DISCUSSION PAPER

July 2016 ■ RFF DP 16-28

How Does Breastfeeding Affect IQ? Applying the Classical Model of Structured Expert Judgment

Abigail Colson, Roger Cooke, and Randall Lutter

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

RESOURCES
FOR THE FUTURE

How Does Breastfeeding Affect IQ? Applying the Classical Model of Structured Expert Judgment

Abigail Colson, Roger Cooke, and Randall Lutter

Abstract

We use the classical model, a method for structured expert judgment (SEJ), to study the effects of breastfeeding on IQ. Data on the link between breastfeeding and IQ are available, e.g., the US National Longitudinal Study of Youth, however, questions about data quality and confounding mean properly interpreting the data is not straightforward, and expert opinions diverge regarding the efficacy of breastfeeding for enhancing IQ in Western cultures. In developing countries, differing demographics and social values combined with scarcity of data render structured expert judgment an attractive method to provide policymakers with quantitative information. We find that early breastfeeding generates most of the IQ gains from full compliance with World Health Organization guidelines, and IQ gains from breastfeeding may be larger in India than the United States.

Contents

How Does Breastfeeding Affect IQ? Applying the Classical Model of Structured Expert Judgment

Abigail Colson, Roger Cooke, and Randall Lutter*

1. Introduction

The long-term effects of nutritional interventions are notoriously difficult to assess in well-controlled, randomized, blinded trials. Random assignment by village (e.g., Hoddinott et al. 2008) or by hospital (e.g., Kramer et al. 2001) is one solution, although the costs of lengthy multiyear follow-up and sample sizes sufficient to evaluate modest effects make such studies both time-consuming and costly. Researchers have estimated long-term effects of malnutrition using unique events, such as the Dutch Hunger Winter of 1944–45, but transferring the estimates derived from such events to practical policy interventions can be challenging (e.g., Almond and Currie 2011; Currie 2011). Not surprisingly, therefore, conventional longitudinal studies frequently underlie most policy recommendations, although confounding always poses threats to findings based on such data (see, e.g., Horta and Victora 2013).

Evaluating the long-term effects of breastfeeding exemplifies the challenges posed by using nonrandomized longitudinal data sets. A growing number of long-term studies of breastfeeding show long-term effects on cognitive performance (commonly called IQ), using a variety of measures of breastfeeding behavior and cognitive performance (Horta and Victora 2013; Horta et al. 2015; Victora et al. 2016), although concerns about confounding persist (e.g., Walfisch et al. 2013). The results of meta-analyses of these studies have found effects of breastfeeding on IQ that are “likely” to be causal, in part because a randomized trial (Kramer et al. 2001) reports effects on IQ. These meta-analyses, however, have left some questions unanswered. First, the underlying longitudinal studies use a variety of measures of breastfeeding (including duration of any breastfeeding and duration of exclusive breastfeeding). The meta-analyses typically lump all types of breastfeeding together and do not identify the types of

* Colson is a research associate and PhD candidate at the University of Strathclyde in Glasgow and 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. Lutter is a visiting fellow at Resources for the Future and professor of public policy at the Frank Batten School of Leadership and Public Policy at the University of Virginia. They appreciate financial support from the Bill and Melinda Gates Foundation through the University of Virginia for this work. They may be reached at randall.lutter@virginia.org. 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 funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

breastfeeding most likely to bring about IQ improvements. Thus they do not address whether breastfeeding that is of shorter duration but more intensive might raise or lower IQ. For example, they do not address whether exclusive breastfeeding for six months followed by no breastfeeding would be expected to have greater IQ effects than exclusive breastfeeding for three months followed by six months of some breastfeeding. Second, they are limited by the availability of data and thus do not speak to the expected IQ gains in countries where no longitudinal data sets linking breastfeeding and IQ exist, such as India. Yet the effects of breastfeeding on cognitive performance as measured in conventional IQ tests may differ by country because educational opportunities and quality differ internationally, and IQ test results reflect skills that are partly acquired in classrooms, such as reading, vocabulary, and arithmetic.

Because high-quality data do not exist to resolve these important issues, in this paper we seek to address these questions by applying structured expert judgment (SEJ) methods to evaluate the effects of the duration of exclusive and any breastfeeding on cognitive performance or IQ. We are interested in the impact of breastfeeding in three countries: the United States, India, and China. The SEJ methods used here have been applied in many other settings where repeated controlled experimentation is difficult or impossible, such as food safety (Aspinall et al. 2016; Hald et al. 2016; Havelaar et al. 2008), prion disease (Tyshenko et al. 2011), and risks of invasive species (Wittman et al. 2014a, 2014b; Rothlisberger et al. 2012). SEJ may be seen as complementary to other data aggregation methods, such as meta-analysis.

We report on a SEJ exercise about the effects of breastfeeding on IQ conducted in 2015 and 2016. This exercise involved detailed one-on-one interviews conducted by an author experienced and trained in such interviews. We asked experts to quantify their uncertainty for the questions of interest and calibration questions for which we know answers post hoc. We then use the answers to these calibration questions to form performance-weighted combinations of the experts' responses. We subject the results to robustness checks and an out-of-sample cross-validation procedure. We find that our results are robust in the sense that loss of either one calibration question or one expert is not an issue.

We study several scenarios or cohorts of interest to policy makers and focus on the differences among no breastfeeding, each of two cohorts that reflect moderate breastfeeding, and breastfeeding compliant with the recommendations of the World Health Organization (WHO). In the moderate cohorts, infants are exclusively breastfed for 3 months and partly breastfed from 3 months to 9 months of age or exclusively breastfed for 6 months with no partial breastfeeding after that. In the compliance cohort, infants are exclusively breastfed to 6 months and partly breastfed from age 6 months to 24 months. Between the first moderate breastfeeding cohort and the compliance cohort, the duration of exclusive breastfeeding doubles from 3 to 6 months, and the duration of any breastfeeding increases by a factor of more than 2.5, growing from 9 months to 24 months.

We find that for the United States, the average IQ for the cohort with 3 months exclusive and 9 months any breastfeeding is three points greater than the average IQ for the cohort with no breastfeeding. The cohort with full compliance with WHO recommendations has an average IQ five points greater than the cohort with no breastfeeding. Further, the IQ gains of the first moderate breastfeeding cohort are more than half (3/5) of the gains from full compliance. In India, the first moderate breastfeeding cohort gains five IQ points relative to the no breastfeeding cohort, while the full compliance cohort gains eight IQ points relative to the no breastfeeding cohort. These results suggest that the effects of breastfeeding on IQ are substantial for moderate levels of breastfeeding but still increase with full compliance to the WHO recommendation. They also suggest that effects are larger in India than in the United States, where several longitudinal studies provide direct, if somewhat varied, empirical support.

In Section 2, we introduce the classical model of SEJ, the experts interviewed in this study, and the protocol questions. Results of the expert weighting are in Section 3. Section 4 presents results of a cross-validation analysis. Appendixes contain a reasonably complete exposition of the mathematics behind the scoring and weighting, a list of all peer-reviewed publications given to experts electronically prior to the interviews as background material, the full elicitation protocol, range graphs summarizing all experts’ responses, and notes on the experts’ rationales.

2. Methods

2.1. The Classical Model

In the classical model of SEJ, experts are asked two types of questions: calibration questions and variables of interest. Calibration questions are those with answers that will be known to the study team within the time frame of the study. Variables of interest are the target questions of the expert elicitation exercise. Experts quantify their uncertainty regarding the answers to both types of questions by providing the 5th, 25th, 50th, 75th, and 95th percentiles of their subjective uncertainty distributions for each question. Experts are scored according to their assessments on the calibration questions. The experts’ assessments are then combined according to both equal weights and performance-based weights. A combination of expert assessments is called a “decision maker.” Full details of the classical model’s procedures for scoring and weighting the expert assessments are provided in Appendix A.

2.2. Experts

We identified experts by first identifying papers cited in recent meta-analyses, including Horta et al. (2007, 2015), Horta and Victora (2013), and Walfisch et al. (2013). We also conducted literature searches using Google Scholar for peer-reviewed papers and key words such as *breastfeeding*, *lactation*, *IQ*, *intelligence*, and *cognitive performance*. We found 16 peer-reviewed articles published in academic journals from 2013 through November 2015 and included copies of these, as well as selected earlier articles, in a briefing book for experts described in Appendix B. The selected additional studies did not all necessarily employ internal comparison groups, measure cognition using standard tests, focus on children older than one year, or adjust for stimulation or interaction with the child. We were careful to include original peer-reviewed studies that reported no statistically significant effects of breastfeeding on IQ. From the authors of papers listed in these meta-analyses and our own literature searches, we identified experts by applying the following criteria:

- authorship of multiple papers about effects of breastfeeding on IQ or cognitive performance, ideally using different data sets (and not multiple papers of the same cohort observed at different ages);
- seniority, using publicly available information on academic titles;
- prominence of journals where research was published; and
- use of IQ or measures of cognitive performance and not outcome measures focused only on early child development.

We concentrated on authors of later papers, believing these authors were more likely to be familiar with the recent literature. We sent out 12 emails inviting the experts we identified to participate in our study. One author of a paper skeptical of the view that breastfeeding duration increases IQ declined to participate despite repeated efforts at recruitment.

Seven experts participated in the study (listed here in alphabetical order):

- Mandy Brown Belfort, Brigham and Women’s Hospital
- Brian Boutwell, College for Public Health & Social Justice, Saint Louis University
- Goeff Der, Institute of Health and Wellbeing, University of Glasgow
- Jordi Julvez Calvo, Centre for Research in Environmental Epidemiology (CREAL)
- Michael Kramer, Faculty of Medicine, McGill University
- Donna Rothstein, US Bureau of Labor Statistics
- Cesar Victora, Federal University of Pelotas

After agreeing to participate, but before the interview, each expert received access to the briefing book mentioned above. One author conducted one-on-one interviews with each of the seven experts using Zoom, a web-based video conferencing platform. During the scheduled interview, the author explained the motivation for using expert judgment and introduced the classical model of SEJ. Experts answered three practice questions to ensure they understood the method.

2.3. Variables

The elicitation protocol included 11 calibration variables (Table 1) and 12 variables of interest (see Appendix C for the full text). The variables of interest asked for the average Wechsler Intelligence Scale for Children, Revised (WISC) score at age 10 for children in each of four cohorts (Table 2) in the United States, India, and China. The WISC score is referred to as IQ throughout this paper.

3. Results

3.1. Expert Scores

Of the seven participating experts, three (experts 2, 5, and 6) were statistically accurate, meaning their calibration scores, (the p-values at which the hypothesis that the expert is statistically accurate would be falsely rejected), are above the traditional threshold (5%) for not rejecting statistical hypotheses (Table 3). Compared with other classical model studies, this is a relatively high number: only 11 of the 33 applications conducted between 2006 and 2015 had three or more statistically accurate experts. The equal weight decision maker (*EW*) has a good calibration score, but its information scores are substantially lower than those of the experts. The performance weighted decision maker without optimization (*PW*noOpt) is slightly more accurate statistically than *EW* (i.e., it has a higher calibration score), but it also has comparably low informativeness. These differences in informativeness are notable, as visible in the range graphs provided in Appendix D. The performance weighted decision maker with optimization (*PW*), is better with regards to both calibration and information. In this case, *PW* coincides with expert 6; all other experts are unweighted in the optimized combination. Concentration of weight in one expert occurs in roughly one third of the applications. The last two columns in Table 3 show the relative information of each expert and decision maker with respect to the equal weight decision maker (*EW*) for all variables and the subset of 11 calibration variables.

Table 1. Calibration Questions

Variable numbers	Variable ID	Question
1	PPVT1st	What is the average Peabody Picture Vocabulary Test (PPVT) score among firstborn children with at least one score?
2	PPVT1stNoBF	What is the average PPVT score among firstborn children who were ever breastfed?
3	PIATMathCorr	What is the correlation among the PPVT and Peabody Individual Achievement Test (PIAT) math scores?
4	PIATReadCorr	What is the correlation among the PPVT and PIAT reading scores?
5	MissingPPVT	In what percentage of the 11,512 records in NLSY79-C is the PPVT never reported?
6	AgeBFEnd	What is the average age in weeks when breastfeeding ended among the 1,583 only children who were breastfed?
7	MomEd1Kid	What is the average years of schooling for the mothers of the 2,900 only children, both breastfed and nonbreastfed?
8	India50	In DHS India, what is the 50th percentile for duration of breastfeeding (months) among children who were breastfed and who were not still breastfeeding?
9	India75	In DHS India, what is the 75th percentile for duration of breastfeeding (months) among children who were breastfed and who were not still breastfeeding?
10	WJScores	In what percentage of completed records is the sum of Woodcock-Johnson scores in 1997 greater than in 2002?
11	PSIDInc	In the US Panel Study of Income Dynamics Child Development Supplement (PSID-C), the average of the reported family incomes (97) is $35,100. What is the average among records in which birth order is reported?

Table 2. Feeding Patterns by Age for Each Cohort

Feeding/food source	Cohort 1	Cohort 2	Cohort 3	Cohort 4
Breastfeeding, exclusive	None	3 months	6 months	6 months
Breastfeeding, any	None	3 to 9 months	None	6 to 24 months
Infant formula, exclusive	6 months	None	None	None
Infant formula, any	6 to15 months	3 to 15 months	6 to 15 months	None
Complementary foods	From 6 months	From 6 months	From 6 months	From 6 months

Table 3. Calibration and Information Scores for Experts and Decision Makers

		Information score			Information relative to *EW*	
Expert ID	Calibration score	All variables	Calibration variables	Combined score	All variables	Calibration variables
1	0.001231	1.483	1.34	0.001649	0.7659	0.8845
2	0.08609	0.7272	0.7368	0.06343	0.4662	0.5146
3	0.004671	1.15	0.951	0.004442	0.5555	0.5141
4	0.002048	0.5076	0.7861	0.00161	0.5858	0.6469
5	0.2306	0.3592	0.4153	0.09578	0.3603	0.3831
6	0.6924	1.031	0.573	0.3968	0.4843	0.3734
7	0.0003015	1.341	1.517	0.0004574	0.8466	0.9427
EW	0.4245	0.3621	0.2942	0.1249	0	0
*PW*noOpt	0.6009	0.5084	0.2945	0.177	0.1779	0.1564
PW	0.6924	1.031	0.573	0.3968	0.4843	0.3734

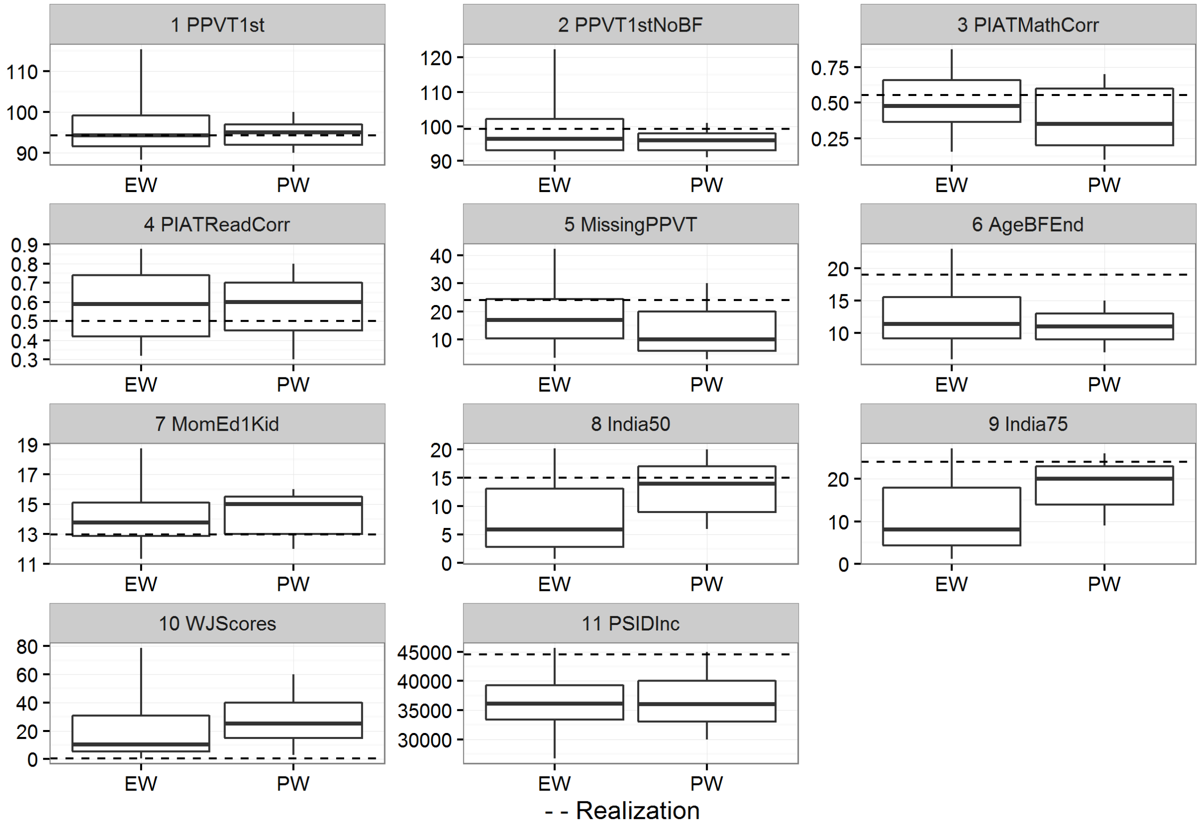
3.2. Decision Maker Assessments

Figures 1 and 2 show the assessments of the *EW* and *PW* decision makers. The difference in the information provided by these two decision makers is easily visible in *PW*'s narrower ranges. The range graphs in Appendix D compare these outputs with the *PW*noOpt combination and the individual experts.

Table 4 provides the values for the 5th, 25th, 50th, 75th, and 95th percentiles for the variables of interest from the *PW* decision maker. The experts thought average IQ would be higher in cohorts with more breastfeeding. The estimated differences in IQ are bigger between the first cohort with moderate breastfeeding (cohort 2) and no breastfeeding (cohort 1) than between compliance with the WHO recommendation for breastfeeding (cohort 4) and moderate breastfeeding in cohort 2. The differences between cohorts are larger in India than in the United States or China.

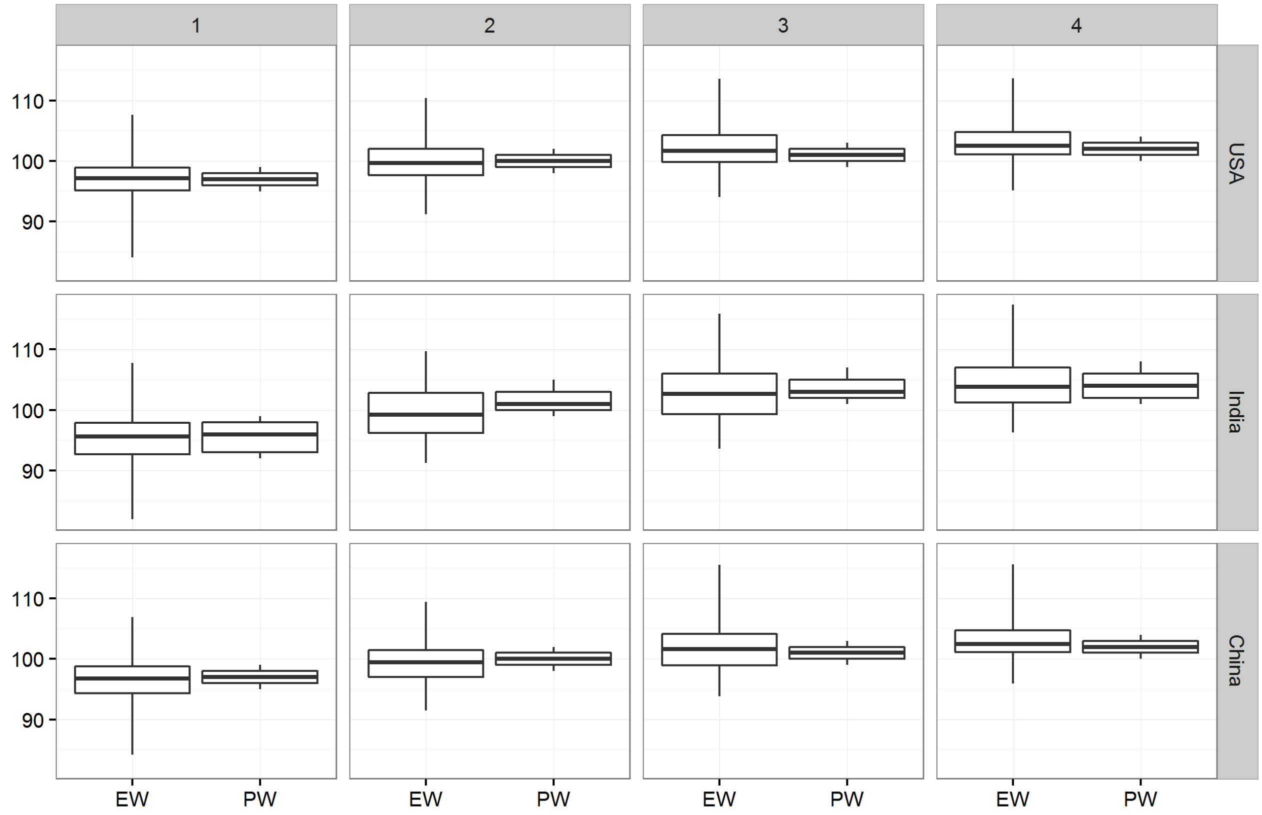
The experts said during the interviews that the current literature does not indicate whether duration of any breastfeeding or duration of exclusive breastfeeding is more important for cognitive development. A few experts said they thought it was unlikely that there would be much cognitive benefit from breastfeeding beyond six months of exclusive breastfeeding, especially in the United States. Experts thought the impact of breastfeeding was likely to be higher in India than the United States or China because breastfeeding can help overcome the developmental setbacks associated with poor prenatal nutrition and low-birth-weight infants. Experts also thought the benefits of breastfeeding could compensate for environmental factors that impact cognitive development in India. Finally, experts thought the health benefits of breastfeeding (for example, fewer infections and less diarrhea) could translate to cognitive benefits in India because infants expending less energy to fight infection would have more energy available for brain development. More information for the rationales provided by the experts is available in Appendix E.

Figure 1. Equal Weight and Performance Weight Decision Maker Assessments for the Calibration Variables



Note: Boxplots show the 5th, 25th, 50th, 75th, and 95th percentiles.

Figure 2. Equal Weight and Performance Weight Decision Maker Assessments for the Variables of Interest



Note: Boxplots show the 5th, 25th, 50th, 75th, and 95th percentiles.

Table 4. Uncertainty Distributions for the Variables of Interest from the Performance Weight Decision Maker

	Average WISC score at age 10				
Cohort	5%	25%	50%	75%	95%
USA cohort 1	95	96	97	98	99
USA cohort 2	98	99	100	101	102
USA cohort 3	99	100	101	102	103
USA cohort 4	100	101	102	103	104
India cohort 1	92	93	96	98	99
India cohort 2	99	100	101	103	105
India cohort 3	101	102	103	105	107
India cohort 4	101	102	104	106	108
China cohort 1	95	96	97	98	99
China cohort 2	98	99	100	101	102
China cohort 3	99	100	101	102	103
China cohort 4	100	101	102	103	104

3.3. Robustness

3.3.1. Robustness on Items

An optimal solution always invites nonrobustness. In computing robustness on items, we remove calibration variables one at a time and compare the scores with the “unperturbed” decision maker (here the optimized *PW*). Comparing the last two columns of Table 5 with the last two columns of Table 3, we see that removing one variable, item 7, perturbs the decision maker more than the differences that exist among the expert themselves. This perturbed decision maker shifts all weight to expert 5. Removing item 6 distributes the weight between experts 2 and 6. In both cases the resulting perturbed decision makers would still return acceptable performance, as determined by the calibration and information scores.

Table 5. Robustness Analysis on Items

		Information score		Information relative to original PW	
Item removed	Calibration score	All variables	Calibration variables	All variables	Calibration variables
1	0.9027	0.7131	0.433	0.3723	0.139
2	0.8444	1.025	0.513	0	0
3	0.5202	1.067	0.6049	0	0
4	0.5202	1.071	0.6146	0	0
5	0.8444	1.053	0.5748	0	0
6	0.8226	0.7025	0.445	0.39	0.1701
7	0.44	0.3652	0.4341	1.348	1.089
8	0.5202	1.064	0.5979	0	0
9	0.8444	1.059	0.5871	0	0
10	0.5845	1.058	0.5851	0	0
11	0.8444	1.06	0.5903	0	0
None	0.6924	1.031	0.573		

3.3.2. Robustness on Experts

We also evaluate robustness with respect to the experts by removing experts one at a time and recomputing the model (Table 6). Unsurprisingly, removal of expert 6 is the only case producing a difference with the unperturbed decision maker, as only expert 6 was weighted in the optimized *PW.* Here again the performance of the perturbed decision maker is still acceptable: the statistical accuracy of the perturbed decision maker remains high, although it is less informative than the unperturbed decision maker. The differences in this case are of the same order as the differences among the experts themselves (again, seen by comparing the final two columns of Table 6 with the final two columns of Table 3).

Table 6. Robustness Analysis on Experts

		Information score		Information relative to original PW	
Expert removed	Statistical accuracy score	All variables	Calibration variables	All variables	Calibration variables
1	0.6924	1.031	0.573	0	0
2	0.6924	1.005	0.518	0	0
3	0.6924	1.031	0.573	0	0
4	0.6924	0.8583	0.5574	0	0
5	0.6924	0.9608	0.481	0	0
6	0.773	0.2651	0.1997	0.9918	0.6094
7	0.6924	1.025	0.5628	0	0
None	0.6924	1.031	0.573		

Overall, we may conclude that robustness in this study against loss of one calibration variable or loss of one expert is not an issue.

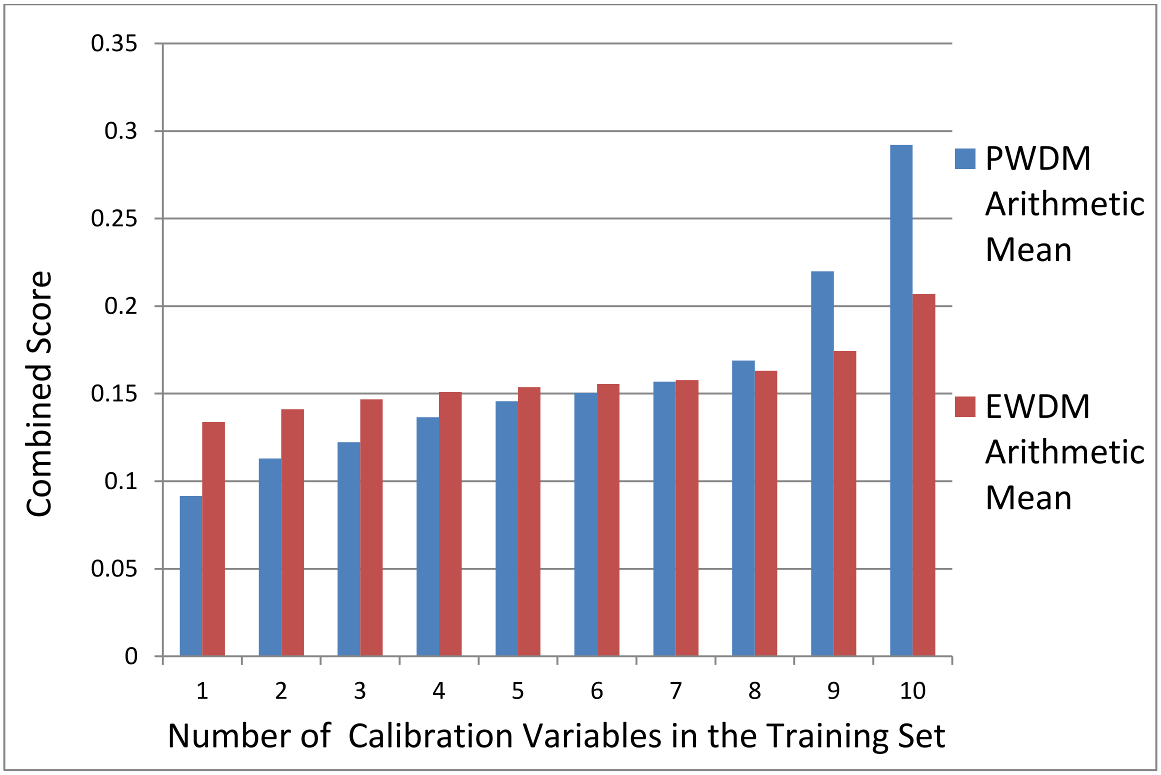
4. Cross Validation

Cross validation is a form of out-of-sample validation. Ideally, we would validate the model out-of-sample by observing values for the variables of interest and comparing how well the observed values were predicted by the *PW* and *EW* combinations. As the variables of interest are usually not observed (which is why expert judgment is needed), out-of-sample validation reduces to cross 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 limits the model's ability to evaluate expert performance and produces *PW* combinations that little resemble those of the whole study. A small test set impairs the ability to discriminate between the performance of *PW* and *EW*.

Based on a large number of out-of-sample validation studies, it is recommended to use 80% of the calibration variables for training sets in cross validation. In this case, that means that training sets of 9 calibration variables are chosen. There are 55 "size 9" subsets of the 11 calibration variables. The *PW* combinations for each subset are computed, and the combined score is computed for the complementary test set. The results are shown in column 9 in Figure 3.

By comparison, the combined scores for the whole study are 0.4 for *PW* and 0.12 for *EW*. The weightings for these 55 training sets involved different combinations of experts 2, 5, and 6.

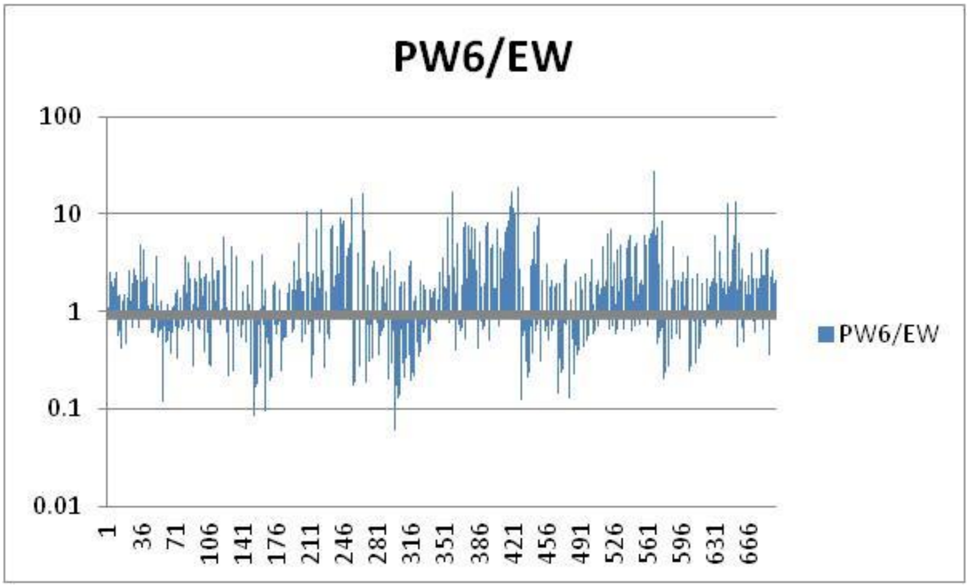
Figure 3. Cross Validation



Note: PWDM denotes performance weight decision maker and EWDM denotes the equal weight decision maker.

Another way to judge out-of-sample validity is to consider all possible training sets that initialize to weight 1 for expert 6. This occurs in 666 of all training sets, or roughly one third of the 2046 ways of splitting the calibration variables into a (non-empty) training set and (non-empty) complementary test set. In each training set, weights are derived for the experts. Experts are combined using these weights and the combination is used to forecast the variables in the test set. Performance on the test set is scored in terms of statistical accuracy and informativeness and compared with the equal weight performance. For the 666 training sets for which expert 6 received weight 1, the ratio of combined scores of performance weighting divided by combined scores for equal weighting is shown in Figure 4. Thus the performance of PW tends to exceed EW in these cross validation cases where the training set performance weights coincide with the study performance weights.

Figure 4. Relative Performance of Performance Weight and Equal Weight Decision Makers



5. Conclusions

We show that structured expert elicitation applied to effects of breastfeeding on IQ finds effects somewhat larger than reported by meta-analyses—five IQ points for the United States from breastfeeding compliant with WHO guidelines relative to no breastfeeding. Our methods suggest that sixty percent (3/5) of these gains come from moderate breastfeeding, i.e., exclusive breastfeeding of 3 months and continued breastfeeding through 9 months of age. We also show total effects in India that are larger than in the United States (8 IQ points). We find that 62.5 percent (5/8) of the total gains in India come from moderate breastfeeding.

We also consider the probability distribution of the IQ increase from increasing breastfeeding for a randomly chosen child from no breastfeeding to breastfeeding that is fully consistent with WHO recommendations—i.e., a shift from cohort one to cohort four. The distributions of mean IQ for cohorts one and four are not likely to be independent, however. If new evidence caused our mean IQ in cohort one to move up toward 100 – indicative of no breastfeeding effect – then our mean for cohort four might plausibly also move downward to 100. Indeed, if lack of breastfeeding has negligible effect on IQ, then plausibly the WHO compliant regime might also have no effect. In this case the two distributions for the mean effect would be countermonotonic and the standard deviation of the difference in mean IQs would be simply twice the standard deviation for each distribution. Since each distribution has a standard deviation of 2, the difference would have a standard deviation of 4. Under these assumptions, we could reject with 95 percent confidence the hypothesis that the true gain in IQ from full compliance with WHO's breastfeeding recommendations, relative to no breastfeeding, is one IQ point or less.

References

Almond, D., and J. Currie. 2011. Killing Me Softly: The Fetal Origins Hypothesis. *Journal of Economic Perspectives* 25: 153–72.

Aspinall, W.P., R.M. Cooke, A.H. Havelaar, S. Hoffmann, and T. Hald. 2016. Evaluation of a Performance-Based Expert Elicitation: WHO Global Attribution of Foodborne Diseases. PLoS ONE 11 (3): e0149817. doi:10.1371/journal.pone.0149817.

Cooke R. 1991.*Experts in Uncertainty; Opinion and Subjective Probability in Science,* Oxford University Press; New York, Oxford, 321 pages. ISBN 0-19-506465-8

Currie, Janet. 2011. Inequality at Birth: Some Causes and Consequences. *American Economic Review* 101 (3): 1–22

Hald, T., W. Aspinall, B. Devleesschauwer, R.M. Cooke, T. Corrigan, A.H. Havelaar, H. Gibb, P. Torgerson, M. Kirk, F. Angulo, R. Lake, N. Speybroeck, and S. Hoffmann. 2016. World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease due to Selected Foodborne Hazards: A Structured Expert Elicitation. PLoS ONE 11 (1): e0145839. doi:10.1371/journal.pone.0145839.

Havelaar, A.H., A. Vargas Galindo, D. Kurowicka, and R.M. Cooke. 2008. Attribution of Foodborne Pathogens Using Structured Expert Elicitation. *Foodborne Pathogens and Disease* 5 (5): 649–59. doi:10.1089/fpd.2008.0115.

Hoddinott, J., J.A. Maluccio, J.R. Behrman, R. Flores, and R. Martorell. 2008. Effect of a Nutrition Intervention during Early Childhood on Economic Productivity in Guatemalan Adults. *Lancet* 371: 411–16.

Horta, B.L., R. Bahl, J.C. Martines, and C.G. Victora. 2007. *Evidence on the Long-Term Effects of Breastfeeding: Systematic Reviews and Meta-analyses*. Geneva: World Health Organization.

Horta, B.L., C. Loret de Mola, and C.G. Victora. 2015. Breastfeeding and Intelligence: A Systematic Review and Meta-analysis. *Acta Paediatrica* 104: 14–19. doi:10.1111/apa.13139.

Horta, B.L., and C.G. Victora. 2013. *Long-Term Effects of Breastfeeding: A Systematic Review*. Geneva: World Health Organization.

Kramer, M.S., B. Chalmers, E.D. Hodnett, et al., and the PROBIT Study Group (Promotion of Breastfeeding Intervention Trial). 2001. Promotion of Breastfeeding Intervention Trial (PROBIT): A Randomized Trial in the Republic of Belarus. *JAMA* 285: 413–20.

Rothlisberger, J.D., D.C. Finnoff, R.M. Cooke, and D.M. Lodge. 2012. Ship-Borne Nonindigenous Species Diminish Great Lakes Ecosystem Services. *Ecosystems* 15: 462–76. doi:10.1007/s10021-012-9522-6.

Tyshenko, M.G., S. ElSaadany, T. Oraby, S. Darshan, W. Aspinall, R. Cooke, A. Catford, and D. Krewski. 2011. Expert Elicitation for the Judgment of Prion Disease Risk Uncertainties. *Journal of Toxicology and Environmental Health, Part A* 74 (2–4): 261–85.

Victora, C.G., R. Bahl, A.J.D. Barros, et al. 2016. Breastfeeding in the 21st Century: Epidemiology, Mechanisms, and Lifelong Effect. *Lancet* 387: 475–90. doi:10.1016/S0140-6736(15)01024-7.

Walfisch, A., C. Sermer, A. Cressman, and G. Koren. 2013. Breast Milk and Cognitive Development—The Role of Confounders: A Systematic Review. *BMJ Open* 3: e003259. doi:10.1136/ bmjopen-2013-003259.

Wittmann, M.E., R.M. Cooke, J.D. Rothlisberger, and D.M. Lodge. 2014a. Using Structured Expert Judgment to Assess Invasive Species Prevention: Asian Carp and the Mississippi–Great Lakes Hydrologic Connection. *Environmental Science & Technology* 48 (4): 2150–56. doi:10.1021/es4043098.

Wittmann, M.E., R.M. Cooke, J.D. Rothlisberger, E.S. Rutherford, H. Zhang, D. Mason, and D.M. Lodge. 2014b. Structured Expert Judgment to Forecast Species Invasions: Bighead and Silver Carp in Lake Erie. *Conservation Biology* 29 (1): 187–97. doi:10.1111/cobi.12369.

Appendix A. Performance Measures and Expert Combination in 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 corresponds, 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%, 50% and 95% values were elicited.

Calibration

For each quantity, each expert divides the range into 4 interquantile intervals for which his or her probabilities are known; namely, p_1 = 0.05: less than or equal to the 5% value, p_2 = 0.45: greater than the 5% value and less than or equal to the 50% value, and so on.

If N quantities are assessed, each expert may be regarded as a statistical hypothesis; namely, each realization falls in one of the four interquantile 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 interquantile 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 three 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 to using the form in (1) (Cooke 1991).

If after a few realizations the expert were to see that all realizations fell outside his or her 90% central confidence intervals, the expert 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 or she learns from the realizations. Expert learning is not a goal of an expert judgment study, and the expert's joint distribution is not elicited. Rather, the decision maker wants experts who do not need to learn from the elicitation. Hence the decision maker scores expert *e* as the statistical likelihood of the hypothesis

H_e: *"the interquantile 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 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 support 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*% 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

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

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 decision maker 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 concentratedness could be contemplated. The above information score is chosen because it is

- familiar;
- tail insensitive;
- scale invariant; and
- slow

The last 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: Decision Maker

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

$$w_{\alpha}(e) = Cal\ (e) \times Inf\ (e) \times 1_{\alpha}(Cal(e) \geq \alpha) \qquad (2)$$

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 cutoff level α, expert *e* is unweighted. The presence of a cutoff level is imposed by the requirement that the combined score be an asymptotically strictly proper scoring rule. That is, an expert maximizes his or her long-run expected score by and only by ensuring that his or her probabilities $p = (0.05, 0.45, 0.45, 0.05)$ correspond to his or 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 "decision maker" (DM). All decision makers discussed here are examples of linear pooling. The classical model is essentially a method for deriving weights in a linear pool. "Good expertise" corresponds with good calibration (high statistical likelihood, high p-value) and high information. We want weights that reward good expertise and that pass these virtues on to the decision maker.

The reward 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 decision maker 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 $1_{\alpha}(calibration\ score)$ 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 decision maker. Let $DM_{\alpha}(i)$ be the result of linear pooling for item *i* with weights proportional to (2):

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

The optimized global weight DM is $DM_{\alpha*}$ where α* maximizes

$$calibration\ score(DM_{\alpha}) \times information\ score(DM_{\alpha}) \qquad (4)$$

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:

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

For each α we define the *item weight DM*$_\alpha$ for item *i* as

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

The optimized item weight DM is $\text{IDM}_{\alpha*}$ where α* maximizes

$$\text{calibration score}(IDM_a) \times \text{information score}(IDM_\alpha) \qquad (7)$$

The nonoptimized versions of the global and item weight DMs are obtained simply by setting α = 0.

Item weights are potentially more attractive, as they allow experts to up- or downweight themselves for individual items according to how much they feel they know about those items. Thus “knowing less” means choosing quantiles farther apart and lowering the information score for a particular item. Of course, good performance of item weights requires that experts can perform this up- or downweighting 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, and 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 decision maker. We should hope that the decision maker 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 on the panel. The global and item weight DMs discussed above (optimized or not) are *performance based DMs.* In general, the optimized global weight DM is used, unless the optimized item weight DM is markedly superior.

Appendix B. Background Material for Experts

We provided all experts with electronic access to the following papers and articles, which may be separated into three groups. The first group includes three systematic reviews. The second group contains eight selected articles cited by the 2013 WHO systematic review. The third group contains articles published in 2011 or thereafter and not cited by the 2013 WHO review. Specifically, it presents 27 peer-reviewed articles with original research assessing the relationship between breastfeeding and various measures of cognitive performance. These articles did not all necessarily employ internal comparison groups, measure cognition using standard tests, focus on children older than one year, or adjust for stimulation or interaction with the child.

Systematic Reviews/Meta-analyses

1. Horta, B.L., and C.G. Victora. 2013. *Long-Term Effects of Breastfeeding: A Systematic Review*. Geneva: World Health Organization.
2. Horta, B.L., C. Loret de Mola, and C.G. Victora. 2015. Breastfeeding and Intelligence: A Systematic Review and Meta-analysis. *Acta Paediatrica* 104: 14–19. doi:10.1111/apa.13139.
3. Walfisch, A., C. Sermer, A. Cressman, and G. Koren. 2013. Breast Milk and Cognitive Development—The Role of Confounders: A Systematic Review. *BMJ Open* 3: e003259. doi:10.1136/ bmjopen-2013-003259.

Selected Articles Cited by the 2013 WHO Systematic Review

1. Clark, K.M., M. Castillo, A. Calatroni, T. Walter, M. Cayazzo, P. Pino, and B. Lozoff. 2006. Breast-feeding and Mental and Motor Development at 5 1/2 Years. *Ambulatory Pediatrics* 6 (2): 65–71.
2. Der, G., G.D. Batty, and I.J. Deary. 2006. Effect of Breast Feeding on Intelligence in Children: Prospective Study, Sibling Pairs Analysis, and Meta-analysis. *BMJ* 333 (7575): 945.
3. Evenhouse, E., and S. Reilly. 2005. Improved Estimates of the Benefits of Breastfeeding Using Sibling Comparisons to Reduce Selection Bias. *Health Services Research* 40 (6, pt. 1): 1781–1802.
4. Gibson-Davis, C.M., and J. Brooks-Gunn. 2006. Breastfeeding and Verbal Ability of 3-Year-Olds in a Multicity Sample. *Pediatrics* 118 (5): e1444–51.
5. Jacobson, S.W., L.M. Chiodo, and J.L. Jacobson. 1999. Breastfeeding Effects on Intelligence Quotient in 4- and 11-Year-Old Children. *Pediatrics* 103 (5): e71.
6. Lucas, A., R. Morley, T.J. Cole, G. Lister, and C. Leeson-Payne. 1992. Breast Milk and Subsequent Intelligence Quotient in Children Born Preterm. *Lancet* 339 (8788): 261–64.

7. Morrow-Tlucak, M., R.H. Haude, and C.B. Ernhart. 1988. Breastfeeding and Cognitive Development in the First 2 Years of Life. *Social Science & Medicine* 26 (6): 635–39.
8. Wigg, N.R., S. Tong, A.J. McMichael, P.A. Baghurst, G. Vimpani, and R. Roberts. 1998. Does Breastfeeding at Six Months Predict Cognitive Development? *Australian and New Zealand Journal of Public Health* 22 (2): 232–36.

Recent Articles Not Referenced in the 2013 WHO Review or Published More Recently

1. Ali, S.S., S.M. Dhaded, and S. Goudar. 2014. The Impact of Nutrition on Child Development at 3 Years in a Rural Community of India. *International Journal of Preventive Medicine* 5 (4): 494–99
2. Belfort, M.B., S.L. Rifas-Shiman, K.P. Kleinman, L.B. Guthrie, D.C. Bellinger, E.M. Taveras, M.W. Gillman, and E. Oken. 2013. Infant Feeding and Childhood Cognition at Ages 3 and 7 Years: Effects of Breastfeeding Duration and Exclusivity. *JAMA Pediatrics* 167 (9): 836–44.
3. Bernard, J.Y., M. De Agostini, A. Forhan, T. Alfaiate, M. Bonet, V. Champion, M. Kaminski, B. de Lauzon-Guillain, M. Charles, and B. Heude. 2013. Breastfeeding Duration and Cognitive Development at 2 and 3 Years of Age in the EDEN Mother–Child Cohort. *Journal of Pediatrics* 163 (1): 36–42.
4. Boutwell, B.B., K.M. Beaver, and J.C. Barnes. 2012. Role of Breastfeeding in Childhood Cognitive Development: A Propensity Score Matching Analysis. *Journal of Paediatrics and Child Health* 48 (9): 840–45.
5. Brion, M.J., D.A. Lawlor, A. Matijasevich, B. Horta, L. Anselmi, C.L. Araujo, et al. 2011. What Are the Causal Effects of Breastfeeding on IQ, Obesity and Blood Pressure? Evidence from Comparing High-Income with Middle-Income Cohorts. *International Journal of Epidemiology* 40: 670–80.
6. Cai, S., W.W. Pang, Y. Ling Low, L.W. Sim, S.C. Sam, M.B. Bruntraeger, E.Q. Wong, D. Fok, B.F.P. Broekman, L. Singh, J. Richmond, P. Agarwal, A. Qiu, M.S. Seang, F. Yap, K.M. Godfrey, P.D. Gluckman, Y. Chong, M.J. Meaney, M.S. Kramer, and A. Rifkin-Graboi. 2014. Infant Feeding Effects on Early Neurocognitive Development in Asian Children. *American Journal of Clinical Nutrition* 101 (2): 326–36.
7. Chiu, W.C., H.F. Liao, P.J. Chang, P.C. Chen, and Y.C. Chen. 2011. Duration of Breast Feeding and Risk of Developmental Delay in Taiwanese Children: A Nationwide Birth Cohort Study. *Paediatric and Perinatal Epidemiology* 25 (6): 519–27.
8. Colen, G.C., and D.M. Ramey. 2014. Is Breast Truly Best? Estimating the Effects of Breastfeeding on Long-Term Child Health and Wellbeing in the United States Using Sibling Comparisons. *Social Science & Medicine* 109: 55–65.
9. Deoni, S.C., D.C. Dean, I. Piryatinsky, J. O'Muircheartaigh, N. Waskiewicz, K. Lehman, M. Han, and H. Dirks. 2013. Breastfeeding and Early White Matter Development: A Cross-sectional Study. *Neuroimage* 82: 77–86.

10. Eriksen, H.F., U.S. Kesmodel, U. Underbjerg, T.R. Kilburn, J. Bertrand, and E.L. Mortensen. 2013. Predictors of Intelligence at the Age of 5: Family, Pregnancy and Birth Characteristics, Postnatal Influences, and Postnatal Growth. *PLoS ONE* 8 (11): e79200.

11. Huang, J., K.E. Peters, M.G. Vaughn, and C. Witko. 2014. Breastfeeding and Trajectories of Children's Cognitive Development. *Developmental Science* 17: 452–61.

12. Jacobson, S.W., R.C. Carter, and J.L. Jacobson. 2014. Breastfeeding as a Proxy for Benefits of Parenting Skills for Later Reading Readiness and Cognitive Competence. *Journal of Pediatrics* 164: 440–42.

13. Jedrychowski, W., F. Perera, J. Jankowski, M. Butscher, E. Mroz, E. Flak, I. Kaim, I. Lisowsk-Miszczyk, A. Skarupa, and A. Sowa. 2012. Effect of Exclusive Breastfeeding on the Development of Children's Cognitive Function in the Krakow Prospective Birth Cohort Study. *European Journal of Pediatrics* 171 (1): 151–58.

14. Jiang, M., E.M. Foster, and C.M. Gibson-Davis. 2011. Breastfeeding and the Child Cognitive Outcomes: A Propensity Score Matching Approach. *Maternal and Child Health Journal* 15 (8): 1296–1307.

15. Julvez, J., M. Guxens, A. Carsin, J. Forns, M. Mendez, M.C. Turner, and J. Sunyer. 2014. A Cohort Study on Full Breastfeeding and Child Neuropsychological Development: The Role of Maternal Social, Psychological, and Nutritional Factors. *Developmental Medicine & Child Neurology* 56 (2): 148–56.

16. Leventakou, V., T. Roumeliotaki, K. Koutra, M. Vassilaki, E. Mantzouranis, P. Bitsios, M. Kogevinas, and L. Chatzi. 2013. Breastfeeding Duration and Cognitive, Language and Motor Development at 18 Months of Age: Rhea Mother–Child Cohort in Crete, Greece. *Journal of Epidemiology & Community Health* 69: 23239.

17. Martin, N.W., B. Benyamin, N.K. Hansell, G.W. Montgomery, N.G. Martin, M.J. Wright, and T.C. Bates. 2011. Cognitive Function in Adolescence: Testing for Interactions between Breast-feeding and FADS2 Polymorphisms. *Journal of the American Academy of Child & Adolescent Psychiatry* 50 (1): 55–62.

18. Oddy, W.H., M. Robinson, G.E. Kendall, J. Li, S.R. Zubrick, and F.J. Stanley. 2011 Breastfeeding and Early Child Development: A Prospective Cohort Study. *Acta Paediatrica* 100 (7): 992–99.

19. Qawasmi, A., A. Landeros-Weisenberger, J.F. Leckman, and M.H. Bloch. 2012. Meta-analysis of Long-Chain Polyunsaturated Fatty Acid Supplementation of Formula and Infant Cognition. *Pediatrics* 6 (1): 1141–49.

20. Quigley, M.A., C. Hockley, C. Carson, Y. Kelly, M.J. Renfrew, and A. Sacker. 2012. Breastfeeding Is Associated with Improved Child Cognitive Development: A Population-Based Cohort Study. *Journal of Pediatrics* 160 (1): 25–32.

21. Rippeyoung, P.L. 2013. Can Breastfeeding Solve Inequality? The Relative Mediating Impact of Breastfeeding and Home Environment on Poverty Gaps in Canadian Child Cognitive Skills. *Canada Journal of Sociology* 38 (1): 65–85.

22. Rothstein, D.S. 2013. Breastfeeding and Children's Early Cognitive Outcomes. *Review of Economics and Statistics* 95 (3): 919–31.

23. Ruiz, P., M. León, P. Herreros, and I. Ibabe. 2013. Breastfeeding and Its Influence into the Cognitive Process of Spanish School-children (6 Years Old), Measured by the Wechsler Intelligence Scale. *Archivos Latinoamericanos de Nutricion* 63 (3): 218–23.

24. Smithers, L.G., R.K. Golley, M.N. Mittinty, L. Brazionis, K. Northstone, P. Emmett, et al. 2012. Dietary Patterns at 6, 15 and 24 Months of Age Are Associated with IQ at 8 Years of Age. *European Journal of Epidemiology* 27: 525–35.

25. Tozzi, A.E., P. Bisiacchi, V. Tarantino, F. Chiarotti, L. D'elia, B. De Mei, M. Romano, F. Gesualdo, and S. Salmaso. 2012. Effect of Duration of Breastfeeding on Neuropsychological Development at 10 to 12 Years of Age in a Cohort of Healthy Children. *Developmental Medicine & Child Neurology* 54 (9): 843–48.

26. Victora, C.G., B.L. Horta, C.L. de Mola, L. Quevedo, R.T. Pinheiro, D.P. Gigante, H. Goncalves, and F.C. Barros. 2015. Association between Breastfeeding and Intelligence, Educational Attainment, and Income at 30 Years of Age: A Prospective Birth Cohort Study from Brazil. *Lancet Global Health* 3 (4): e199–e205.

27. Von Stumm, S., and R. Plomin. 2015. Breastfeeding and IQ Growth from Toddlerhood through Adolescence. *PLoS ONE* 10 (9): e0138676. doi:10.1371/journal.pone.0138676.

Appendix C. Elicitation Protocol

Introduction

Structured expert judgment is an accepted tool in risk analysis for supplementing data shortfalls, quantifying uncertainty, and building rational consensus. It has been used in studies sponsored by the European Union, the US National Oceanographic and Atmospheric Administration, the US Environmental Protection Agency, Health Canada, and the Robert Wood Johnson Foundation, among many others, to characterize uncertainty in a wide variety of relationships not amenable to repeated experimentation. To pick a few examples, these include the effects of medical procedures, risks from nuclear power plants, and risks of invasive species.

Selected 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 "decision maker." 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:

Regarding the US National Longitudinal Survey of Youth 1979 Matched Mother and Child Data (NLSY79-C), (11,512 records), the mean year of birth is 1986 (Range = 1970–2011) What is the mean year of birth of children who have four older brothers or sisters?

What is the mean year of birth of children who have four older brothers or sisters?

___	___	___	___	___
5%	***25%***	***50%***	***75%***	***95%***

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

The 50%-tile is that number for which you judge the chance ½ that the true value is above or below.

The 25%-tile 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%-tile 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.

ALWAYS: 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 have four older brothers or sisters?

1980	***1983***	***1986***	***1988***	***1990***
5%	***25%***	***50%***	***75%***	***95%***

This means that the true value is equally likely to be above or below 1986; there is a 50% chance that it lies between 1983 and 1988, and a 90% 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 that 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%-tile. If the expert's assessments are *statistically accurate*, then in the long run, 5% of the answers should fall within this interpercentile interval. Similarly, 90% of the answers should fall between the 5%-tile and the 95%-tile, etc.

In gauging overall performance, statistical accuracy is more important than informativeness. Noninformative 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 fact by itself 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 exert more influence on the decision maker. But 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 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 11,512 records in NLSY79-C is the week when breastfeeding ended NOT reported?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

B) Of the 11,512 records in NLSY79-C, how many are firstborn?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

C) Of the 11,512 records in NLSY79-C, how many are 4th born?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

Elicitation

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" (Dunn and Dunn 1981). 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.[1]

Calibration Questions

1. In NLSY79-C the average Peabody Picture Vocabulary Test (Revised Form L) (PPVT) mean score, among the children with scores, is 90.660. What is the average among firstborn children with at least one PPVT score?

5% 25% 50% 75% 95%

2. In NLSY79-C the average PPVT mean score, among the children with scores, is 90.660. What is the average among firstborn children who were ever breastfed?

5% 25% 50% 75% 95%

*3. In NLSY79-C, 1,706 children have PPVT scores recorded for 1986 and Peabody Individual Assessment Test **math** scores for 1986. What is the correlation between these scores?*

5% 25% 50% 75% 95%

*4. In NLSY79-C, 1,700 children have PPVT scores recorded for 1986 and Peabody Individual Assessment Test **reading** recognition scores for 1986. What is the correlation between these scores?*

5% 25% 50% 75% 95%

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

5. In what percentage of the 11,512 records in NLSY79-C is the Peabody Picture Vocabulary Test (PPVT) never reported?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

6. In NLSY79-C, the average age in weeks when breastfeeding ended is 9.12. What is the average age in weeks when breastfeeding ended among the 1,583 only children who were breastfed?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

7. In NLSY79-C the average years of schooling for the mothers was 12.86 years. What is the average years of schooling for the mothers of the 2900 only children (i.e., those without siblings), both breastfed and non-breastfed?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

8. In the 2005–06 Demographic Health Survey for India, what is the 50th percentile for duration of breastfeeding (in months) among children who were breastfed and who were not still breastfeeding at the time of the survey? This data excludes children who died while breastfeeding.[2]

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

9. In the 2005–06 Demographic Health Survey for India, what is the 75th percentile for duration of breastfeeding (in months) among children who were breastfed and who were not still breastfeeding at the time of the survey? This data excludes children who died while breastfeeding. [2]

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

[2] These data include individuals reported as breastfed but with a duration of 0 months (approximately 3%).

10. The US Panel Study of Income Dynamics Child Development Supplement (PSID-C) data set has 3,563 records. In what percentage of completed records is the sum of Woodcock-Johnson Word Scores and Woodcock-Johnson Applied Problem Scores in 1997 greater than in 2002?

________	________	________	________	________
5%	***25%***	***50%***	***75%***	***95%***

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

________	________	________	________	________
5%	***25%***	***50%***	***75%***	***95%***

Variables of Interest

Questions 12 through 23 concern a hypothetical ideal perfectly randomized experiment with a very large number of subjects from each of three countries listed below. We select India and China because their populations are important from a global health perspective and yet estimates of effects of breastfeeding on cognitive performance from long-term longitudinal studies appear to be sparse for these countries. We include the United States because the published literature includes multiple studies of associations between breastfeeding and cognitive performance, using different data.

All infants (also siblings) are randomly assigned to one of four feeding cohorts.

Feeding/food source	Feeding patterns by age			
	Cohorts			
	1	2	3	4
Breastfeeding, exclusive	None	3 months	6 months	6 months
Breastfeeding, any	None	3 to 9 months	None	6 to 24 months
Infant formula, exclusive	6 months	None	None	None
Infant formula, any	6 to15 months	3 to 15 months	6 to 15 months	None
Complementary foods	From 6 months	From 6 months	From 6 months	From 6 months

All formula is approved by the US Food and Drug Administration and provided by the mother while holding the infant in a position where breastfeeding could have occurred. All children are tested at age 10 with the Wechsler Intelligence Scale for Children, Revised (WISC) or its foreign equivalent, properly normed. The overall average WISC (IQ) score (within each country and cohorts) is 100, st. dev. = 15.

You may consider the following data while developing your responses. The reported values are for the most recent data that are publicly available.

Variable	Source	China	India	USA
Youth literacy rate, population 15–24, both sexes, %	UNESCO Institute for Statistics http://data.uis.unesco.org/	99.73	90.178	NA
Dropout rates through middle school, both sexes, %	UNESCO Institute for Statistics http://data.uis.unesco.org/	7.96	2.987	NA
Educational attainment (percentage of 15–19-year-olds for whom primary education is the highest level attained, data from 2010)	Barro and Lee, v. 2.0, 06/14 http://www.barrolee.com/	6.5	18.2	7.4
Infant mortality (per thousand live births)	UNESCO Institute for Statistics http://data.uis.unesco.org/index.aspx?queryid=190&lang=en	12.1	43.8	6
Stunting, percentage of all children under the age of five who are more than 2 standard deviations below international norms of height for age	UN http://data.un.org/Data.aspx?d=SOWC&f=inID:106	10 %	48 %	3%
Pupils per teacher in primary school	http://data.uis.unesco.org/	16.9	35.2	14.4
Government expenditures per primary student (purchasing power parity)	UNESCO Institute for Statistics http://www.uis.unesco.org/Education/Pages/education-finance.aspx	NA	$433	$10,237

Please assess your uncertainty regarding the average scores within the different cohorts.

USA

For ***the USA*** *what is the average WISC score at age 10 for children*

12. …who were never breastfed, and received infant formula from birth to 15 months and complementary foods from 6 months (cohort 1)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

13. …who were exclusively breastfed from birth until 3 months and breastfed until 9 months, received formula from 3 to 15 months, and received complementary foods from 6 months (cohort 2)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

14. …who were exclusively breastfed from birth until 6 months when breastfeeding stopped, received formula from 6 months to 15 months, and received complementary foods from 6 months (cohort 3)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

15. …who were exclusively breastfed from birth until 6 months, when complementary foods were introduced, and partly breastfed until 24 months (cohort 4)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

India

*For **India** what is the average WISC score at age 10 for children*

16. …who were never breastfed, and received infant formula from birth to 15 months and complementary foods from 6 months (cohort 1)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

17. …who were exclusively breastfed from birth until 3 months and breastfed until 9 months, received formula from 3 to 15 months, and received complementary foods from 6 months (cohort 2)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

18. …who were exclusively breastfed from birth until 6 months when breastfeeding stopped, received formula from 6 months to 15 months, and received complementary foods from 6 months (cohort 3)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

19. …who were exclusively breastfed from birth until 6 months, when complementary foods were introduced, and partly breastfed until 24 months (cohort 4)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

China

*For **China** what is the average WISC score at age 10 for children*

20. …who were never breastfed, and received infant formula from birth to 15 months and complementary foods from 6 months (cohort 1)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

21. …who were exclusively breastfed from birth until 3 months and breastfed until 9 months, received formula from 3 to 15 months, and received complementary foods from 6 months (cohort 2)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

22. …who were exclusively breastfed from birth until 6 months when breastfeeding stopped, received formula from 6 months to 15 months, and received complementary foods from 6 months (cohort 3)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

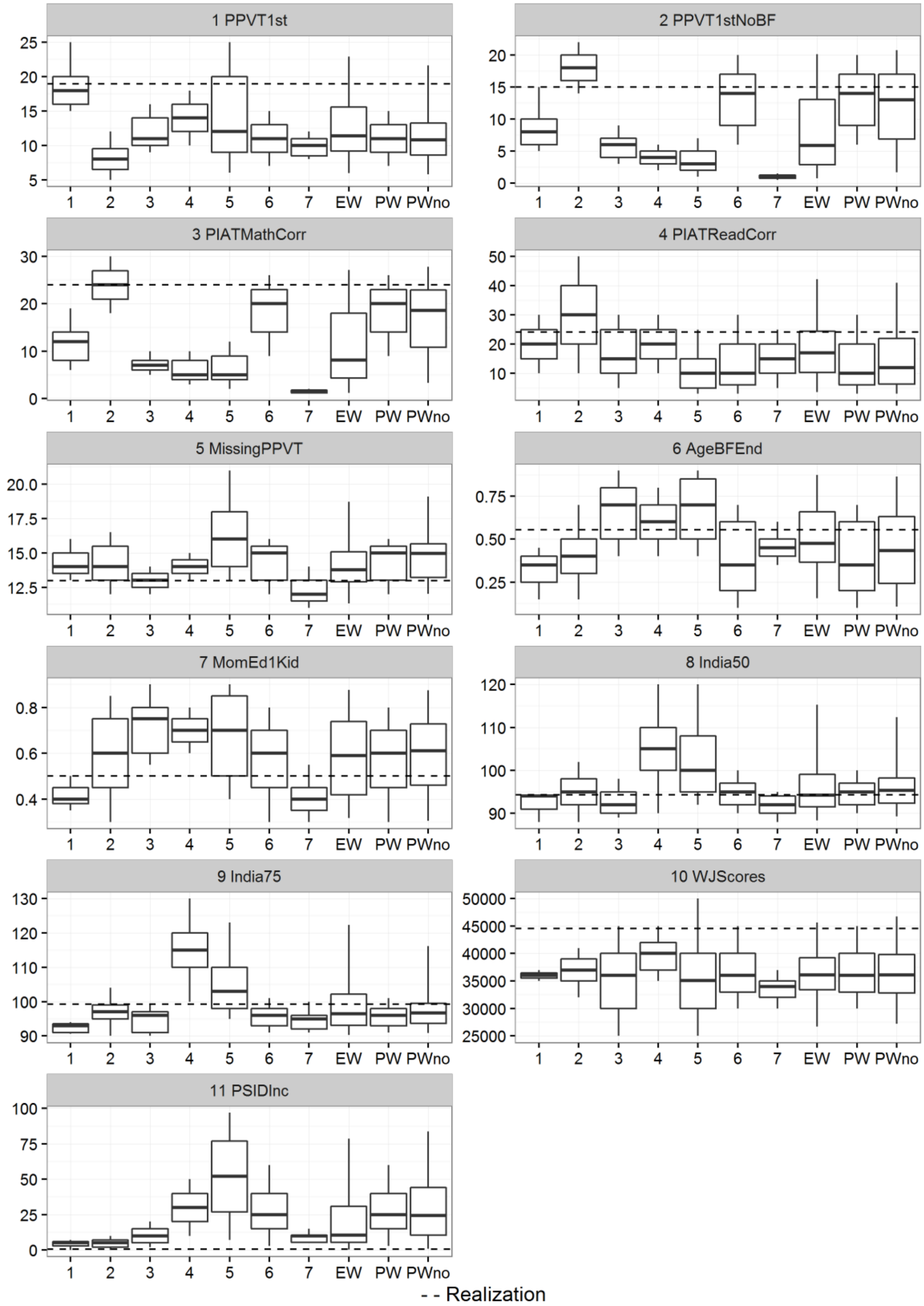
23. …who were exclusively breastfed from birth until 6 months, when complementary foods were introduced, and partly breastfed until 24 months (cohort 4)?

__________ __________ __________ __________ __________

5% 25% 50% 75% 95%

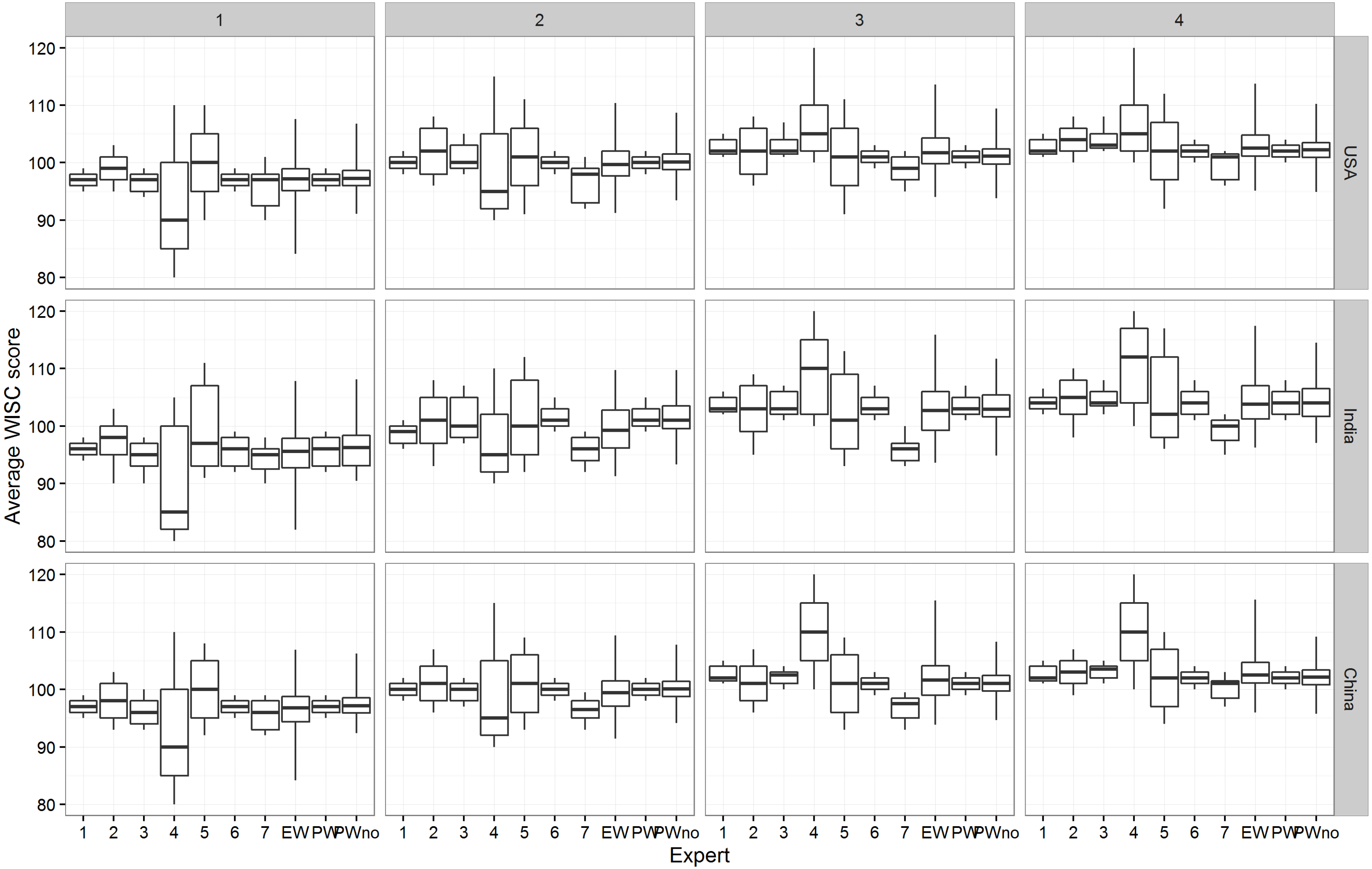
Appendix D. Range Graphs

Figure D1. Range Graphs from All Experts and Decision Makers for the 11 Calibration Variables



Note: Boxplots show the 5th, 25th, 50th, 75th, and 95th percentiles.

Figure D2. Range Graphs from All Experts and Decision Makers for the Variables of Interest



Note: Boxplots show the 5th, 25th, 50th, 75th, and 95th percentiles

Appendix E. Notes on Experts' Rationales

Below are the rationales provided by experts during the elicitation interviews. Note that the order here corresponds to the expert number used in figures and tables but does not reflect the alphabetical ordering of experts listed in the main text.

Expert 1

There's no additional cognitive benefit to be had from breastfeeding beyond 6 months of exclusive breastfeeding, so cohorts 3 and 4 will have similar results. There's a "window of opportunity" for breastfeeding to impact cognitive development. It's possible that excessive breastfeeding could even lower cognitive development, due to too much exposure to the organic compounds in the fatty acids of breastmilk or possibly the transfer of contaminants.

The difference between cohorts would be larger in India than in the US. In lower socioeconomic classes, there's a bigger difference between kids who were breastfed versus not breastfed. Kids also need more breastmilk in lower socioeconomic classes. Breastfeeding can have more of an impact for these kids. Kids in higher socioeconomic classes can benefit more from their environment, but if environment is lacking, breastfeeding can have a big impact. In India, there will be a big difference in stunting rates between cohorts 1 and 4, as kids in cohort 4 will be getting proper nutrition from breastmilk. Breastfeeding can help an infant recover from poor prenatal nutrition, which will be a factor in India. Infants with low birth weight who are breastfed will recover faster to normal weight ranges, and a similar dynamic could happen with brain development, too.

The impact of breastfeeding on cognition will be lower in high income, highly developed settings.

The impact of breastfeeding in China will be similar to the impact in the US.

There's a dose-response from breastfeeding from 0–6 months, but the effect plateaus at 6 months.

The contact hypothesis could have an impact on child cognition, but it's tough to measure. The idea is that skin-to-skin contact between the mother and infant helps the baby relax, and that has an impact. If this happens, though, this benefit would also decrease after 6 months, because by that time the baby's response is set.

Exclusive breastfeeding for 6 months is the key to seeing an impact. This effect isn't the result of maternal confounding.

Expert 2

There's less uncertainty about the impact of breastfeeding on IQ in the US than in India or China. However, the difference between cohorts 2 and 3 is tough to think about. Exclusive breastfeeding and duration of any breastfeeding both have an impact, but it's not clear which is more important. More data exists looking at the impact of duration, and duration may be more important that exclusivity. Cohort 2 received any breastfeeding only 3 months more than cohort 3, though, and that may not be big enough to see a difference. Cohort 2 would likely have higher IQ if they received any breastfeeding until 12 months, though.

The effect of breastfeeding on IQ is larger in India than in the US. Exclusive breastfeeding is also more important in India, because kids who are exclusively breastfed will get less diarrhea than other kids. In India, kids in cohort 2 will have more diarrhea than kids in cohort 3, and that will have an impact on cognition. If a child has a lot of infections, they spend energy fighting infection instead of promoting brain health. This means that longer breastfeeding—and longer exclusive breastfeeding—has a bigger impact on IQ in India. Poorer mothers in India are also likely to overdilute the formula, which could have an impact.

In China, hygiene and water quality are better, so formula will be safe and properly administered. There isn't the same infection risk that exists in India for kids who aren't breastfed. This means there isn't likely to be a difference between cohorts 2 and 3 in China.

Expert 3

The main questions for thinking about the outcomes of the different cohorts are

- What is the impact of the different doses?
- What period of time is most critical?

The critical period for breastfeeding is likely early, so there will be a big difference between no breastfeeding and 3 months exclusive breastfeeding. There will also likely be more benefit from early breastfeeding than from continued breastfeeding.

Breastfeeding will have a bigger impact on IQ in India than the United States. The main reason is that the population is at a higher nutritional risk, so breastfeeding provides a larger benefit. A secondary reason is that there is more environmental and developmental deprivation in the population, so there's more room for the effect of breastfeeding to come through.

China has less nutritional deprivation than India, so the effect of breastfeeding is likely between that seen in the USA and India.

Expert 4

A fully randomized trial won't have selection issues, which is the problem with existing studies. There is a chance that breastfeeding has no effect. Everything observed in current studies could be due to selection issues. On the other hand, the impact of breastfeeding could be large, so all of these estimates are uncertain.

In the USA, the benefit of continued breastfeeding after 6 months exclusive breastfeeding would be minuscule if it existed at all, so there won't be a difference between cohorts 3 and 4.

In India, there could be a greater impact of breastfeeding on IQ, driven by the health benefits of breastfeeding. Breastfeeding in the first 3 months in India will have a large health impact, so there will be an IQ difference between cohorts 1 and 2. There could also be a health impact of continued breastfeeding after 6 months in India, so there could be a slight difference in IQ between cohorts 3 and 4 here.

Expert 5

There's virtually no difference in IQ between breastfed and non-breastfed kids. The difference could be 1–2 points. Breastfeeding has very little effect on cognition in normal weight kids. There may be an effect in pre-term or low birthweight kids.

There isn't much literature on the impact of breastfeeding duration, although the literature is growing. Some funny things happen in the data when looking at long durations. Confounding for duration is different than the confounding for initiation. Maternal employment practices are the big confounder for duration, over and above the confounders for breastfeeding initiation. Some studies show IQ drops with breastfeeding longer than 12 months; other literature shows IQ keeps increasing as duration of breastfeeding increases. There isn't a consistent dose-response, though.

Breastfeeding will have a bigger impact in India than in the US or China because a greater proportion of infants there are low birthweight.

Factors like birth order and number of kids (which are related ideas) were not considered for these estimates. The impact of these factors, if it exists, would be small.

Expert 6

In the United States, the difference between cohorts 1 and 4 will be very small.

In India, kids get less schooling, which could result in a bigger impact of breastfeeding. Breastfeeding protects against infection, so that could also lead to a bigger difference between cohorts in India than in the US. Current average breastfeeding practices in India are probably most similar to cohort 2. There's more uncertainty about outcomes in India than in the US.

There is no clear indication that China should be any different from the US. China has more infection than the US, which could result in a bigger impact of breastfeeding. However, the parents and schools in China work the kids hard, which could result in a smaller impact of breastfeeding. The increased standardization in China may mean the Chinese population is less variable. Estimating any actual difference from the US would be artificial, so the estimates from China will match those from the US.