WORKING PAPER

September 2017 ■ RFF WP 17-18

Effects of Increases in IQ in India on the Present Value of Lifetime Earnings

A Structured Expert Judgment Study

Randall Lutter, Abigail Colson, and Roger Cooke

1616 P St. NW
Washington, DC 20036
202-328-5000 www.rff.org

RESOURCES FOR THE FUTURE

Effects of Increases in IQ in India on the Present Value of Lifetime Earnings: A Structured Expert Judgment Study

Randall Lutter, Abigail Colson, Roger Cooke

Abstract

Policymakers considering interventions that improve cognitive performance (IQ) need estimates of the value of IQ improvements. Using the classical model of structured expert judgment, we develop estimates of the percentage increase in earned income from interventions to increase IQ in India. Our estimates vary with age, for boys and girls, for urban areas and all of India, and also reflect expert's uncertainty. We combine these estimates with data on wages and labor force participation from the Indian Human Development Survey and use lifetables for India to calculate the expected discounted gain in lifetime earnings from a hypothetical gain of one IQ point (per capita), in both Indian Rupees and US dollars. We contrast our estimates with earlier estimates for the United States, including those used by regulatory agencies in the United States. Our results (which range from low tens to low hundreds of dollars per capita, depending on gender and discount rates), suggest that large scale interventions that are effective in raising cognitive performance in India would have large economic benefits.

Key Words: IQ, cognitive performance, lifetime earnings, India

JEL Classification Numbers: I15, J18, J24

Contents

Effects of Increases in IQ in India on the Present Value of Lifetime Earnings: A Structured Expert Judgment Study

Randall Lutter, Abigail Colson, and Roger Cooke*

1. Introduction

Many public health interventions targeted at infants and young children are thought to have the effect of raising cognitive performance, which is typically measured as IQ. Such interventions include 1) nutrition programs to supplement caloric intake (e.g., Pollitt et al., 1993 and Pollitt et al., 1995); to provide essential micronutrients (e.g., Black 2003 and Jauregui-Lobera, 2014); and to improve breastfeeding practices (e.g., Victora et al., 2016, Lutter and Lutter 2012); 2) environmental programs aimed at reducing exposure to lead or mercury during infancy and early childhood (e.g., Levin 1986; US EPA 2001, Shimshack and Ward 2010); and 3) early childhood education programs (Karoly et al. 2005; Elango et al. 2015). In the United States and some other developed countries, regulatory agencies and independent analysts have conducted estimates of the benefits (as well as costs) of interventions that raise IQ, using estimates of the effect of IQ gains on lifetime earnings (e.g., Levin 1986; US EPA 2001). In developing countries, however, such approaches appear rare, and we are unaware of any accepted estimates of the monetary value of IQ gains used for policy development in major developing country settings (see, e.g., World Bank 2016).

In this paper we develop estimates of the expected discounted gains in lifetime earnings that would result from a gain in IQ in India. Our approach has two parts. First, we conduct a structured expert judgment exercise consisting of one-on-one interviews with experts asked to quantify effects of IQ on earned income in India under specific scenarios. The structured expert judgment method is described in detail in Appendix A. In this exercise, we distinguish among effects of IQ at ages 25, 40, and 55, because labor market research in the United States suggests that associations between earnings and IQ grow over the lifecycle. Thus, any percentage gains on

* Lutter is professor of public policy at the Frank Batten School of Leadership and Public Policy at the University of Virginia and visiting fellow at Resources for the Future; corresponding author: lutter@rff.org. Colson is a research associate and PhD candidate at the University of Strathclyde in Glasgow and a resident scholar at the Center for Disease Dynamics, Economics & Policy in Washington, DC. Cooke is the Chauncey Starr Senior Fellow at Resources for the Future. They appreciate financial support from the Bill and Melinda Gates Foundation through the University of Virginia for this work. All views expressed in this paper are the sole responsibility of the authors and not necessarily those of the funder or any other organization. The authors had sole responsibility for study design, data collection and analysis, decision to publish, and preparation of the manuscript.

earned income observed early (e.g., at ages 25–30) may understate the effects at later ages. We elicited estimates of effects on both men and women, because labor force participation and compensation differ so much by gender. Finally, we sought information about effects for all of India and urban areas of India separately. Rural and urban India differ greatly—the former provides more limited educational opportunities and workforce mobility and (not coincidentally) there is significant out-migration to urban areas, particularly of capable young adults.

Second, we use data from the Indian Human Development Survey to estimate the present value of lifetime earnings for men and women, in light of labor force participation as well as survival probabilities from the abridged Indian lifetables for 2010–2015. We then apply the experts' estimates of effects of IQ gains on earned income at three different ages to derive an estimate in monetary terms of the gains from boosting IQ (or avoiding IQ decrements). We present estimates that reflect the uncertainty in the experts' judgments of expected gains in income at specific ages, as well as alternative assumptions about real discount rates, and variability (e.g., by gender). Our estimates of the value of IQ gains in monetary terms differ by gender, urban, or national setting as well as discount rates, and capture experts' uncertainty in the nature of the IQ earnings effects by age.

Our results suggest that a single IQ point gain for boys would raise the expected present value of lifetime income per child by between $109 and $249 dollars, depending on whether the discount rate is 3 percent or 7 percent. For girls, an equivalent range would be from $30 to $64, a range much less than that for boys because of lower propensity to participate in the labor force outside the home and lower earnings conditional on paid employment. To derive these estimates, we discounted to age 16 (i.e., approximately when labor force participation begins and opportunities to improve IQ end).

These estimates may be used to help identify interventions that might generally be expected to have positive or negative net benefits. Specifically, interventions that cost less than the lower bound estimate of the effects of IQ gains on the present value of income may be expected to offer net benefits. Interventions that cost more than the upper bound estimate of the gains in the present value of earnings from small increases in IQ may be expected to have negative net benefits. All other interventions are in an ambiguous category where more discerning research is needed.

2. Experts and Questions

2.1. Structured Expert Judgment

In this study, we use the Classical Model of structured expert judgment to quantify uncertainty about the impact of IQ on earnings in India. In the Classical Model, experts quantify their uncertainty on two types of questions: calibration questions and variables of interest (Cooke 1991). Calibration questions are items from the experts' field that are unknown to the expert but known by the study team. The variables of interest are the elicitation's target question.

Experts are scored on their performance on the calibration questions according to their statistical accuracy (measured as P-values) and information. The experts' assessments are then weighted by performance and combined into a performance-weighted decisionmaker (PW). The PW assessments are compared to the equally weighted decisionmaker (EW) that assigns all experts equal weight regardless of performance. More information on these scores and the Classical Model's weighting mechanism is provided in Appendix B.

2.2. Experts

The experts participating in this elicitation and their affiliations are listed in Table 1. We identified these experts by conducting a review of the economics literature estimating the relationship between various measures of IQ and labor market earnings in adulthood, and inviting participation from authors of papers in highly ranked journals based on our subjective assessment of the impact and importance of such papers.

Table 1. Experts and Affiliations

Name	**Affiliation**
David Deming	Harvard University
Kevin Lang	Boston University
Richard Murnane	Harvard University
Ronni Pavan	Rochester University
Rodrigo Pinto	UCLA
Catherine Weinberger	UCSB
Shintaro Yamaguchi	McMaster University
Jeffrey Zax	University of Colorado

We interviewed all experts individually using a video-conference program with document sharing capabilities. Interviews lasted one to two hours. To help the experts prepare for the interviews, we sent them a briefing book in advance, describing the methods and purpose of the study. To promote comparable familiarity with recent literature, we also provided previously published studies germane to the research question.

2.3. Variables

Calibration questions for this study asked about data from surveys that are frequently used in the IQ and earnings literature. The variables of interest asked experts about the results from a hypothetical, perfectly designed and implemented randomized controlled trial that raised the IQ of the intervention group two points at age twelve. The elicitation protocol is available in Appendix C.

In Table 2 we present a list of variables, along with short "code" names. In Table 2 calibration variables have realized values, but variables of interest do not.

Table 2. Calibration Variables and Variables of Interest

Variable No.	Short Name	Realized Values	Description
1	inc_reported	7.85E+01	In the NLSY79 representative sample (without the Hispanic and African-American oversamples), there were initially 6111 subjects of whom 5751 have AFQT scores. What percent of these 5751 subjects have data for earned income in 2008?
2	ppvt_1stborn	9.43E+01	In the NLSY79-Children data the average observed Peabody Picture Vocabulary Test (PPVT) mean score is 90.660. What is the average observed PPVT score among first-borns?
3	inc_25_afqt	1.16E+04	What was the average earned income at age 25 among those whose AFQT score, tested at age 22, was above the average of this group?
4	inc_50_afqt	6.96E+04	What was the average earned income at age 50 among those whose AFQT score, tested at age 22, was above the average of this group?
5	inc_25_12th	1.15E+04	What was the average earned income at age 25 among those mothers who completed (at least) the 12th grade?
6	inc_25_2col	1.24E+04	What was the average earnings at age 25 among those who completed (at least) 2 years of college?
7	afqt89_80	1.35E+00	In the NLSY79-Children data, what is the ratio of the average mothers' AFQT scores for children born in 1989 / 1980?
8	afqt06_89	1.05E+00	In the NLSY79-Children dataset, what is the ratio of the average mothers' AFQT scores for children born in 2006 / 1989?
9	age_reported	6.52E+01	In what percentage of the PSID-C records is mother's age at birth reported?

10	avg_inc	4.45E+04	In the PSID-C dataset, the average of the reported family income (97) is $35,100. What is the average among records in which birth order is reported?
11	4th_avgincom e	8.50E-01	In the PSID-C dataset, consider the average reported family income for children for whom birth order is recorded. What is the ratio of the average family income at time of 4th birth relative to the above average family income?
12	all_25_male		All India: % change in earnings at age 25, males
13	all_25_female		All India: % change in earnings at age 25, females
14	all_40_male		All India: % change in earnings at age 40 males
15	all_40_female		All India: % change in earnings at age 40, females
16	all_55_male		All India: % change in earnings at age 55, males
17	all_55_female		All India: % change in earnings at age 55, females
18	urb_25_male		Urban India: % change in earnings at age 25, males
19	urb_25_femal e		Urban India: % change in earnings at age 25, females
20	urb_40_male		Urban India: % change in earnings at age 40, males
21	urb_40_femal e		Urban India: % change in earnings at age 40, females
22	urb_55_male		Urban India: % change in earnings at age 55, males
23	urb_55_femal e		Urban India: % change in earnings at age 55, females

3. Expert Elicitation Results

3.1. Solution

We present the PW's 5, 25, 50, 75, and 95 percentiles in Table 3.[1] The distributions from each individual expert and the EW assessment are presented in Appendix A.

Table 3. Optimized Performance Weight Solution

Variable No.	Short Name	5%	25%	50%	75%	95%
1	inc_reported	25.66	41.77	63.96	80.37	90.03
2	ppvt_1stborn	83.89	87.65	91.52	93.63	101.5
3	inc_25_afqt	6237	1.17E+04	1.44E+04	2.17E+04	3.81E+04
4	inc_50_afqt	4.34E+04	5.87E+04	6.87E+04	8.29E+04	1.18E+05
5	inc_25_12th	9107	1.18E+04	1.26E+04	1.67E+04	2.49E+04
6	inc_25_2col	6501	1.20E+04	1.45E+04	1.98E+04	3.33E+04
7	afqt89_80	0.8505	0.989	1.09	1.203	1.443
8	afqt06_89	0.7071	0.99	1.086	1.246	1.883
9	age_reported	30.44	72.22	87.92	98.39	99.89
10	avg_inc	2.41E+04	3.34E+04	3.60E+04	3.97E+04	5.22E+04
11	4th_avgincome	0.3876	0.7386	0.9464	1.169	2.245
12	all_25_male	1.41E-05	0.3184	1.485	4.922	17.47
13	all_25_female	1.20E-05	0.2778	1.825	5.08	17.07
14	all_40_male	2.74E-05	0.4414	2.747	8.293	26.2
15	all_40_female	1.25E-05	0.4365	3.144	9.284	30.3
16	all_55_male	1.45E-05	0.3698	2.817	8.58	25.75
17	all_55_female	1.25E-05	0.3438	2.626	8.258	33.1
18	urb_25_male	2.26E-05	0.3747	1.713	5.27	20.97
19	urb_25_female	2.32E-05	0.3176	1.922	5.338	20.54
20	urb_40_male	2.59E-05	0.4725	2.949	8.664	29.87
21	urb_40_female	1.27E-05	0.4635	3.309	9.565	34.07
22	urb_55_male	1.40E-05	0.3842	3.163	9.169	33.59
23	urb_55_female	1.24E-05	0.3761	3.159	8.559	40.25

3.2. Expert and Decisionmaker Scores

Five of the eight experts (numbers 1–4 and 6 in Table 4) were weighted in the PW. Three of these exhibited P-values above the traditional 5 percent threshold for rejecting a null hypothesis (in this case, the null hypothesis is that an expert's probabilistic assessments are

[1] As described in Appendix B, a performance-weighted decisionmaker can be based on item weights or global weights. In this study, item weights perform better and are used throughout the results.

statistically accurate—hence we do not want to reject the null hypothesis, and high P-values reflect high statistical accuracy). Experts 5, 7, and 8 were not weighted in the PW. Scores in the "Mean information" columns in Table 4 give the average information with respect to the uniform background measure for all variables and for calibration variables (columns 3 and 4 respectively). These values indicate no great differences in informativeness between the calibration variables and the variables of interest. Column 5 ("Combined score") gives the product of column 2 and column 4. This is the expert's un-normalized weight, but the actual weights vary per item, being the product of the P-Value and the informativeness per item.

Most informativeness scores vary within a factor 3. The ratio of informativeness scores corresponds very roughly to the ratio of the 90 percent confidence intervals. In this case, both the EW and the PW exhibit good statistical accuracy and roughly comparable informativeness. Although PW's combined score (column 5) is slightly greater than that of EW, the difference is small. Both scores are substantially above those of the experts. The two rightmost columns give the relative information of each expert with respect to the EW, for all variables and for calibration variables only. These numbers indicate the amount of agreement and are best interpreted visually in Appendix A. The lower these numbers are, the more the experts resemble each other. Note also that the unweighted experts are more unlike the EW decisionmaker than the weighted experts.

Table 4. Expert and Decisionmaker Performance Results

1.	2.	3.	4.	5.	6.	7.
		Mean Information			Information relative to EW	
		All variables	Calibration variables		All variables	Calibration variables
1	1.94E-02	1.321	1.177	2.29E-02	0.464	0.520
2	5.18E-02	2.276	1.907	9.87E-02	1.483	1.118
3	4.54E-01	0.587	0.731	3.32E-01	0.337	0.345
4	1.05E-01	0.885	0.300	3.15E-02	0.354	0.413
5	2.92E-03	1.726	1.791	5.24E-03	0.716	0.996
6	3.20E-03	2.111	1.576	5.05E-03	0.831	0.676
7	6.97E-07	1.690	1.738	1.21E-06	0.652	0.828
8	3.02E-04	1.227	1.123	3.39E-04	0.780	0.843
EW	0.7046	0.776	0.575	0.405	0.000	0.000
PW	0.7046	0.743	0.623	0.439	0.247	0.205

Notes: EW denotes equal weighting and PW denotes performance weighting, with item weights. The experts that were not weighted are shaded. Expert numbers here do not correspond with the order in Table 1.

3.3. Robustness on Items

An optimal solution always invites non-robustness. In computing robustness on items, we remove calibration variables one at a time, and we compare the scores with the "unperturbed" decisionmaker (here, the PW). Comparing the last two columns of Table 5 with the last two columns of Table 4, we see that perturbations due to loss of a calibration variable are less than the differences among the experts themselves.

Table 5. Robustness on Calibration Variables

	Mean information			Information relative to original PW	
Excluded item	All variables	Calibration variables	P-Value	All variables	Calibration variables
inc_reported	0.952	0.835	0.823	0.128	0.152
ppvt_1stborn	0.799	0.615	0.720	0.062	0.073
inc_25_afqt	0.597	0.768	0.499	0.209	0.215
inc_50_afqt	0.590	0.753	0.531	0.213	0.224
inc_25_12th	0.572	0.712	0.499	0.206	0.208
inc_25_2col	0.763	0.634	0.756	0.064	0.064
afqt89_80	0.788	0.715	0.756	0.050	0.057
afqt06_89	0.751	0.553	0.756	0.042	0.052
age_reported	0.878	0.873	0.562	0.170	0.210
avg_inc	0.565	0.698	0.499	0.208	0.213
4th_avgincome	0.770	0.796	0.756	0.165	0.186
None	0.743	0.623	0.705		

3.4. Robustness on Experts

We also compute robustness relative to choice of experts by removing experts one at a time and re-computing the model, as shown in Table 6. Again, we compare the scores with the "unperturbed" PW. Unsurprisingly, removing the unweighted experts has negligible effects. (There may be a small effect on the uniform background measure, which is computed from all the experts' assessments.) The rightmost columns show that these results are very robust against loss of an expert. Overall, we conclude that lack of robustness in this study against the loss of one calibration variable or loss of one expert is not an issue.

Table 1. Robustness on Experts

Excluded expert	Mean information All variables	Mean information All variables	P-Value	Information relative to original PW All variables	Information relative to original PW Calib Vbls
1	0.740	0.618	0.705	0.022	0.023
2	0.631	0.529	0.705	0.091	0.100
3	0.721	0.836	0.669	0.298	0.323
4	0.658	0.542	0.669	0.142	0.199
5	0.743	0.623	0.705	0.000	0.000
6	0.730	0.599	0.705	0.009	0.009
7	0.739	0.614	0.705	0.000	0.000
8	0.733	0.601	0.705	0.004	0.009
None	0.743	0.623	0.705	0.000	0.000

3.5. Expert Rationales

Each expert provided a qualitative rationale for his or her quantitative assessments. These rationales are provided in Appendix D. Generally, experts thought the impact of IQ gains on earnings in India would depend on the extent to which educational opportunities are made available and how efficiently the labor market provides returns to skill. Experts who compared the impact of IQ on earnings in India to the relevant literature from the United States thought the impact would likely be lower in India due to lower access to higher education and lower general returns to skill in the Indian labor market; they also thought the impact was more uncertain in India. Experts agreed that the impact of IQ on earnings would be lowest at age 25 and would generally increase with age. However, some experts thought that by age 55, a higher proportion of the remaining labor force could work in agriculture and other low-tech fields with lower returns to skill, so the impact of IQ could be lower at age 55 than at age 40. Most experts thought the lower frequency of agricultural work in urban India would result in a higher impact of IQ in urban India than in the country at large. The experts were divided on the different impact for men and women. Some thought the low labor force participate rate of women would depress the impact on earnings, while others thought an increase in IQ could subsequently increase women's labor force participation, and it would thus have a larger impact on women's earnings.

4. Labor Income in India

To express these estimates of percent change in monetary terms, we need to consider available information on labor market earnings over the lifecycle, by gender. We use the Indian Human Development Survey (IHDS), a nationally representative survey described elsewhere, as well as the abridged SRS-Based Life Table for 2011–2015, from the Office of the Registrar General and Census Commissioner within India's Ministry of Home Affairs.[2] The IHDS provides valid measures of annual cash wages for 53,404 cases out of a total of 204,569 respondents.[3] Table 7 presents summary statistics for individuals older than 14 and younger than 65.

Table 7. Selected Summary Statistics for Labor Income in India

Variable	N	Mean	Standard Deviation	Min	Max
Annual cash wages	51,279	51733.23	79348.32	0	2400000

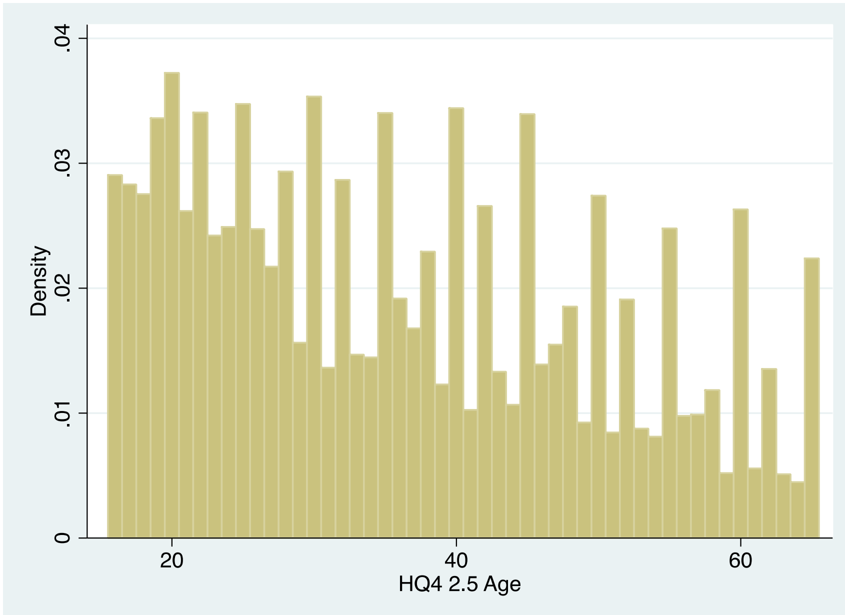
We exclude from consideration private business income, because many small businesses are supported by work from family members who do not receive an explicit or regular wage or salary for their work. In such cases, private business income is difficult or impossible to attribute to any individual. We also do not consider the value of any fringe benefits provided to employees as such information is not presented in the IHDS. As a result of our inability to include private business income and fringe benefits, our final estimates understate effects of IQ gains on total expected lifetime income.

The IHDS also reveals substantial clustering of reported ages at notable milestones divisible by five, such as 25 and 30. For example, the codebook reports without comment that the percent of valid responses for one's own age (out of 204,565), for ages 29, 30, and 31 is 1, 2.3, and 0.9 percent respectively (IHDS Codebook, 20). For ages 39, 40, and 41, the percent of valid responses is 0.8, 2.2, and 0.7, respectively (IHDS Codebook, 20). Such clustering is not surprising in a population with limited literacy and personal documentation, but it creates challenges for estimating earnings by age. Figure 1 illustrates the clustering of respondents' reported ages and shows that it also exists at ages not divisible by five.

[2] For the life tables, please see http://www.censusindia.gov.in/Vital_Statistics/SRS_Life_Table/Srs_life_Table_2011-15.html. For more information on the India Human Development Survey, please see the India Human Development Survey Web site at www.ihds.umd.edu .

[3] It also provides information on annual wages through the Mahatma Gandhi National Rural Employment Guarantee Scheme, but only 9,077 cases have valid nonmissing data and for 4203 of these, the values for income from the NREG program is the same as for the annual cash wages, suggesting substantial redundancy.

Figure 1. Ages of Respondents to IHDS Who Are Between 16 and 65 Years of Age



We elect to focus on age groups defined as 15–19, 20–24, etc., since such groups correspond to the age groups in the Indian abridged life tables that present data on mortality risks. Tables 8a and 8b present non-missing annual cash wages for men and women, by these age groups.

Table 8a. Annual Cash Wages, by Age Group, Men

Age Group	Min Age	Max Age	Number of Observations	Mean	SD	Min	Max
1	15	19	2,140	24200.03	23702.8	0	432000
2	20	24	4,739	40135.71	43471.17	0	960000
3	25	29	5,220	54856.98	59593.12	0	720000
4	30	34	4,598	60501.54	80048.32	100	2400000
5	35	39	4,591	63554.02	75181.83	360	1200000
6	40	44	4,011	68185.18	83369.34	0	966000
7	45	49	3,903	76399.59	99984.67	130	1800000
8	50	54	3,005	93432.25	123779.1	180	1200000
9	55	59	2,192	99063.35	135234.4	0	1080000
10	60	64	1,255	49439.76	81899.02	240	1200000

Note: Units are Indian Rupees. The total number of observations is 35654.

Table 8b. Annual Cash Wages by Age Group, Women

Age Group	Min Age	Max Age	Number of Observations	Mean	SD	Min	Max
1	15	19	914	10494.72	13135.83	140	150000
2	20	24	1,428	25425.74	43105.98	75	360000
3	25	29	1,972	27773.83	56138.06	150	960000
4	30	34	2,090	26569.56	52864.15	100	588000
5	35	39	2,438	27849.05	57767.08	100	1020000
6	40	44	2,140	27785.44	55009	70	600000
7	45	49	1,867	31441.98	64692.5	100	600000
8	50	54	1,252	30457.71	69946.33	0	600000
9	55	59	895	33755.46	81541.4	120	840000
10	60	64	629	19525.13	44428.51	200	444000

Note: Units are Indian Rupees. The total number of observations is 15,625.

We develop estimates of lifetime income that reflect nonparticipation in the work force and mortality risks. For simplicity, we do not address part-time versus full-time work, but instead assess the probability of working using a simple de minimis rule: all individuals who work more than 12 days in the preceding year (i.e., one day per month) are counted as working. In Tables 9a and 9b we present this measure of labor force participation for men and women, based on data from IHDS.[4] We assume conservatively that IQ has no effect on labor force participation because we are unaware of any Indian data suitable to characterize such an effect, although Lin, Lutter, and Ruhm (2016) show that higher cognitive performance predicts greater work in the United States, using the National Longitudinal Survey of Youth (1979).

[4] We also consider an alternative measure based on the hours worked for adults of different ages in the year prior to administration of the survey. This alternative de minimis rule—which defines "working" as at least 50 hours per year--results in rates of labor force participation that are similar. For example, for men in age groups 2, 3, and 4, the percent of individuals working more than 50 hours per year is 70.4, 87.4, and 92.3, respectively.

Table 9a. Men's Labor Force Participation, by Age Group

Age Group	Mean workdays	Mean work hours	Percent of individuals working more than 12 days per year
1	69	445	0.434
2	161	1216	0.712
3	227	1777	0.876
4	247	1930	0.924
5	257	2013	0.940
6	255	1979	0.935
7	253	1945	0.937
8	239	1823	0.915
9	217	1611	0.871
10	161	1108	0.722
Note: Age groups are ages 15–19, 20–24, etc.			

Table 9b. Women's Labor Force Participation, by Age Group

Age Group	Mean workdays	Mean work hours	Percent of individuals working more than 12 days per year
1	31	155	0.263
2	46	279	0.299
3	71	445	0.410
4	91	570	0.495
5	109	687	0.561
6	107	670	0.558
7	102	633	0.538
8	92	555	0.515
9	76	454	0.447
10	59	330	0.371
Note: Age groups are ages 15–19, 20–24, etc. Hours worked for women exclude childrearing and housekeeping.			

We also incorporate the probability of surviving to different ages, which we take from the 2011–2015 abridged life tables for India and present in Table 10. We ignore any beneficial effect that greater IQ may have on survival probabilities.

Table 10. Probability of Dying, By Age Group and Gender

Age Group	Ages	All	Men	Women
	0–1	0.04200	0.04101	0.04313
	1–5	0.00883	0.00709	0.01081
	5–10	0.00409	0.00409	0.00414
	10–15	0.00334	0.00349	0.00315
1	15–20	0.00539	0.00529	0.00549
2	20–25	0.00742	0.00797	0.00688
3	25–30	0.00827	0.00951	0.00693
4	30–35	0.01010	0.01252	0.00757
5	35–40	0.01361	0.01745	0.00966
6	40–45	0.01859	0.02334	0.01341
7	45–50	0.02657	0.03395	0.01874
8	50–55	0.04230	0.04928	0.03415
9	55–60	0.06137	0.07555	0.04838
10	60–65	0.09344	0.10524	0.08088

With these data, we can use a standard present value formula to calculate expected lifetime income given labor force participation rates and survival probabilities. In these calculations, we use 3 percent and 7 percent real discount rates and the mean values of annual cash wages. We discount to age 16, which is very roughly the latest age when schooling can improve cognitive test results and just before labor force participation. For the age group 15–20, we use a linear interpolation procedure to focus only on income in years 18 and 19, since labor force participation is quite low below age 18. We present the results in Table 11.

Table 11. Present Value Lifetime Earnings, by Gender and Discount Rate in Indian Rupees

At age	Discount Rate	Men	Women
	3 percent	1,152,410	274,177
	7 percent	531,812	129,897

Source: We derive these estimates from the cross-sectional Indian Health and Demographic Survey data and the abridged life tables for India. They reflect the present value at age 16 of all future wages and salaries, excluding the value of income derived from small businesses and any fringe benefits. The units are Indian Rupees for 2011 and 2012.

To calculate the expected gain in the present value of lifetime income we begin with the results of our expert elicitation, which apply to ages 25, 40, and 55, and make linear interpolations for the intervening ages divisible by five. The 50th percentile results of the expert elicitation for the all India scenario and the interpolated values are presented in Table 12.

Table 12. Expected Increases in Earned Income from a 2 IQ Point Gain at Selected Ages

Age	Men	Women
15	0.00	0.00
20	0.74	0.91
25	**1.49**	**1.83**
30	1.91	2.26
35	2.33	2.70
40	**2.75**	**3.14**
45	2.77	2.97
50	2.79	2.80
55	**2.82**	**2.63**
60	1.88	1.75
65	0.94	0.88
70	0.00	0.00

Note: The bolded values for ages 25, 40 and 55 represent the 50th percentile estimate for the all-India scenario presented to our expert panel. The other values are linear interpolations, which in some cases are anchored by conservative assumptions of no effects at ages 15 and 70.

With these estimates of the gains in per capita earned income at different ages we can calculate the gain in present value of discounted income that would result from an increase in IQ of 2 points throughout India, the increment assumed in the expert elicitation (Table 13).

Table 13. Expected Gains in Present Value Lifetime Earnings in India, from an Increase in IQ of 2 Points, in Rupees

Discount Rate	Men	Women
3 percent	26,352	6,833
7 percent	11,522	3,152
Source: Authors' calculations. Note that these are present value at age 16.		

Interpreting these estimates may be helped by a conversion to dollars. Google reports the current rate in July 2017 was 64.1054 Indian Rupees to the dollar.[5] The survey was conducted in 2011 and 2012. During this period, the exchange rate went from about 45 in January 2011 to a temporary high of about 53 in January 2012 to a sustained plateau of 55 in late 2012 until the middle of 2013, when the exchange rate continued its rise. We adopt a rate of 53 rupees per

[5] See https://www.google.com/finance?q=usdinr&ei=IeR5WYDqMIWnmAGgr7KQBg.

dollar, which results in estimates in US dollars for 2011–2012. In Table 14, we present dollar estimates of gains for men and women at 3 and 7 percent discount rates. In Table 14, we also converts the gains to those expected to result from an increase of a single IQ point, so as to facilitate comparisons with other estimates in the literature.[6]

Table 14. Expected Gains in Present Value Lifetime Earnings in India, from an Increase in IQ of 1 Point, in US dollars

Discount Rate	Men	Women
3 percent	$249	$64
7 percent	$109	$30

Source: Authors' calculations, assuming 53 rupees to the dollar and discounting to age 16.

The expert elicitation also provided 5th and 95th percentile values. The former, a lower bound is essentially zero, being less than one hundredth of the 50th percentile value. The latter, however, is relatively high and reflects some experts' expressed belief that liberalization of the educational system and labor markets will substantially increase the compensation for more able workers. Table 15 shows the results using the 95th percentile values from the expert elicitation.

Table 15. Upper Bound for Gains in PV Lifetime Earnings in India, from an Increase in IQ of 1 Point, in US dollars

Discount Rate	Men	Women
3 percent	$2429	$671
7 percent	$1092	$299

Source: Authors' calculations, assuming 53 rupees to the dollar and discounting to age 16.

The choice of age at which to express the expected gains in income is somewhat arbitrary. If we focused instead on age 1, one would have to discount these estimates for the additional 15 year delay as well as the probability of not surviving from age 1 to age 16. We present the discount factors for such a procedure in Table 16.

[6] Other work (e.g., Lin, Lutter, and Ruhm 2016), suggests that associations between IQ or comparable measures of cognitive ability and earnings are approximately linear, so we divide by two to get effects of a gain of one IQ point.

Table 16. Discount Factors to Convert Values at Age 16 to Values at Age 1

Discount Rate	Men	Women
3 percent	0.634	0.632
7 percent	0.358	0.357

Source: Authors' calculations.

We refrain from estimating population benefits because they would depend on the scope, scale, and effectiveness of any intervention that might raise IQ. For illustration, however, one might consider a hypothetical intervention that raises the IQ of 5 million infants in India by 2 IQ points each (e.g., through improved breastfeeding or reduced exposure to environmental lead). In this instance, the estimate of annual social benefits for boys involves the product of the number of male births (approximately 2.5 million), the IQ gain of 2 points, and the dollar value (e.g., $230). This example suggests that large-scale effective interventions could provide large economic benefits.

At this point, one might consider how these estimates compare with extrapolations of comparable estimates from developed countries such as the United States. For example, in its analysis of possible revisions to national standards for airborne lead levels, EPA estimated the present value lifetime income effects of a 1-unit increase in IQ to be between $8,760 and $12,512 (US EPA 2008). Converting these figures to 2014 dollars using the CPI-U, and discounting them from age 3 to age 1 (at 3 percent) would give adjusted estimates $9,740 and $13,911, respectively. Applying the discount factors for 3 percent that appear in Table 16 to the estimates of benefits (for a 3 percent discount rate) from Table 14, and taking the mean of the values for men and women would give an estimate of $99.16. This estimate (on the order of 1 percent of the adjusted estimates) is much less than might be expected by multiplying the EPA values by the ratio of real income in India to real income in the United States. Such adjustments might be appropriate, if the estimate of the income gains from IQ improvements in different countries in fact were proportional to the per capita gross domestic product. Robinson (2017) suggests that such proportionality is reasonable for estimates of willingness to pay to reduce mortality risk in developing countries. To illustrate, the CIA World Factbook reports that the estimated per capita gross domestic product for the United States in 2016 was $57,300, while a comparable number for India, using an exchange rate that reflects purchasing power parity, was only $6,700.[7] This suggests that mean GDP per capita is about 8.55 times greater in the United States than in India. Applying that ratio to the adjusted EPA estimates would give estimates of

[7] See https://www.cia.gov/library/publications/the-world-factbook/rankorder/2004rank.html.

$1,139 to $1,627. Thus, the estimates that we have derived appear quite low relative to EPA's estimates—after simple (and simplistic) proportional adjustments for income differences between the United States and India.

5. Conclusions

We have used expert elicitation and household survey data to calculate the expected present value of increases in lifetime earnings from interventions that might increase IQ. Our estimates are robust to our choice of experts and to calibration questions. They indicate that much of the benefits of such interventions may depend on future success in liberalizing educational opportunities and labor markets in India.

References

Black, Maureen M. 2003. "Micronutrient Deficiencies and Cognitive Functioning." *The Journal of Nutrition* 133(11): 3927S–3931S.

Cooke, Roger M. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press, ISBN 0-19-506465-8.

Colson, A., and R.M. Cooke. 2017. "Cross Validation for the Classical Model of Structured Expert Judgment." *Reliability Engineering and System Safety* 163(July): 109–120. http://dx.doi.org/10.1016/j.ress.2017.02.003

Eggstaff, J.W., T.A. Mazzuchi, and S. Sarkani, S. 2014. "The Effect of the Number of Seed Variables on the Performance of Cooke's Classical Model." *Reliability Engineering and System Safety* 121: 72–82. DOI: 10.1016/j.ress.2013.07.015

Elango, S., J. Garcia, J. Heckman, and A., Hojman. 2015. Early Childhood Education. Working paper available at https://heckmanequation.org/assets/2017/01/FINALMoffitt-ECE-Paper2015.pdf.

Jauregui-Lobera, Ignacio, 2014, Iron Deficiency and Cognitive Functions, *Neuropsychiatric Disease and Treatment*. 10: 2087–2095. 10.2147/NDT.S72491

Karoly, Lynn A., M. Rebecca Kilburn, and Jill S. Cannon. 2005. "Proven Benefits of Early Childhood Interventions." Santa Monica, CA: RAND Corporation. http://www.rand.org/pubs/research_briefs/RB9145.html.

Levin, Ronnie. 1986. Reducing Lead in Drinking Water: A Benefits Analysis. EPA, EPA-230-09—86-019, Washington, DC

Lin, Dajun, Randall Lutter, and Christopher J. Ruhm. 2016. "Cognitive Performance and Labor Market Outcomes". NBER Working Paper No. 22470. http://www.nber.org/papers/w22470.

Lutter, Chessa K., and Randall Lutter. 2012. "Fetal and Early Childhood Undernutrition, Mortality, and Lifelong Health." *Science* 337(6101): 1495–1499.

Pollitt, E., K. S. Gorman, P.L. Engle, R. Martorell, and J. Rivera. 1993. "Early Supplementary Feeding and Cognition: Effects over Two Decades." *Monographs of the Society for Research in Child Development*; 58(7): 1–99; discussion 111–8.

Pollitt, E., K. S. Gorman, P.L. Engle, J. Rivera, and R., Martorell. 1995. "Nutrition in Early Life and the Fulfillment of Intellectual Potential." *Journal of Nutrition* April 125 (4 Suppl): 1111S–1118S.

Robinson, Lisa. 2017. "Estimating the Values of Mortality Risk Reductions in Low- and Middle-Income Countries." *Journal of Benefit-Cost Analysis* 1–10, doi:10.1017/bca.2017.14.

Shimshack, Jay P., and Michael B. Ward. 2010. "Mercury Advisories and Household Health Trade-offs." *Journal of Health Economics* 29(5): 674–685.

US Environmental Protection Agency (EPA). 2001. Lead; Identification of Dangerous Levels of Lead: Final Rule. *Federal Register* 66, 1206-1240. January 5. https://www.gpo.gov/fdsys/pkg/FR-2001-01-05/pdf/01-84.pdf.

———. 2008. Regulatory Impact Analysis of the Proposed Revisions to the National Ambient Air Quality Standards for Lead: Chapter 5. Benefits Analysis Approach and Results. http://www.epa.gov/ttnecas1/regdata/RIAs/finalpbriach5.pdf.

Victora, Cesar G., Rajiv Bahl, Aluísio JD Barros, Giovanny VA França, Susan Horton, Julia Krasevec, Simon Murch et al., 2016. "Breastfeeding in the 21st Century: Epidemiology, Mechanisms, and Lifelong Effect." *The Lancet* 387(10017): 475–490.

World Bank. 2016. Nine Countries Pledge Greater Investments in Children, Powering Economies for Long-Term Growth. Press release, October 6. http://www.worldbank.org/en/news/press-release/2016/10/06/nine-countries-pledge-greater-investments-in-children-powering-economies-for-long-term-growth.

Appendix A. Performance Measures and Combination: The Classical Model

There are two generic, quantitative measures of performance, *calibration* and *information*. Loosely, calibration measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with the expert's assessments. Information measures the degree to which a distribution is concentrated. To simplify the exposition we assume that the 5 percent, 50 percent and 95 percent values were elicited.

Calibration

For each quantity, each expert divides the range into 4 inter quantile intervals for which his/her probabilities are known, namely p_1 = 0.05: less than or equal to the 5 percent value, p_2 = 0.45: greater than the 5 percent value and less than or equal to the 50 percent value, etc.

If N quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each realization falls in one of the four inter-quantile intervals with probability vector

$$p = (0.05, 0.45, 0.45, 0.05).$$

Suppose we have realizations $x_1, \ldots x_N$ of these quantities. We may then form the sample distribution of the expert's inter quantile intervals as:

$$s_1(e) = \#\{\ i \mid x_i \leq 5\%\ quantile\}/N$$
$$s_2(e) = \#\{\ i \mid 5\%\ quantile < x_i \leq 50\%\ quantile\}/N$$
$$s_3(e) = \#\{\ i \mid 50\%\ quantile < x_i \leq 95\%\ quantile\}/N$$
$$s_4(e) = \#\{\ i \mid 95\%\ quantile < x_i\ \}/N$$
$$s(e) = (s_1, \ldots s_4)$$

Note that the sample distribution depends on the expert *e*. If the realizations are indeed drawn independently from a distribution with quantiles as stated by the expert then the quantity

$$2NI(s(e) \mid p) = 2N\ \Sigma_{i=1..4}\ s_i \ln(s_i / p_i) \qquad (1)$$

is asymptotically distributed as a chi-square variable with 3 degrees of freedom. This is the so-called likelihood ratio statistic, and I(s | p) is the relative information of distribution *s* with respect to *p*. If we extract the leading term of the logarithm we obtain the familiar chi-square test statistic for goodness of fit. There are advantages in using the form in (1) (Cooke 1991).

If after a few realizations the expert were to see that all realization fell outside his 90 percent central confidence intervals, he might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are *not* independent, and he learns from the realizations. Expert learning is not a goal of an expert judgment study and his joint distribution is not elicited. Rather, the decisionmaker wants experts who do not need to learn from the elicitation. Hence the decisionmaker scores expert *e* as the statistical likelihood of the hypothesis

H_e: "the inter quantile interval containing the true value for each variable is drawn independently from probability vector p."

A simple test for this hypothesis uses the test statistic (1), and the likelihood, or p-value, or **calibration score** of this hypothesis, is:

$$Cal(e) = p\text{-}value = Prob\{\ 2NI(s(e) \mid p) \geq r \mid H_e\}$$

where *r* is the value of (1) based on the observed values $x_1, \ldots x_N$. It is the probability under hypothesis H_e that a deviation at least as great as *r* should be observed on *N* realizations if H_e were true. Calibration scores are absolute and can be compared across studies. However, before doing so, it is appropriate to equalize the power of the different hypothesis tests by equalizing the effective number of realizations. To compare scores on two data sets with *N* and *N' realizations*, we simply use the minimum of *N* and *N'* in (1), without changing the sample distribution *s*. In some cases involving multiple realizations of one and the same assessment, the effective number of seed variables is based on the number of assessments and not the number of realizations.

Although the calibration score uses the language of simple hypothesis testing, it must be emphasized that we are not rejecting expert-hypotheses; rather we are using this language to measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct.

Information

The second scoring variable is information. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or "spread out" is measured relative to some other distribution.

Measuring information requires associating a density with each quantile assessment of each expert. To do this, we use the unique density that complies with the experts' quantiles and is minimally informative with respect to the background measure. This density can easily be found with the method of Lagrange multipliers. For a uniform background measure, the density is constant between the assessed quantiles, and is such that the total mass between the quantiles agrees with *p*. The background measure is not elicited from experts as indeed it must be the same for all experts; instead it is chosen by the analyst.

The uniform and log-uniform background measures require an *intrinsic range* on which these measures are concentrated. The classical model implements the so-called *k* percent overshoot rule: for each item we consider the smallest interval $I = [L, U]$ containing all the assessed quantiles of all experts and the realization, if known. This interval is extended to

$$I^* = [L^*, U^*]; L^* = L - k(U\text{-}L)/100; U^* = U + k(U\text{-}L)/100.$$

The value of k is chosen by the analyst. A large value of *k* tends to make all experts look quite informative, and tends to suppress the relative differences in information scores. The **information score** of expert *e* on assessments for uncertain quantities 1…N is

$$Inf\,(e) = Average\ Relative\ information\ wrt\ Background = (1/N)\ \Sigma_{i\,=\,1..N}\ I(f_{e,i} \mid g_i)$$

where g_i is the background density for variable *i* and $f_{e,i}$ is expert *e*'s density for item i. This is proportional to the relative information of the expert's joint distribution given the background, under the assumption that the variables are independent. As with calibration, the assumption of independence here reflects a desideratum of the decisionmaker and not an elicited feature of the expert's joint distribution. The information score does not depend on the realizations. An expert can give himself a high information score by choosing his quantiles very close together.

Evidently, the information score of *e* depends on the intrinsic range and on the assessments of the other experts. Hence, information scores cannot be compared across studies.

Of course, other measures of concentrated-ness could be contemplated. The above information score is chosen because it is

- familiar
- tail insensitive
- scale invariant
- slow

The latter property means that relative information is a slow function; large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the calibration score, which is a very fast function. This causes the product of calibration and information to be driven by the calibration score.

Combination: Decisionmaker

The **combined score** of expert *e* will serve as an (unnormalized) weight for *e*:

$$w_\alpha(e) = Cal\,(e) \times Inf\,(e) \times 1_\alpha(Cal(e) \geq \alpha), \quad (2)$$

where $1_\alpha(Cal(e)\alpha) = 1$ if $Cal(e) \geq \alpha$, and is zero otherwise. The combined score thus depends on α. if *Cal(e)* falls below cut-off level α expert *e* is unweighted. The presence of a cut-off level is imposed by the requirement that the combined score be an asymptotically strictly proper scoring rule. That is, an expert maximizes his/her long run expected score by and only by ensuring that his probabilities *p= (0.05, 0.45, 0.45, 0.05)* correspond to his/her true beliefs. α is similar to a significance level in simple hypothesis testing, but its origin is indeed different. The goal of scoring is not to “reject” hypotheses, but to measure “goodness” with a strictly proper scoring rule.

A combination of expert assessments is called a "decisionmaker" (DM). All decisionmakers discussed here are examples of linear pooling. The classical model is essentially a method for deriving weights in a linear pool. "Good expertise" corresponds to good calibration (high statistical likelihood, high p-value) and high information. We want weights which r*EW* ard good expertise and which pass these virtues on to the decisionmaker.

The r*EW* ard aspect of weights is very important. We could simply solve the following optimization problem: find a set of weights such that the linear pool under these weights maximizes the product of calibration and information. Solving this problem on real data, one finds that the weights do not generally reflect the performance of the individual experts. As we do not want an expert's influence on the decisionmaker to appear haphazard, and we do not want to encourage experts to game the system by tilting their assessments to achieve a desired outcome, we must impose a strictly scoring rule constraint on the weighing scheme.

The scoring rule constraint requires the term 1_α*(calibration score),* but does not say what value of α we should choose. Therefore, we choose α so as to maximize the combined score of the resulting decisionmaker. Let $DM_\alpha(i)$ be the result of linear pooling for item *i* with weights proportional to (2):

$$DM_\alpha(i) = \Sigma_{e=1,..E}\, w_\alpha(e)\, f_{e,i} \,/\, \Sigma_{e=1,..E}\, w_\alpha(e) \quad (3)$$

The optimized global weight DM is DM_{α^*} where α* maximizes

$$\textit{calibration score}(DM_\alpha) \times \textit{information score}(DM_\alpha). \tag{4}$$

This weight is termed global because the information score is based on all the assessed seed items.

A variation on this scheme allows a different set of weights to be used for each item. This is accomplished by using information scores for each item rather than the average information score:

$$w_\alpha(e,i) = 1_\alpha(\textit{calibration score}) \times \textit{calibration score}(e) \times I(f_{e,i} \mid g_i) \tag{5}$$

For each α we define the Item weight DM_α for item i as

$$IDM_\alpha(i) = \Sigma_{e=1,..E}\, w_\alpha(e,i)\, f_{e,i} \,/\, \Sigma_{e=1,..E}\, w_\alpha(e,i) \tag{6}$$

The *optimized item weight DM* is IDM_{α^*} where α* maximizes

$$\textit{calibration score}(IDM_\alpha) \times \textit{information score}(IDM_\alpha). \tag{7}$$

The non-optimized versions of the global and item weight DM's are obtained simply by setting α = 0.

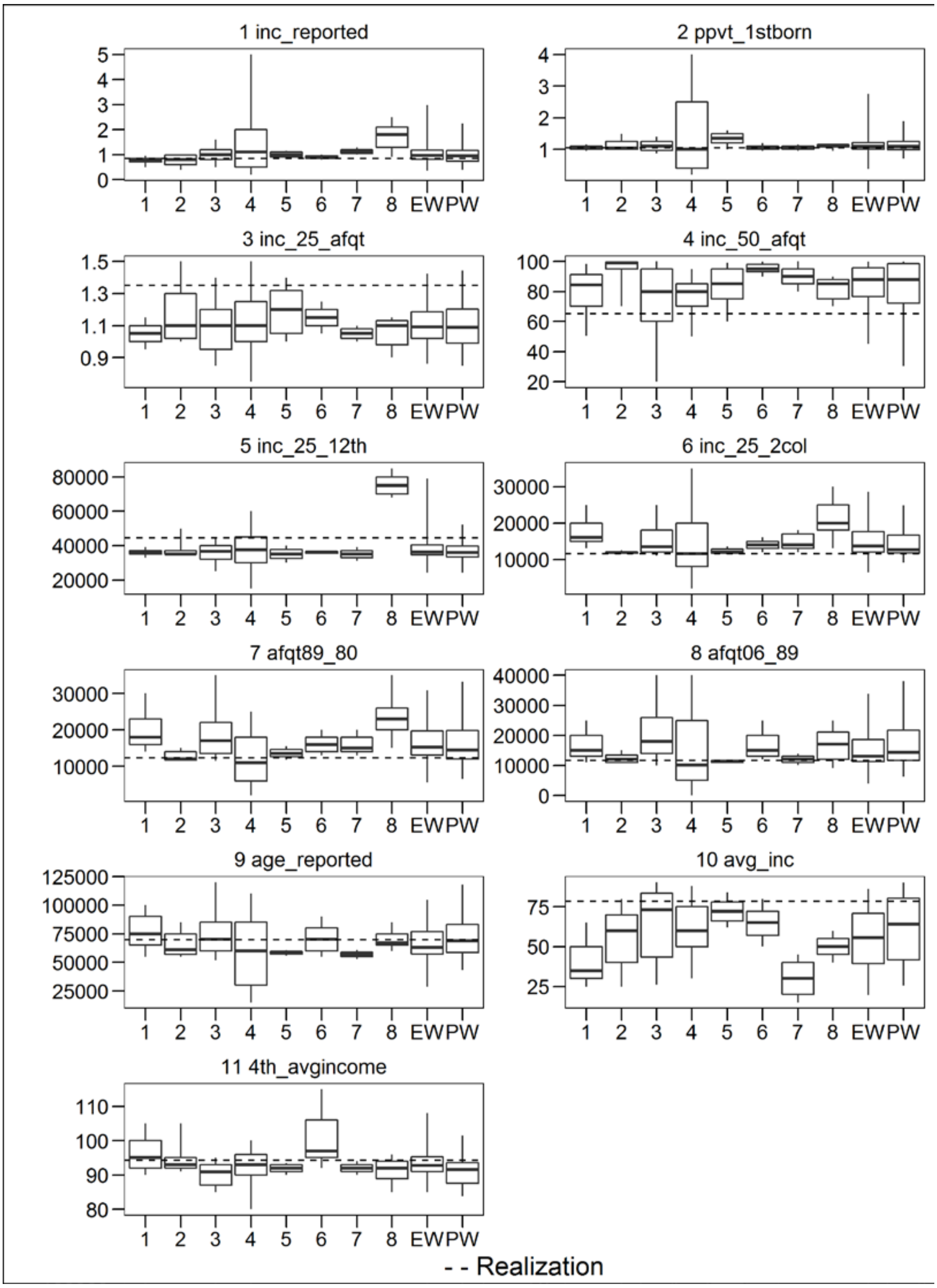
Item weights are potentially more attractive as they allow an expert to up- or down- weight him/herself for individual items according to how much (s)he feels (s)he knows about that item. "knowing less" means choosing quantiles further apart and lowering the information score for that item. Of course, good performance of item weights requires that experts can perform this up- down weighting successfully. Anecdotal evidence suggests that item weights improve over global weights as the experts receive more training in probabilistic assessment. Both item and global weights can be pithily described as optimal weights under a strictly proper scoring rule constraint. In both global and item weights calibration dominates over information, information serves to modulate between more or less equally well calibrated experts.

Since any combination of expert distributions yields assessments for the seed variables, any combination can be evaluated on the seed variables. In particular, we can compute the calibration and the information of any proposed decisionmaker. We should hope that the decisionmaker would perform better than the result of simple averaging, called the *equal weight DM*, and we should also hope that the proposed DM is not worse than the best expert in the panel. The global and item weight DM's discussed above (optimized or not) are *Performance based DM's.* In general the optimized global weight DM is used, unless the optimized item weight DM is markedly superior.

Appendix B. Ranges and Cross-Validation

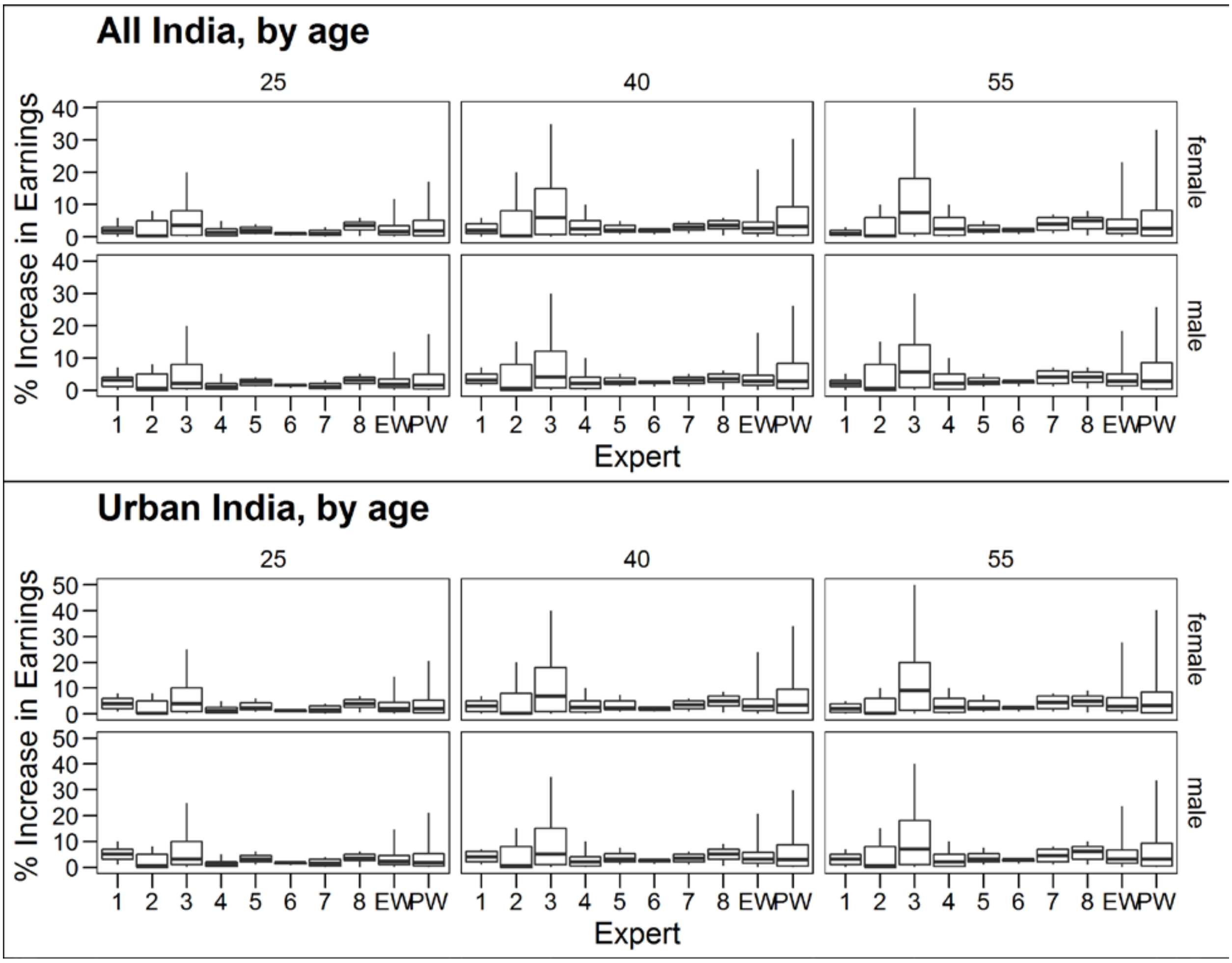
Range Graphs

Figure A1. Range Graphs for Experts and Decisionmakers on the Calibration Questions



Note: The boxplots show the 5, 25, 50, 75, and 95 percentiles. EW is the equal-weight DM and PW is the performance-weight DM with item weights.

Figure A2. Range Graphs for Experts and Decisionmakers on the Variables of Interest



Note: The boxplots show the 5, 25, 50, 75, and 95 percentiles. EW is the equal-weight DM and PW is the performance-weight DM with item weights.

Cross Validation

The results reported in section 3 are in-sample; that is, performance of the PW is scored on the same set of calibration variables used to initialize the model. Out-of-sample validation requires using one set of variables to initialize the model and a different set to validate performance. Out-of-sample validation using the variables of interest is seldom feasible as indeed we resort to expert judgment because values for the variables of interest are not available. Cross validation is a form of out-of-sample validation, whereby a subset of the calibration variables (training set) is used to initialize the model, and the complementary set (test set) is used to score performance. A small training set impedes the model's ability to resolve expert performance and produces combinations that little resemble that of the whole study. A small test set impairs the ability to discriminate between the performance of PW and EW. Colson and Cooke (2017) study all 33 studies conducted between 2006 and March 2015 and found that a training set based on 80 percent of the calibration variables best balances the competing demands of expert resolution and DM resolution. The cross-validation software graciously provided by Lt Col. Justin Eggstaff (Eggstaff et al. 2014) does not presently perform cross validation on item specific weights (the PW combination used here).

Whereas in the paper PW referred to the item weights-based decisionmaker, PW refers to the global weights-based decisionmaker throughout Appendix A. Details on the differences between these two decisionmakers are in Appendix B.

Figure A3 shows the ratio of combined scores, or un-normalized weights, for the PW and EW for all training sets, starting with the training sets of size 1 at the left and proceeding to training sets of size 10 at the right. With 11 calibration variables, a training set of size 9 contains 80 percent of the calibration variables. With this ordering, training sets of at least 80 percent size begin at index 1881. In Figure A3 we see that PW under-performs relative to EW until index about 1700. Beyond this number PW is able to resolve expert performance and the ratio PW/EW is mostly greater than 1.

Figure A3. Ratios of Combined Scores PW / EW for All Training Sets

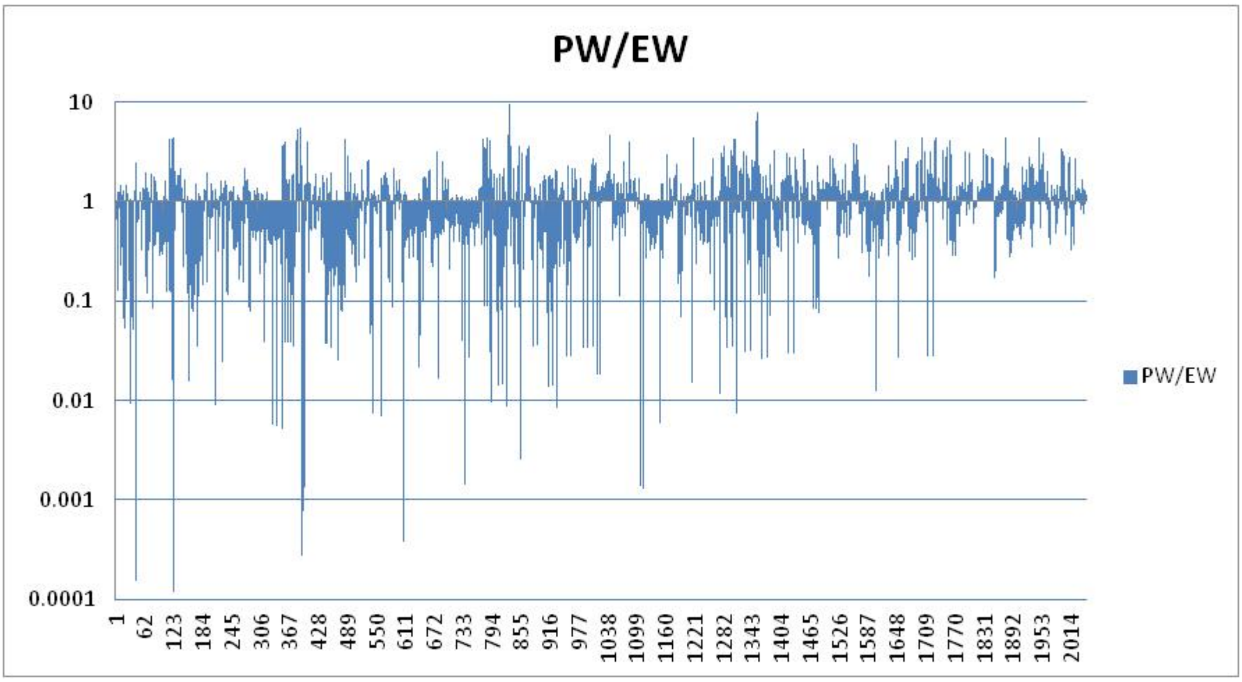


Figure A4 shows the separate combined scores for PW and EW averaged over same sized training sets. Size 9 corresponds to 80 percent of the calibration set. According to Colson and Cooke (2017) this is the preferred size for judging out-of-sample performance.

Figure A4. Combined Scores of PW and EW Averaged over Same-Sized Training Sets

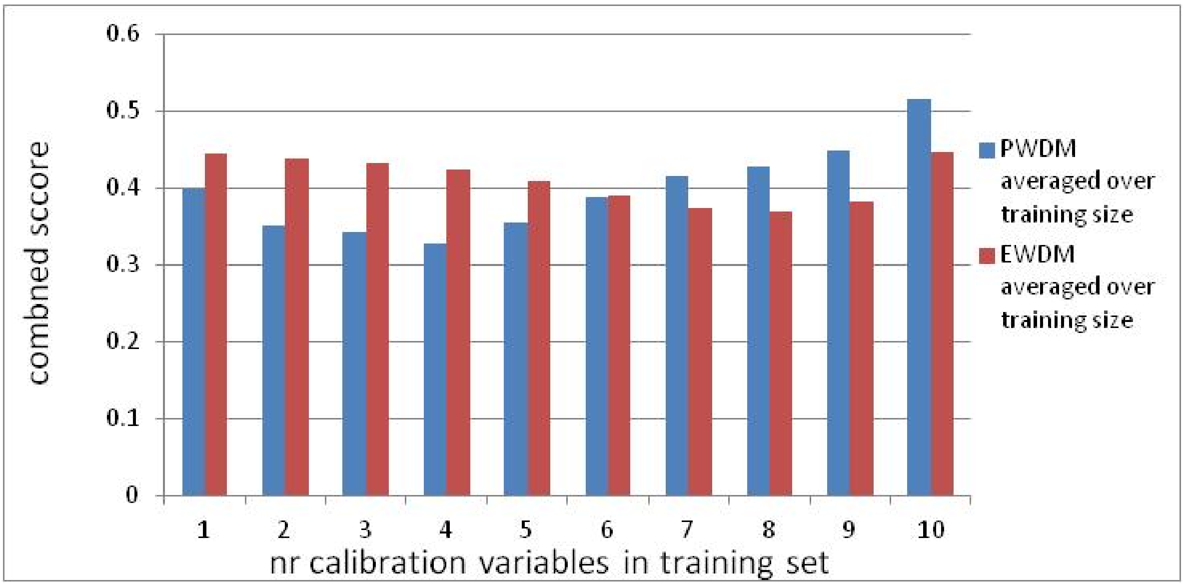
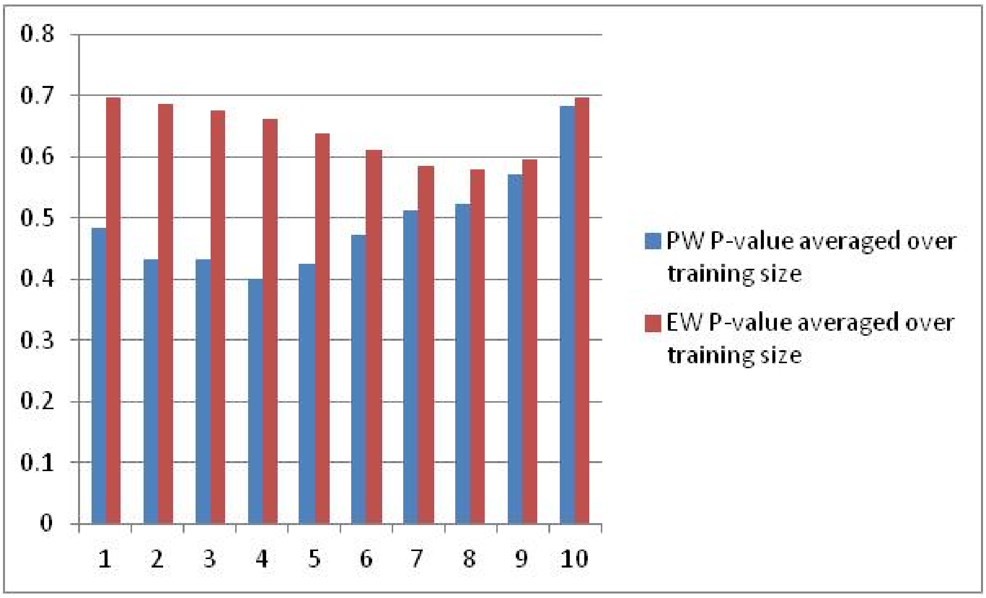
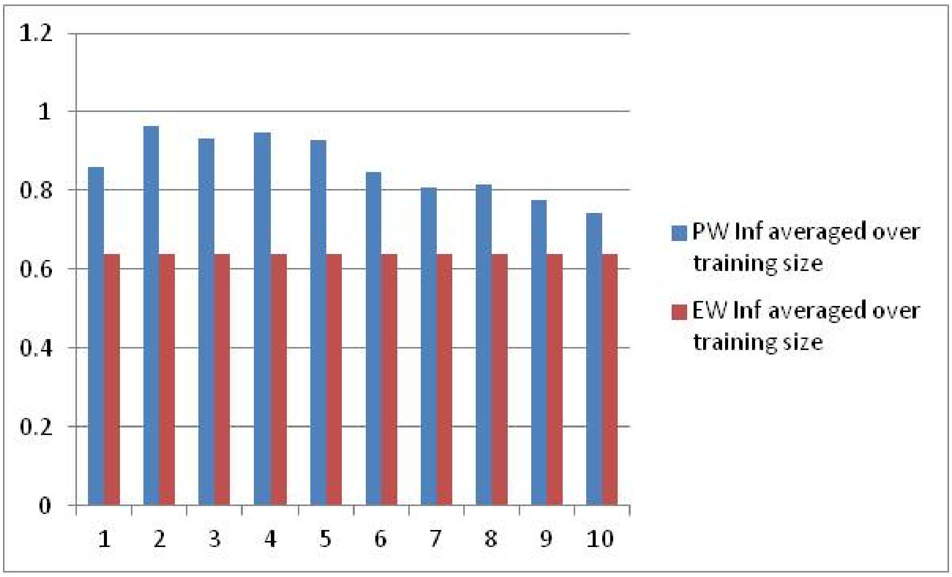


Figure A5 shows the components of the combined score, i.e. statistical accuracy (P-value) and informativeness, for PW and EW, averaged over same-sized training sets. Note that all the statistical accuracy scores are high and their differences are negligible.

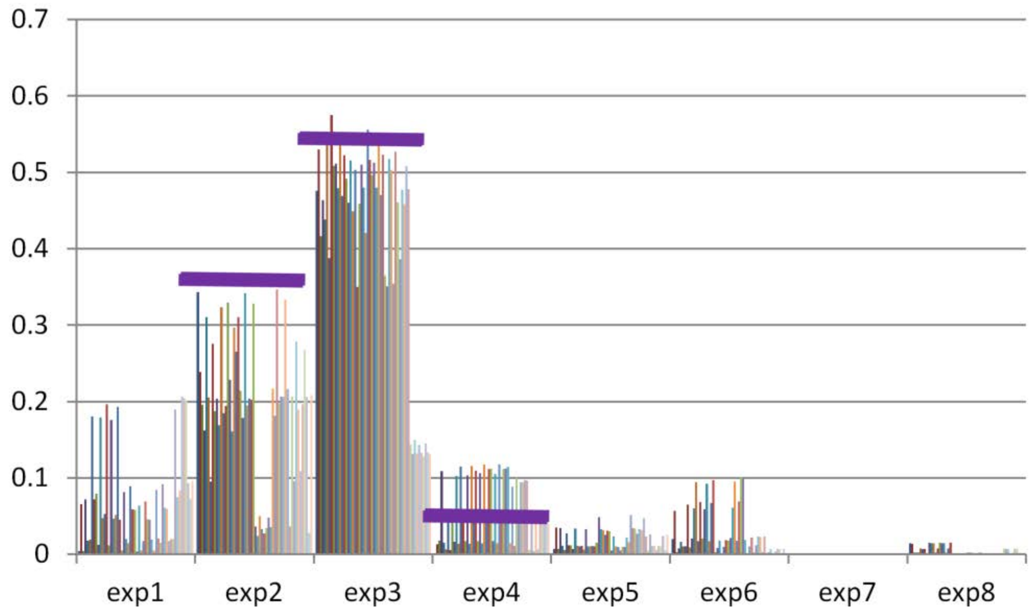
Figure A5. Statistical Accuracy (left) and Informativeness (right) of PW and EW Averaged over Same-Sized Training Sets.





The PW used in this cross validation study (i.e., the global weights-based decisionmaker) assigned non-zero weights to experts 2 (0.37), 3 (0.57) and 4 (0.06). Figure 5 shows the normalized weights assigned to all experts for all training sets of size 9. The purple bars indicate the weights using all 11 calibration variables. In as much as the weights based on 9 training variables resemble the weights based on all calibration variables, the out-of-sample performance at 80 percent training set size should predict out-of-sample performance of the PW for the whole study.

Figure 5. Expert Normalized Weights for All Training Sets of Size 9



Note: Purple bars give the weights based on all 11 calibration variables in which only experts 2, 3 and 4 were weighted.

The out-of-sample predictive performance of the performance based decisionmaker, relative to the equal weight decisionmaker, has been validated to the extent possible without actually observing the variables of interest.

Appendix C. Elicitation Protocol

Expert Judgment: Evaluation of the Effect of IQ on Earnings

Introduction

Structured expert judgment is an accepted tool in risk analysis for supplementing data shortfalls and quantifying uncertainty. Experts quantify uncertainty with regard to variables of interest and calibration variables from the subject area. Experts are treated as statistical hypotheses and combined so as to maximize the statistical accuracy and informativeness of the "decisionmaker." Expert names are preserved to enable competent peer review, but are not associated with responses in any open documentation. Expert reasoning is captured during the elicitation and becomes, where indicated, part of the published record. Elicitation is done by specifying percentiles of uncertain quantities, as illustrated below.

Elicitation Format

You are presented with an uncertain quantity:

In the NLSY79 data set (National Longitudinal Survey of Youth 79- Children, 11512 records), the mean year of birth is 1986 (Range = 1970 - 2011). What is the mean year of birth of children who are the fifth child?

What is the mean year of birth of children who are the fifth child?				
5%	***25%***	***50%***	***75%***	***95%***

You are asked to quantify your uncertainty by specifying percentiles of your subjective uncertainty:

- The 50 percentile is that number for which you judge the chance ½ that the true value is above or below
- The 25 percentile is that number for which the chance that the true value is BELOW is ¼, and the chance that the true value is ABOVE is ¾.
- The 5 percentile is that number for which the chance that the true value is BELOW IS 0.05 and the chance that the true value is ABOVE is 0.95.
- Etc.
- It is always true that 5%-tile < 25%-tile < 50%-tile < 75%-tile < 95%-tile.

Suppose you respond as shown below:

What is the mean year of birth of children who are the fifth child?				
1980	***1983***	***1985***	***1988***	***1990***
5%	***25%***	***50%***	***75%***	***95%***

This means that the true value is equally likely to be above or below 1985, there is a 50% chance that it lies between 1983 and 1988, and a 90 percent chance that it lies between 1980 and 1990.

A *good probability assessor* is one whose assessments capture the true values with the long run correct relative frequencies (**statistically accurate)**, with distributions which are as narrow as possible (**informative).** Informativeness is gauged by 'how far apart the percentiles are' relative to an appropriate background (Shannon relative information).

Measuring statistical accuracy requires the true values for a set of assessments. The true value for the above question is 1991.32. It falls above the 95 percentile. If the expert's assessments are *statistically accurate*, then in the long run, 5 percent of the answers should fall within this inter-percentile interval. Similarly, 90 percent of the answers should fall between the 5 percentile and the 95 percentile, etc.

In gauging overall performance, statistical accuracy is more important than informativeness. Non-informative but statistically accurate assessments are useful, as they sensitize us to how large the uncertainties may be; highly informative but statistically very inaccurate assessments are not useful. Do not shy away from wide distributions if that reflects your real uncertainty.

If you have little knowledge about an item, this does NOT disqualify you as an uncertainty assessor. Knowing little means that your percentiles should be 'far apart'. If other experts are more informative, without sacrificing accuracy, then they will most influence the decisionmaker. If there are no statistically accurate experts with more informative assessments, then the uninformative assessments accurately depict the uncertainty. That in itself is VERY important information.

The **variables of interest** concern an ideal experiment involving fully randomized trials. Like thought experiments in physics, these exercises focus attention on unobservable causal relations.

Training

Below are a few practice elicitations to familiarize you with the format and performance concepts.

A) In what percentage of the 11512 records in the National Longitudinal Survey of Youth79-Children (NLSY79-C) is the earned income in 2008 NOT reported?
__________ __________ __________ __________ __________ ***5% 25% 50% 75% 95%***

B) In what percentage of the 11512 records in NLSY79-C is the Armed Forces Qualification Test (AFQT) score for the mother in 1980 NOT reported?
__________ __________ __________ __________ __________ ***5% 25% 50% 75% 95%***

C) Of the 11512 records in NLSY79-C, how many are 4th born?
__________ __________ __________ __________ __________ ***5% 25% 50% 75% 95%***

Elicitation

Calibration Questions

1. In the NLSY79 representative sample (without the Hispanic and African-American oversamples), there were initially 6111 subjects of whom 5751 have AFQT scores. What percent of these 5751 subjects have data for earned income in 2008?
__________ __________ __________ __________ __________ ***5% 25% 50% 75% 95%***
2. In the NLSY79-Children data the average observed Peabody Picture Vocabulary Test (PPVT) mean score is 90.660.[8] What is the average observed PPVT score among first-borns?
__________ __________ __________ __________ __________ ***5% 25% 50% 75% 95%***

[8] The NLSY 79 reports that "The Peabody Picture Vocabulary Test, revised edition (PPVT) 'measures an individual's receptive (hearing) vocabulary for Standard American English and provides, at the same time, a quick estimate of verbal ability or scholastic aptitude'". It goes on to add, "The PPVT was designed for use with individuals aged 2½ to 40 years. The English language version of the PPVT consists of 175 vocabulary items of generally increasing difficulty. The child listens to a word uttered by the interviewer and then selects one of four pictures that best describes the word's meaning." See

https://www.nlsinfo.org/content/cohorts/nlsy79-children/topical-guide/assessments/peabody-picture-vocabulary-test-revised.

3. Of the 1137 mothers in the NLSY79-Children data who were born in 1958 with AFQT scores and earned income data[9] at age 25, the average earned income at age 25 was $10,094.

What was the average earned income at age 25 among those whose AFQT score, tested at age 22, was above the average of this group?

___________	___________	___________	___________	___________
5%	*25%*	*50%*	*75%*	*95%*

4. Of the 628 mothers in the NLSY79-Children data who were born in 1958 with AFQT scores and earned income data at age 50, the average earned income at age 50 was $52,759.

What was the average earned income at age 50 among those whose AFQT score, tested at age 22, was above the average of this group?

___________	___________	___________	___________	___________
5%	*25%*	*50%*	*75%*	*95%*

5. Of the 434 mothers in the NLSY79-Children data who were born in 1958 with years of schooling and earned income data at age 25, the average earned income at age 25 was $11,506.

What was the average earned income at age 25 among those mothers who completed (at least) the 12th grade?

___________	___________	___________	___________	___________
5%	*25%*	*50%*	*75%*	*95%*

[9] Earned income is total income from wages, salary, commissions, and tips from all jobs. It does not include dividends or profit from businesses or farms. It is individual income.

6. Of the 434 mothers in the NLSY79-Children data who were born in 1958 with years of schooling and earned income data at age 25, the average earned income at age 25 was $11,506.

What was the average earnings at age 25 among those who completed (at least) 2 years of college?

5%	25%	50%	75%	95%

7. In the NLSY79-Children data, what is the ratio of the average mothers' AFQT scores for children born in 1989 / 1980?

5%	25%	50%	75%	95%

8. In the NLSY79-Children dataset, what is the ratio of the average mothers' AFQT scores for children born in 2006 / 1989?

5%	25%	50%	75%	95%

9. The Panel Study of Income Dynamics Child Development Supplement (PSID-C) dataset has 3563 children.

In what percentage of the PSID-C records is mother's age at birth reported?

5%	25%	50%	75%	95%

10. In the PSID-C dataset, the average of the reported family income (97) is $35,100. What is the average among records in which birth order is reported?

5%	25%	50%	75%	95%

11. In the PSID-C dataset, consider the average reported family income for children for whom birth order is recorded. What is the ratio of the average family income at time of 4th birth relative to the above average family income?

___________	___________	___________	___________	___________
5%	*25%*	*50%*	*75%*	*95%*

Variables of Interest

Questions 12- concern earned income at different ages, by gender, in all of India and in urban regions, as shown below:

Region	Gender	Age
		25
		40
		55
		25
		40
		55
		25
		40
		55
		25
		40
		55

Improved infant nutrition, improved maternal health, and improved early childhood education have all been linked to increases in IQ. In each of the 12 questions you are asked to consider a policy intervention that raises IQ at age 12 by 2 IQ points, in all urban settings and throughout India, for boys and girls. The intervention that is in urban settings does not affect boys and girls residing outside urban areas who might move later to urban settings, while the intervention that is throughout India affects all boys and girls regardless of where in India they live.

In each case imagine an ideal measurement with treatment (policy intervention) and control groups of very large size, and individuals assigned randomly to such groups and then selected entirely at random. Please assess your uncertainty in the percentage change in annual earned income of the treatment group relative to the control group in the cases described below.

By earned income we mean:

- all wages, salaries, tips and fringe benefits
- all income from self-employment
- all business income, including any dividends from businesses that the individual may have created.

General background data on India are available in the online briefing book.

School enrollment, attendance, and literacy data:
https://virginia.box.com/s/43imtt988wyykmqhifs9zg9d86i0lime
Labor market data: https://virginia.box.com/s/d4q9vqzcros5u39n4091p7nu53vk5kcs

All India

% increase in earnings = [earnings with IQ+2 – earnings with IQ] / earnings with IQ

15. % increase in earnings at age 25, males

______ ______ ______ ______ ______

5% 25% 50% 75% 95%

16. % increase in earnings at age 25, females

______ ______ ______ ______ ______

5% 25% 50% 75% 95%

17. % increase in earnings at age 40, males

______ ______ ______ ______ ______

5% 25% 50% 75% 95%

18. % increase in earnings at age 40, females

______ ______ ______ ______ ______

5% 25% 50% 75% 95%

19. % increase in earnings at age 55, males

______ ______ ______ ______ ______

5% 25% 50% 75% 95%

20. % increase in earnings at age 55, females

______ ______ ______ ______ ______

5% 25% 50% 75% 95%

Urban India

% increase in earnings = [earnings with IQ+2 – earnings with IQ] / earnings with IQ

21. % increase in earnings at age 25, males

_________ _________ _________ _________ _________

5% 25% 50% 75% 95%

22. % increase in earnings at age 25, females

_________ _________ _________ _________ _________

5% 25% 50% 75% 95%

23. % increase in earnings at age 40, males

_________ _________ _________ _________ _________

5% 25% 50% 75% 95%

24. % increase in earnings at age 40, females

_________ _________ _________ _________ _________

5% 25% 50% 75% 95%

25. % increase in earnings at age 55, males

_________ _________ _________ _________ _________

5% 25% 50% 75% 95%

26. % increase in earnings at age 55, females

_________ _________ _________ _________ _________

5% 25% 50% 75% 95%

Answers to Training Questions

A) 52.3490462 %

B)6.0854485 %

C) 657

Acronyms and Explanations

- Earned income is total income from wages, salary, commissions, and tips from all jobs. It does not include dividends or profit from businesses or farms.
- The PSID data are taken from Rothstein, REStat, July, 2014.

- PPVT: Peabody Picture Vocabulary Test
- NLSY: National Longitudinal Study of Youth
- WI Wechsler Intelligence Test 'digit span recall used in NLSY'
- WJLetterwordscores: Wood Johnson letter word score
- WJappliedprobscores
- bio_mom_id: Biological Mother's ID
- AFQT: Armed Forces Qualification Test
- PIAT: Peabody Individual Achievement Test

Appendix D. Experts' Rationales

Table B1. Experts' Rationales for the Variables of Interest

Expert No.	General notes	Impact in women vs. men	Impact at different ages	Impact in urban vs. rural India
1	The largest impact of IQ on earnings in the "all India" intervention will be for males at age 40.	The impact of IQ on earnings will be weaker for women. Women generally have lower income than men in India, so the returns to education and skills will be compressed. This could be due to discrimination or other factors. Women work in different sectors/industries in India where the return to skill is lower.	By age 55, people are beginning to retire. Older workers tend to be in agriculture or other sectors where returns to skill are lower, so the overall returns to IQ are lower at age 55 than at age 40. The difference between males and females is smaller at age 25 as the wage gap is smaller for younger workers.	Returns to IQ should be higher in urban India than in all India because the industry composition in urban areas (where lots of people are employed in software, IT, and other high-tech industries) has higher returns to skill. In urban India, returns to IQ will be higher for younger workers, as they are more likely to work in the industries with higher returns to skill. Returns to skill will decrease with age.
2	There is no definitive evidence that IQ has an impact on earnings, in India, the US, or elsewhere. There is a strong relationship between the two, but it is impossible to experimentally manipulate IQ while holding everything else constant. Most of the "natural experiments" we rely on for evidence might actually be shifting the relationship between IQ and measurable outcomes such as test scores and productivity, rather than actually changing IQ. Impact in India vs. the US: The impact of IQ on earnings would be smaller in India than the US. The US has a lot of free or	In India the impact of IQ on earnings is likely lower for girls, who are less likely to work for pay. However, there are examples in which increasing girls' schooling drastically changed the dynamics of a village, so my upper tail estimates of the impact of IQ on earnings for women are especially high, especially at age 40 when I have more uncertainty about how many might be in the labor market. The impact depends on how IQ is related to the opportunity structure, such as opportunities to work for pay and the availability of		There could be different changes in educational opportunities in rural vs. urban India, but the improvement could be in urban or rural settings. Overall, my uncertainty surrounding the impact of IQ on earnings isn't different for urban India compared to all India.

	subsidized educational opportunities, whereas in India I believe a family's financial situation is more likely to determine educational opportunities A child with a given IQ might have a greater chance to develop that potential in the US, and the link between IQ and earnings would also be stronger in the US. The Indian labor market has more competition than the US, and India has a lot more nutritional deprivation as a general constraint on the population.	scholarships supporting higher education, either of which could increase the monetary returns to IQ. Fewer women are still working at age 55, so if the effect of IQ on earnings was large, it would be dampened by age 55.		
3	The questions are talking about raising IQ 2 points, or .15 standard deviations. Increasing IQ 1 standard deviation is associated with a 10% increase in earnings in the US. 1 year additional schooling has a similar effect to 1 standard deviation increase in IQ. India is developing and has the potential for large growth, and its labor market has imperfections, so the returns to IQ are largely uncertain. Impact in India vs. US: In India, part of what determines the returns to IQ is that returns are lower in low tech fields. However, increasing IQ probably increases school participation and achievement, and therefore	Labor participation is lower for women than men in India. Education experience of women is also lower. The direct effect of IQ on women at work is small, but there is a big indirect impact through increasing schooling and education, which will lead to an increase in labor market participation.	Returns to IQ increase with age. The upper bound for the impact of IQ on earnings will be higher for women than men at age 40 and 55. The constraints for women in the labour market could be loosened for high ability women such that there is a big impact on earnings for women with increased ability.	Urban areas have more diverse job options, which enables better matching of the diversity in skills than is possible in rural areas. Thus, the impact of IQ on earnings will be higher in urban areas than rural areas.

	increases labor market participation.			
4	A lot of the returns to IQ is achieved through education. The impact of IQ itself on earnings is more limited. The association between IQ and increased schooling is the main channel for impacting earnings. 2 IQ points is not that much.	Literature shows that returns to education are higher for women than men.	In India, IQ has a larger impact on earnings at age 40 than 25. The age-earnings curve starts to flatten after age 40, though. If we were talking about the US, the impact at age 25 would be higher due to more efficient returns to increased schooling.	
5	• India has fairly high labor market participation for males, but not for females. • 2 IQ points won't have a large impact on earnings. • The service sector in India will continue to expand, creating more potential for growth.	• IQ has a larger impact on education for women than men. • Women are less likely to be working than men in India, so this compresses the potential impact of IQ on earnings.	There's not much variation in wages at age 25. With age, there are more workers in the upper tail, and the earnings distribution becomes more skewed.	Returns to education and returns to knowing English are higher in urban areas than rural areas. The point estimates (i.e., medians) won't change much, but the upper bounds increase in urban areas relative to urban because of the potential impact of better returns to education/knowing English.
6	The impact of IQ on earnings in India depends a lot on the impact IQ will have on education. How IQ affects educational attainment in India is a key uncertainty in these questions. In the US, most of the impact of IQ on earnings is through education. The correlation between IQ and education is lower in India than the US, but it is uncertain. Generally, the highest returns to education for IQ are for tertiary education, not secondary. Returns to skill is pretty high in India, and it is increasing recently. However, agriculture is a big industry in India, and the correlation between IQ and	The impact of IQ on education is even lower for girls than boys in India. Women have less strong attachment to work and tend to work more in agriculture, where returns to skill are low. Women will have increased labor force participation (at different ages) due to increased IQ.	At age 25, the increase in earnings associated with IQ would just be due to increased education. In India, though, people enter the labor market earlier than in the US. At age 25 most males will have 10 years of work experience in India. In US, returns to IQ increases with age. It's lower when young and increases. In the US, between 25 and 55, return to IQ doubles from 8 to 16% From 40 to 55 it increases 12-14%. The responses for 55-year olds are conservative	In general, returns to skill is higher in urban settings. Females work less in urban areas than rural areas in India now, but if they work, they have normal jobs. The difference between male and female workers will be smaller in urban settings than rural settings.

	earnings is lower in agriculture than other industries. Returns to skill generally increases with age and is higher in urban settings than rural. Returns to skill is greater for men than women. The impact of IQ on earnings in India is likely to be smaller than in the US because returns to skill is lower there, but there is uncertainty.		because of the uncertainty about life cycle dynamics in India. In the US, IQ is correlated with many other factors, so the impact of these correlates is uncertain here.	
7	Virtually all evidence on the link between IQ and earnings is from developed countries, mostly the US. The extent to which you can extrapolate that to India is hugely uncertain. There are currently enormous obstacles to trade in India. It's a tough market for entrepreneurs and investors. The current Prime Minister is working to improve this and modernize the Indian economy, which will enable the labor market to better reward cognitive skills, but the extent to which these efforts have been successful is uncertain.	At age 25, workers are still young. The extent to which cognitive skills will be rewarded depends on how well the labor markets are organized. As workers age, the impact of IQ on earnings will increase.	In the US, the payoff from increased cognitive skills is higher for women than for men. This could also hold in India, but India is also a more patriarchal society, which could mean that cognitive skills have less of an impact on earnings for women than men. The net of these two uncertainties is that the overall uncertainty distribution about the impact of IQ on earnings is the same for women and men in India. The role of women is very traditional in India, and while there is pressure for that the change, the extent to which that has happened is unclear.	Urban areas have less agriculture and more formal labor markets, which could be better organized to reward skills. Urban areas also have more opportunities for entrepreneurship. Thus, the impact of IQ on earnings is likely to be greater in urban India than all of India. The extent to which increased IQ will produce increased earnings in rural areas depends on the extent to which there is access to new technology to cultivate crops. Research has shown that in areas without technology, cognitive skills are not rewarded by the labor markets in rural areas.
8	It's hard to think about an intervention that just changes IQ. Normally these interventions affect IQ and something else, and it's tough to tease out the impact of the something else versus the impact in the IQ change.	Men and women have different return to skills. An increase in IQ for women will lead to a bigger impact on their education outcomes than the same IQ increase for men.		Individuals in urban areas may be more driven and have more access to jobs that use a lot of cognitive skills. Again, there's a lot of uncertainty about the impact of an intervention that impacts only IQ, as most

	For a lot of these types of interventions, the impact of the associated change in non-cognitive skills is much higher than the impact of IQ itself. For example, changes in non-cognitive skills are shown to have a big impact on jailing, which in turn has a big impact on earnings.	It's harder to raise the IQ of females than males, but that's an issue that isn't captured in this hypothetical experiment, where we're assuming everyone, male and female, increases IQ 2 points. There would be a huge variance in these outcomes. The complex dynamics of life contribute to the different impact and its uncertainty. At age 40, women in India are doing more chores and working in the household and working less in the labor force. At age 55, however, women could return to the labour market.		interventions impact IQ and other factors. That makes the impact of this intervention very uncertain.