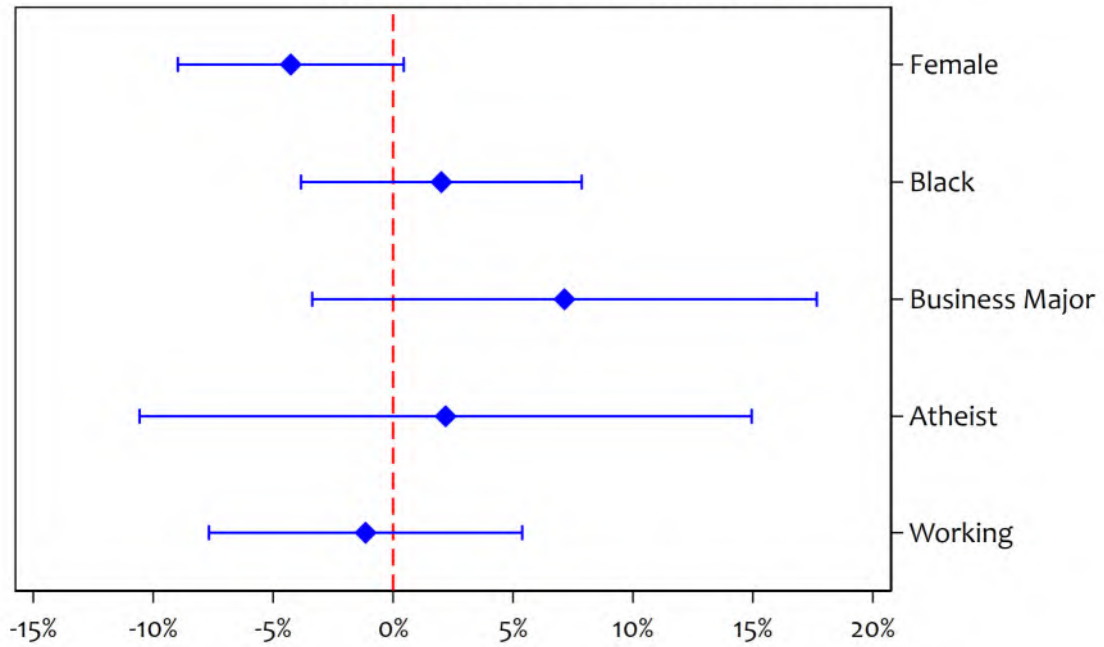


Figure D4: Effect of Increasing Financial Incentives
on Accuracy and Efficiency

Solely comparing results within the Progressive Treatment

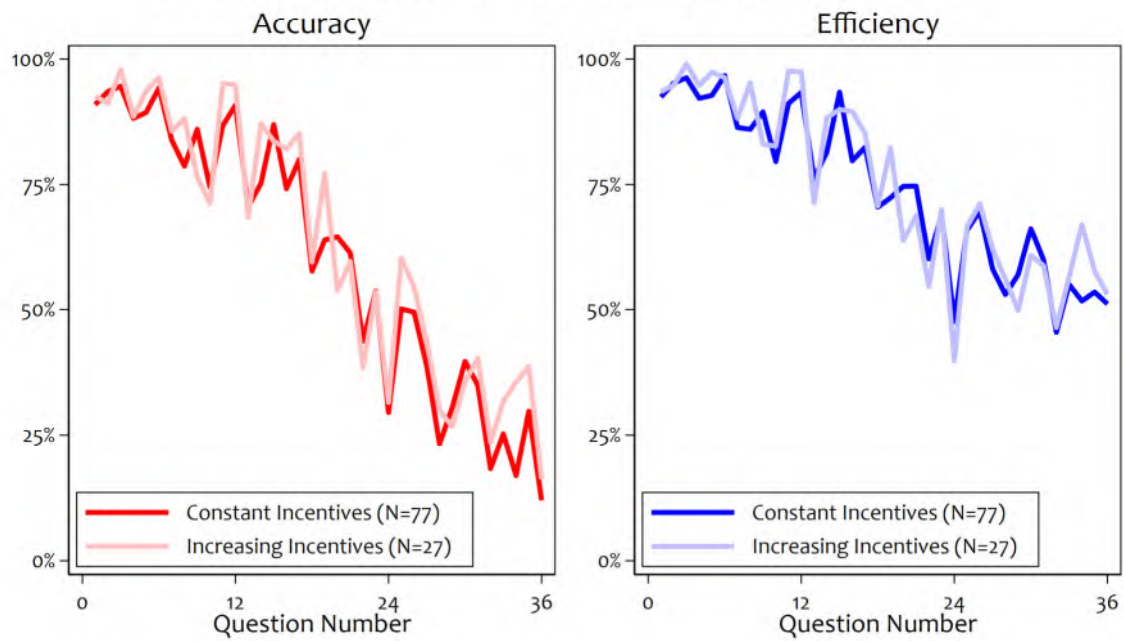
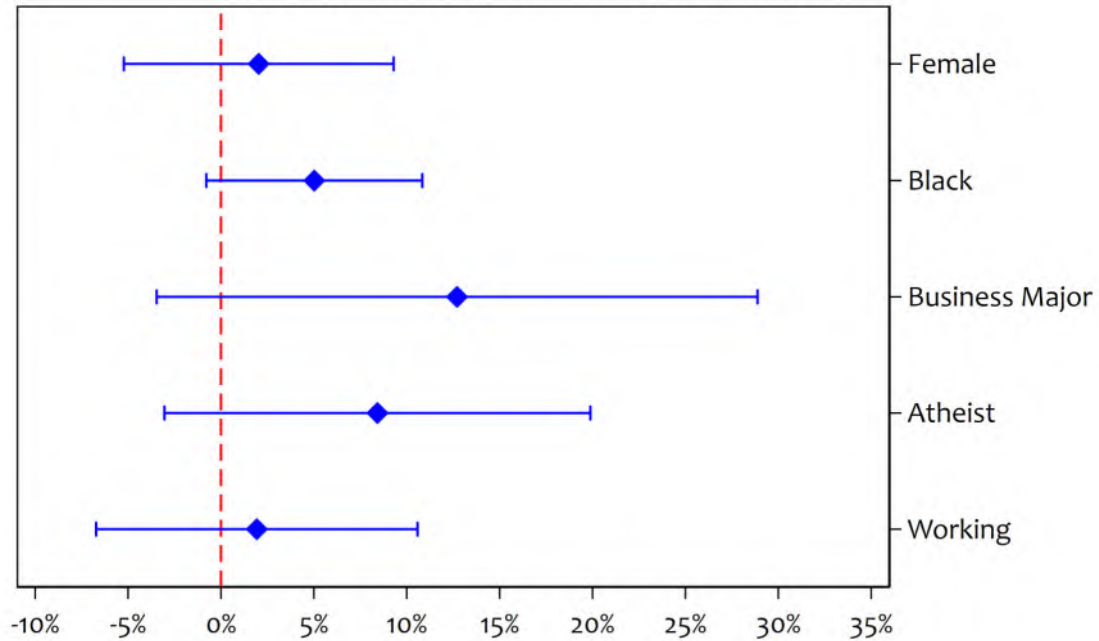


Figure D5: Effects of Increasing Incentives on Efficiency

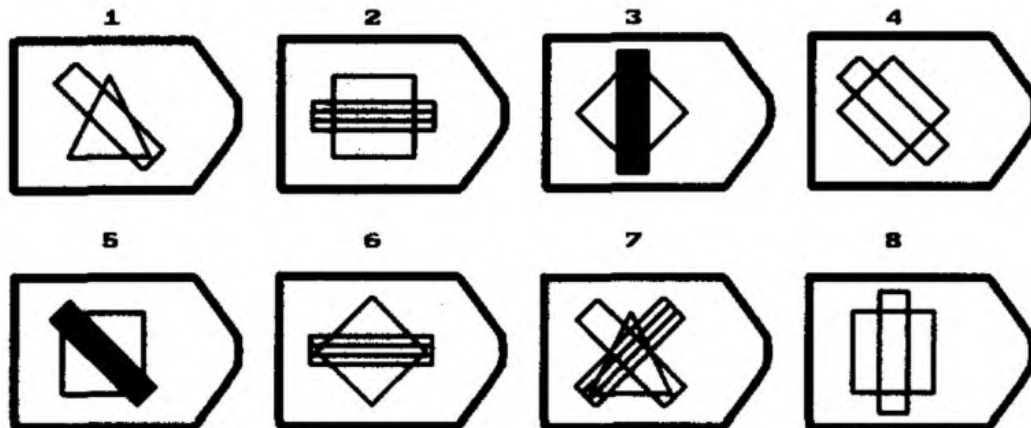
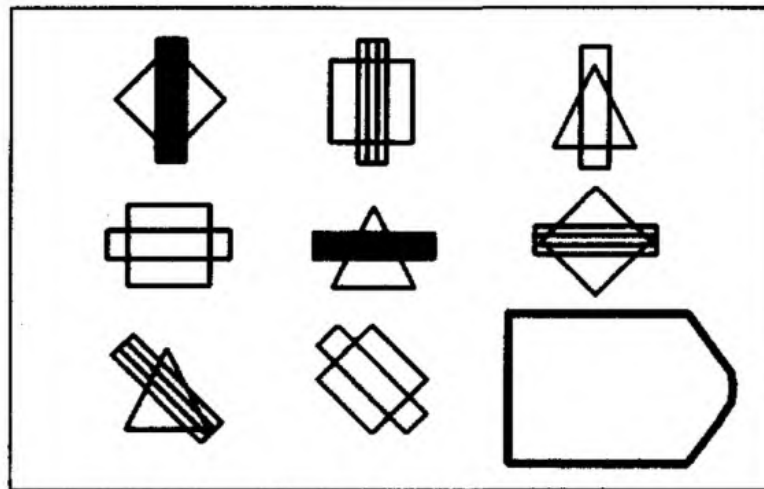
Average marginal effects of progressively increasing incentives from Fractional Regression
Solely comparing results within the Progressive Treatment
27 subjects with increasing incentives, and 77 with constant incentives



Your Instructions

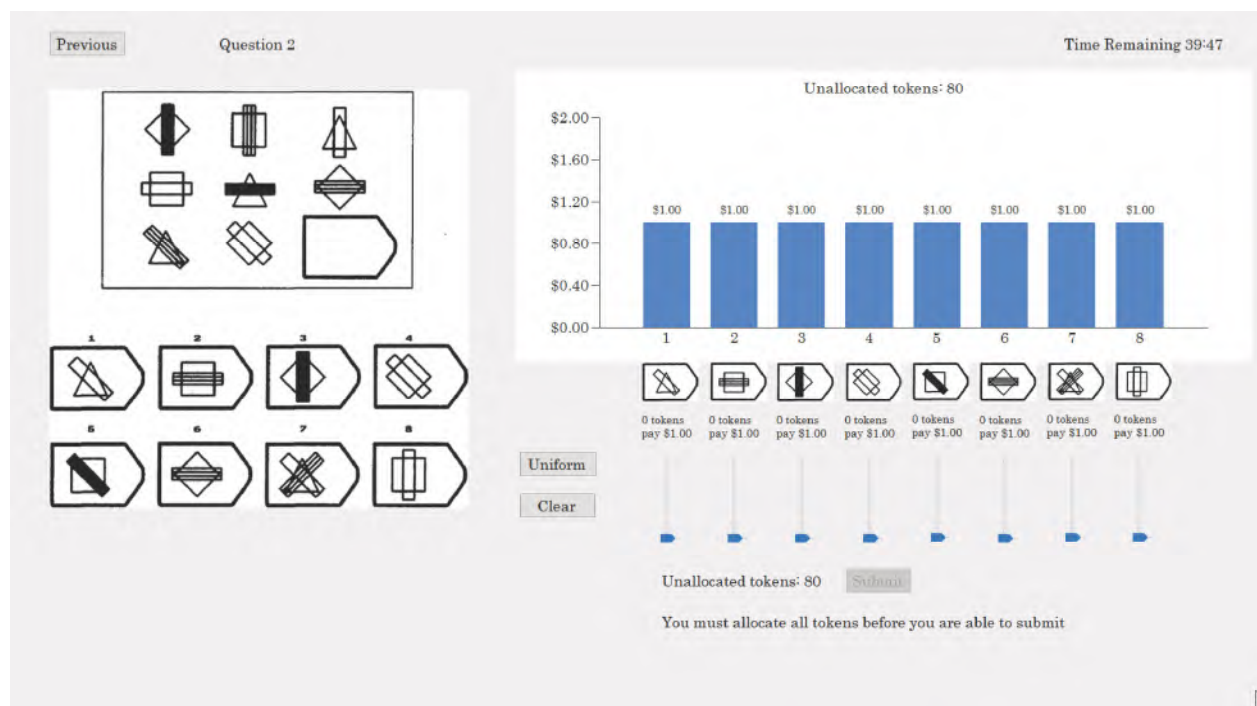
This task is a test of perception and clear-thinking. You have already completed the first part of the task, in a previous session, when you were given 12 similar problems. We will now consider a fresh set of 36 problems.

Consider a similar problem, shown here.

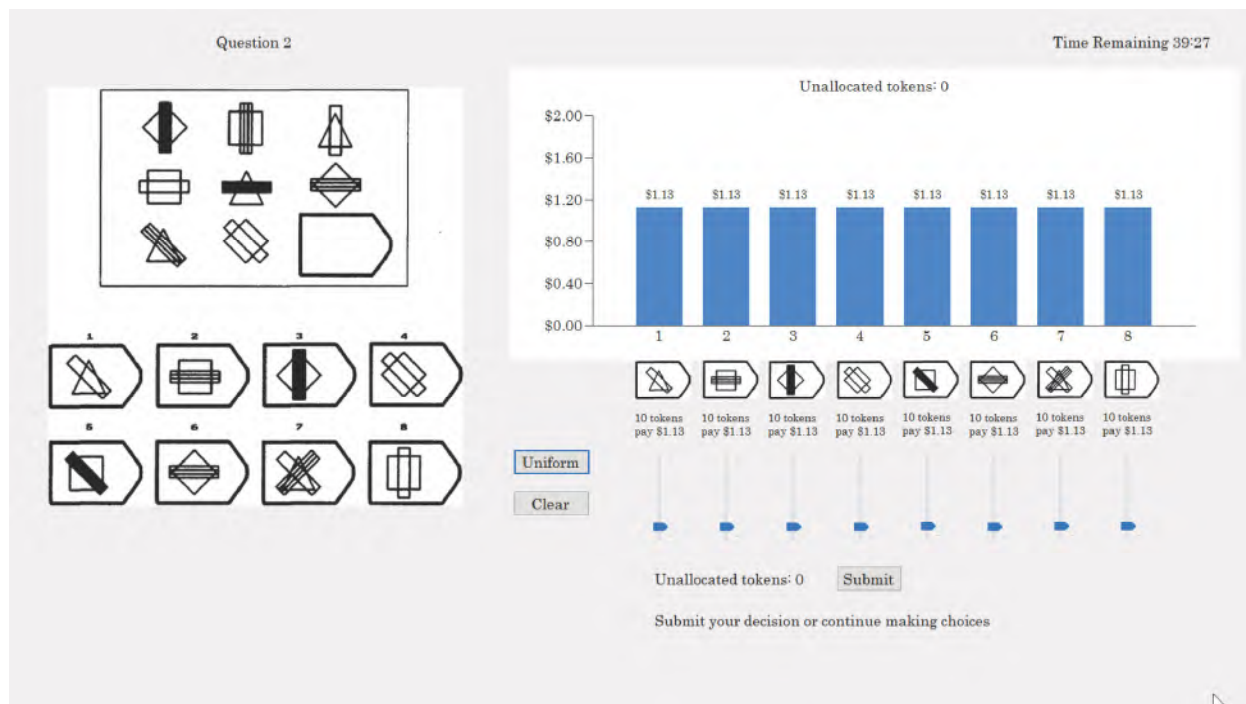


The top part of this problem is a pattern with a bit cut out of it. Look at the pattern, and think what the piece needed to complete the pattern correctly both across and down must be like. Then find the right piece out of the eight bits shown, and numbered, in the bottom part of this problem.

You will be asked to report your beliefs about the correct answer using an interface like this one, which is also generally familiar to you from a previous session. The version you will see on the computer will be larger and easier to read.



The problem and possible solutions are shown on the left of the screen, in the usual manner. On the right of the screen you have 80 tokens to allocate across the 8 possible answers. We start off with 0 tokens allocated to each of the possible answers. If you wanted to change this initial allocation so that there were 10 tokens allocated to each possible answer, just click on the **Uniform** button, and you will see this display:

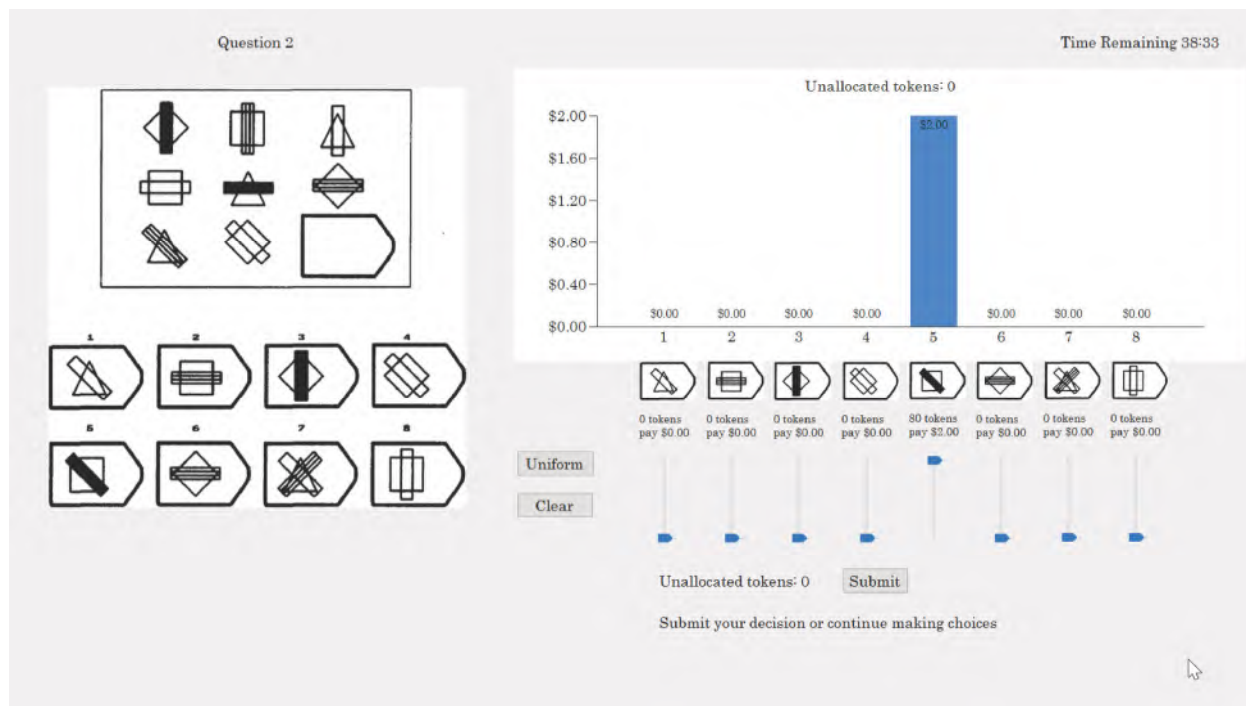


Here you have allocated 10 tokens to each possible answer, and you would earn \$1.13 if you reported this allocation of tokens, since only one of the 8 possible answers is correct. You can return to the initial allocation of 0 tokens for each possible answer by clicking on the **Clear** button.

As you allocate tokens, by moving the sliders up or down, the earnings will change above each bar. These are the earnings that you will receive for this problem if that bar refers to the correct answer to the problem. **You will be paid for all 36 problems, and each problem will pay between \$0 and \$2 depending on your answer.**

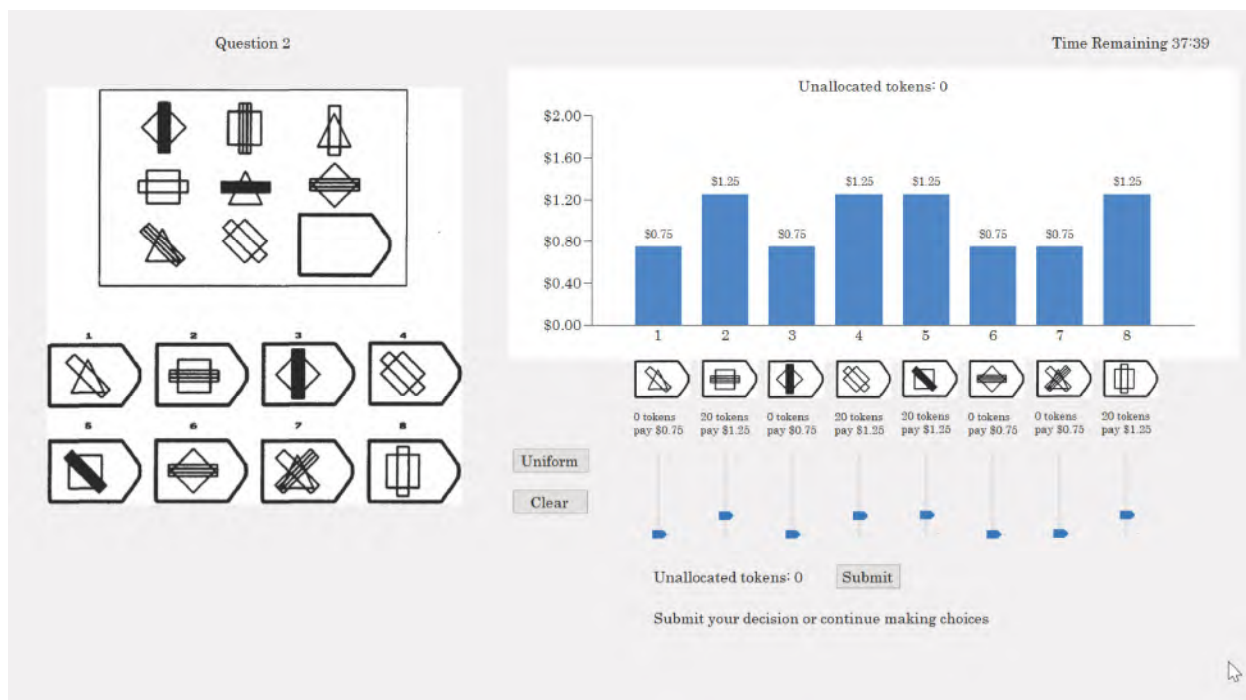
Return now to the problem itself. Only one of these pieces is perfectly correct. Pieces #2, #5 and #8 complete the problem correctly going down, in the sense that they have a square as the larger piece. Pieces #1, #4 and #5 are correct going across, in the sense that they have the right slope for the 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. In this case, if you were certain that piece #5 is the single best answer, you should allocate all 80 tokens to the bin representing piece #5 as in this display:



So in this case you would earn \$2.00 if indeed the correct answer was #5. Of course, if any of the other pieces turned out to be the correct answer you would, in this case, earn \$0.

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:



You can see that even if you eliminate some pieces, such as #1, #3, #6 and #7, that are clearly wrong, you give yourself a 1-in-4 chance of earning more money than if you guessed across all 8 pieces. In this case you would expect to earn \$1.25 if indeed one of pieces #2, #4, #5 or #8 had been correct. Recall that if you had allocated the tokens roughly equally across all 8 bars, thinking that any of the 8 pieces might be correct, you would only earn \$1.13 for this problem.

This illustrative problem is relatively easy, but some of the later problems will not be so easy, and it might be harder to identify the one correct answer. You would still benefit by ruling out certain pieces so that you can allocate more tokens to the ones that you have not ruled out.

You will find that the problems soon get difficult. Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. The important thing is to notice how the problems develop and to learn the method of solving them. In each problem you look across each row, or down each column, and decide what the missing figure is like. Then look for the answer that is right both ways, across **and** down, from the options you have to select from.

As you recall from allocating tokens in a similar task in a previous session, you can change your token allocation as many times as you like before you hit the **Submit** button and **Confirm** your token allocation. You can also review and modify your prior decisions. You will have navigation buttons in the top left corner of the screen. These navigation buttons are only available before you start moving the sliders for a problem. Once you move any sliders for a problem, the navigation buttons disappear and you must submit your answer and move to the next problem in order to see the navigation buttons again.

You will be allowed 40 minutes to complete this task. Your screen displays how many

minutes (and seconds) are remaining in the top, right corner. Remember it is accurate work that counts. Attempt each problem in turn. Do your best to eliminate the incorrect pieces and find the correct piece to complete it before going on to the next problem. You will have the option to go back to previous problems if you want to. After you have completed the last question you must confirm if you are finished. When you are finished you will be told your total earnings, and will not have the option to go back to any of the problems. If you run out of time, you will automatically be treated as having finished, and told your total earnings.

Where you position each slider depends on your beliefs about the correct answer to the question. Again, each bar shows the amount of money you could earn if the true outcome corresponds to the possible solution shown under the bar.

You will be **paid for each of the 36 problems**, so you should think carefully about each problem. Since you can earn up to \$2 for each problem, you could earn up to \$72 over all 36 problems. You will not earn anything on any problems for which you have not confirmed an allocation of tokens.

It is up to you to balance the strength of your personal beliefs with the possibility of them being wrong. There are several important points for you to keep in mind when making your decisions:

- First, your belief about the correct answer to each problem is your personal judgment.
- Second, you have up to 40 minutes to complete this task.
- Third, you will not earn anything on any problems for which you have not confirmed an allocation of tokens.
- Fourth, depending on your choices and the correct answer you can earn up to \$2 for each problem, or up to \$72 over all 36 problems.
- Finally, your choices might also depend on your willingness to take risks or to gamble.

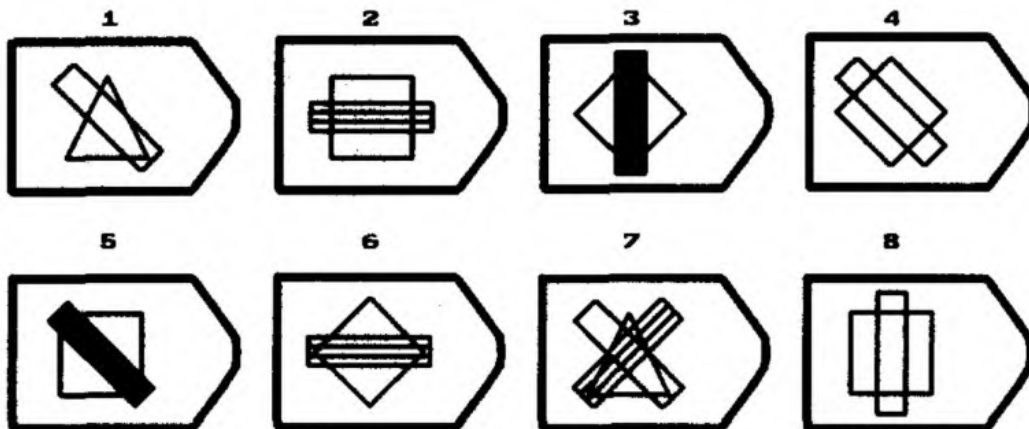
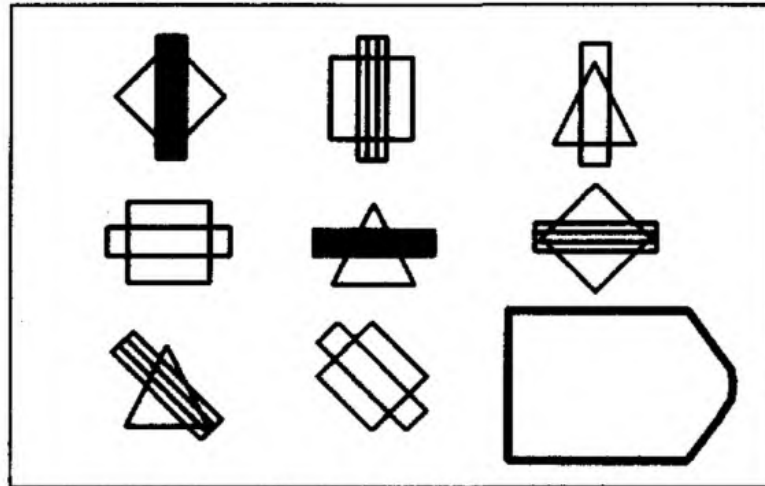
The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are finished you will see a summary of your responses and your total earnings from this task. Your earnings are in addition to the show-up payment you receive for participating.

Your Instructions

This task is a test of perception and clear-thinking. You have already completed the first part of the task, in a previous session, when you were given 12 similar problems. We will now consider a fresh set of 36 problems.

Consider a similar problem, shown here.



The top part of this problem is a pattern with a bit cut out of it. Look at the pattern, and think what the piece needed to complete the pattern correctly both across and down must be like. Then find the right piece out of the eight bits shown, and numbered, in the bottom part of this problem.

You will be asked to report your beliefs about the correct answer using an interface like this one, which is also generally familiar to you from a previous session. The version you will see on the computer will be larger and easier to read.

Previous

Question 2

Time Remaining 89:47

1

2

3

4

5

6

7

8

Uniform

Clear

Unallocated tokens: 80

1

2

3

4

5

6

7

8

0 tokens pay \$1.00

0 tokens pay \$1.00

0 tokens pay \$1.00

0 tokens pay \$1.00

0 tokens pay \$1.00

0 tokens pay \$1.00

0 tokens pay \$1.00

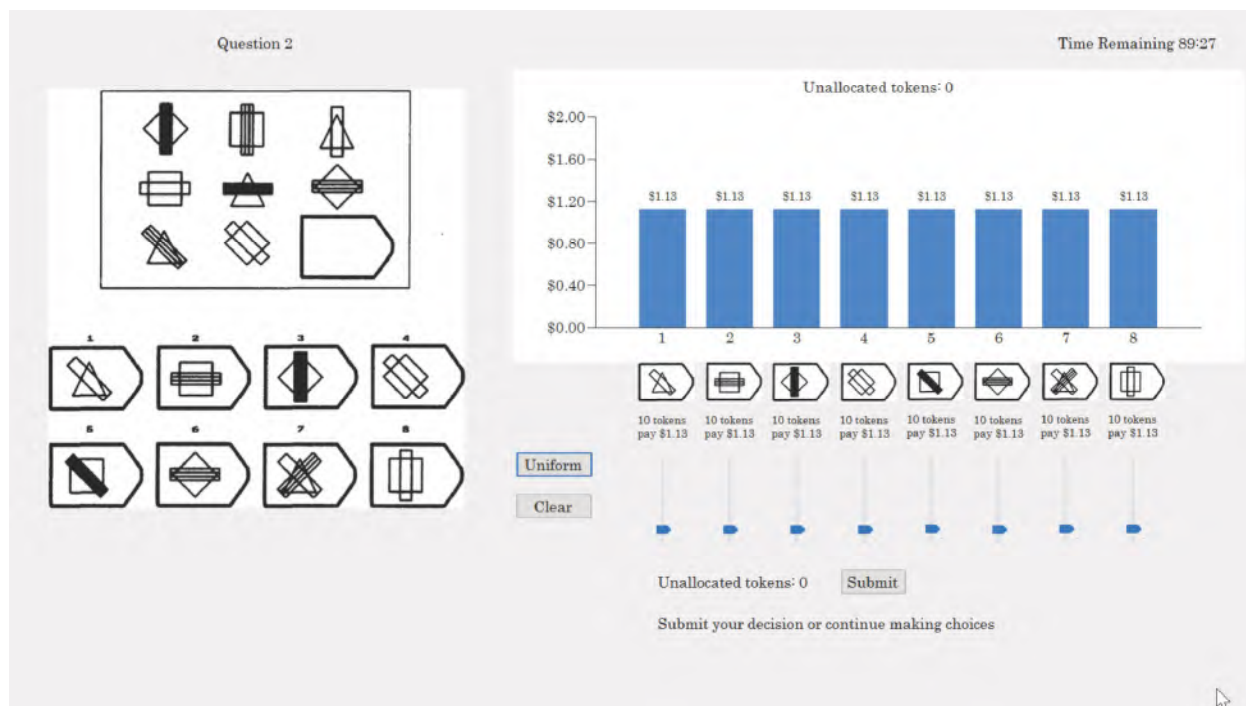
0 tokens pay \$1.00

Unallocated tokens: 80

Submit

You must allocate all tokens before you are able to submit

The problem and possible solutions are shown on the left of the screen, in the usual manner. On the right of the screen you have 80 tokens to allocate across the 8 possible answers. We start off with 0 tokens allocated to each of the possible answers. If you wanted to change this initial allocation so that there were 10 tokens allocated to each possible answer, just click on the **Uniform** button, and you will see this display:

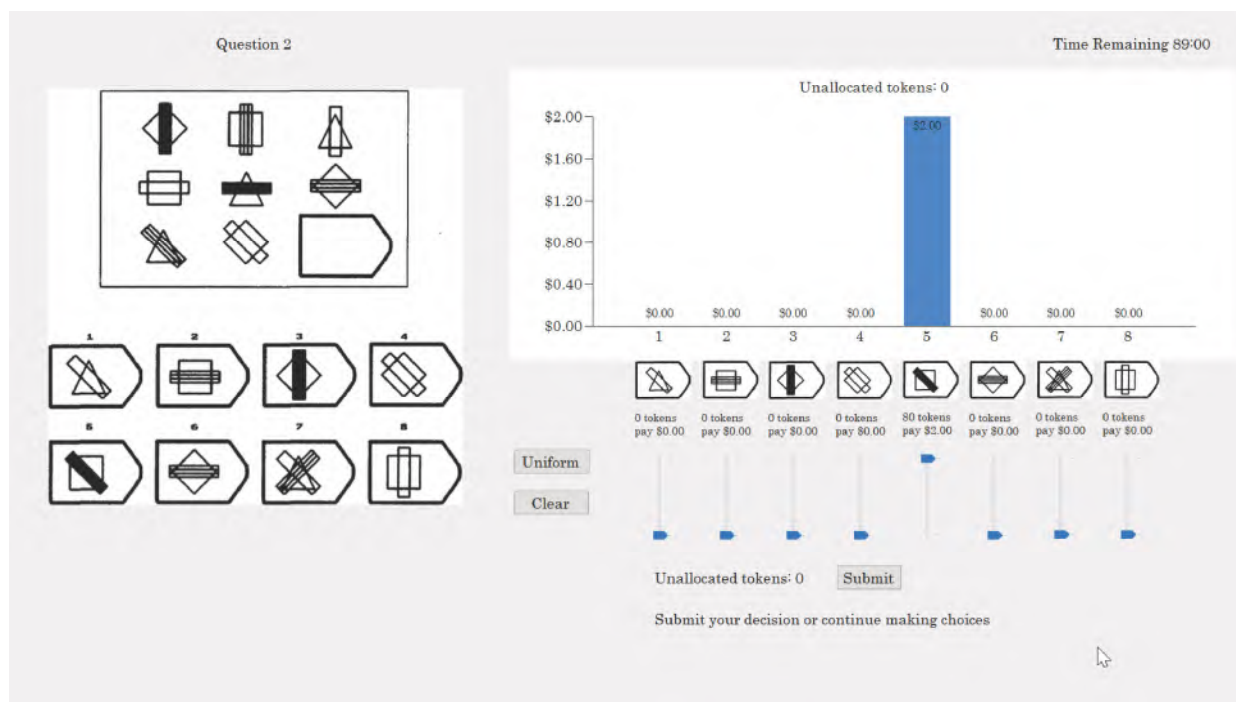


Here you have allocated 10 tokens to each possible answer, and you would earn \$1.13 if you reported this allocation of tokens, since only one of the 8 possible answers is correct. You can return to the initial allocation of 0 tokens for each possible answer by clicking on the **Clear** button.

As you allocate tokens, by moving the sliders up or down, the earnings will change above each bar. These are the earnings that you will receive for this problem if that bar refers to the correct answer to the problem. **You will be paid for all 36 problems. The first 12 problems will pay between \$0 and \$2 depending on your answer, the next 12 problems will pay between \$0 and \$3 depending on our answer, and the last 12 problems will pay between \$0 and \$5 depending on your answer.**

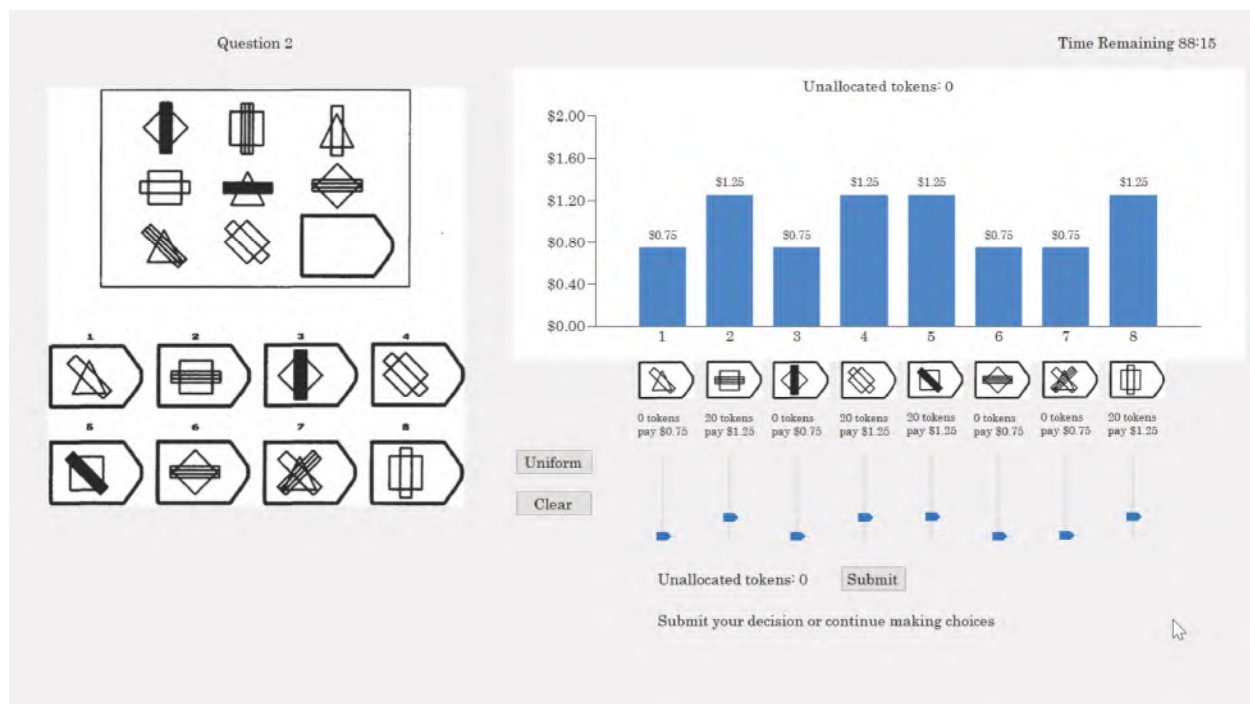
Return now to the problem itself. Only one of these pieces is perfectly correct. Pieces #2, #5 and #8 complete the problem correctly going down, in the sense that they have a square as the larger piece. Pieces #1, #4 and #5 are correct going across, in the sense that they have the right slope for the 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. In this case, if you were certain that piece #5 is the single best answer, you should allocate all 80 tokens to the bin representing piece #5 as in this display:



So in this case you would earn \$2.00 if indeed the correct answer was #5 and this was one of the first 12 problems. Of course, if any of the other pieces turned out to be the correct answer you would, in this case, earn \$0. If this was one of problems 13 through 24 you would earn \$3.00 if indeed the correct answer was #5. And if this was one of problems 25 through 36 you would earn \$5.00 if indeed the correct answer was #5.

Assume that this was one of the first 12 problems. 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:



You can see that even if you eliminate some pieces, such as #1, #3, #6 and #7, that are clearly wrong, you give yourself a 1-in-4 chance of earning more money than if you guessed across all 8 pieces. In this case you would expect to earn \$1.25 if indeed one of pieces #2, #4, #5 or #8 had been correct. Recall that if you had allocated the tokens roughly equally across all 8 bars, thinking that any of the 8 pieces might be correct, you would only earn \$1.13 for this problem.

This illustrative problem is relatively easy, but some of the later problems will not be so easy, and it might be harder to identify the one correct answer. You would still benefit by ruling out certain pieces so that you can allocate more tokens to the ones that you have not ruled out.

You will find that the problems soon get difficult. Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. The important thing is to notice how the problems develop and to learn the method of solving them. In each problem you look across each row, or down each column, and decide what the missing figure is like. Then look for the answer that is right both ways, across **and** down, from the options you have to select from.

As you recall from allocating tokens in a similar task in a previous session, you can change your token allocation as many times as you like before you hit the **Submit** button and **Confirm** your token allocation. You can also review and modify your prior decisions. You will have navigation buttons in the top left corner of the screen. These navigation buttons are only available before you start moving the sliders for a problem. Once you move any sliders for a problem, the navigation buttons disappear and you must submit your answer and move to the next problem in order to see the navigation buttons again.

You can work at your own speed, although we have to be out of the room in 90 minutes. Your screen displays how many minutes (and seconds) are remaining in the top, right corner. Remember it is accurate work that counts. Attempt each problem in turn. Do your best to eliminate the incorrect pieces and find the correct piece to complete it before going on to the next problem. You will have the option to go back to previous problems if you want to. After you have completed the last question you must confirm if you are finished. When you are finished you will be told your total earnings, and will not have the option to go back to any of the problems. If you run out of time, you will automatically be treated as having finished, and told your total earnings.

Where you position each slider depends on your beliefs about the correct answer to the question. Again, each bar shows the amount of money you could earn if the true outcome corresponds to the possible solution shown under the bar.

You will be **paid for each of the 36 problems**, so you should think carefully about each problem. Since you can earn up to \$2 for each of problems 1 through 12, \$3 for each of problems 13 through 24, and \$5 for each of problems 25 through 36, you could earn up to \$120 over all 36 problems. You will not earn anything on any problems for which you have not confirmed an allocation of tokens.

It is up to you to balance the strength of your personal beliefs with the possibility of them being wrong. There are several important points for you to keep in mind when making your decisions:

- First, your belief about the correct answer to each problem is your personal judgment.
- Second, you have up to 90 minutes to complete this task.
- Third, you will not earn anything on any problems for which you have not confirmed an allocation of tokens.
- Fourth, depending on your choices and the correct answer you can earn up to \$2, \$3 or \$5 for each problem, or up to \$120 over all 36 problems.
- Fifth, your potential earnings are higher as you progress through the problems.
- Finally, your choices might also depend on your willingness to take risks or to gamble.

The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are finished you will see a summary of your responses and your total earnings from this task. Your earnings are in addition to the show-up payment you receive for participating.