William R Maas
Private Citizen

Subject: Flawed Measurement of IQ Outcome in the MIREC Studies

I am writing to express concerns about studies relying on the Maternal-Infant Research on Environmental Chemicals (MIREC) data set to assess the association of fluoride with IQ of preschoolers. The NTP report cites two studies using these data, Green et al. (2019) and Till et al. (2020). At least one other study has been published, Farmus et al. (2021), too late for consideration in the report under review, but it and future reports that may be included in future NTP summaries share the same problem.

The National Academies Committee in its Letter Report (page 8) had raised a concern about studies that were classified as having lower risk of bias when measurement of a neuro-developmental or cognitive outcome was flawed. The MIREC study trained a single staff person from each of the study sites to administer in-person assessments and thereby created the potential for differential assessment among cities. The authors did not provide interrater reliability data to judge if the variation in IQ scores is due to the rater or not.

These MIREC studies were rated as "probably low risk of bias based on indirect evidence that the outcome was assessed using instruments that were valid and reliable in the study population, and that the outcome assessors were blind to participants' fluoride exposure." While the outcome assessors were blind to fluoride exposure, and while the instruments themselves were valid and reliable, the way the instruments were used did not result in valid and reliable measurement of outcomes in the study population.

The NIEHS report on Evaluating Features and Application of Neurodevelopmental Tests in Epidemiological Studies provides examples of factors that could present problems and introduce bias in the test results.[1] Styck and Walsh (2016) report "examiner errors occur frequently and impact index and FSIQ scores," concluding that "current estimates for the standard error of measurement of popular IQ tests may not adequately capture the variance due to the examiner".[2] Therefore, the bias in the standard error derived from a convenience sample is compounded in these MIREC studies.

The MIREC data subset used by Green, Till, and Farmus provided data from a convenience sample recruited from 7 healthcare facilities in 6 cities. Three cities intentionally provide higher fluoride exposure to mothers and children through the practice of community water fluoridation. Because the MIREC data collection plan was formulated before researchers decided to use these data to assess whether IQ differences might be associated with different levels of fluoride exposure, no steps were taken to determine how much of any IQ difference measured between cities was the result of examiner variation rather than actual difference in IQ.

---

[1] National Institute of Environmental Health Sciences, Public Health Service. NIEHS Report on Evaluating Features and Application of Neurodevelopmental Tests in Epidemiological Studies NIEHS Report 01, June 2022, U.S. Department of Health and Human Services ISSN: 2768-5632.

[2] Styck KM, Walsh SM. Evaluating the prevalence and impact of examiner errors on the Wechsler scales of intelligence: A meta-analysis. Psychol Assess. 2016 Jan;28(1):3-17. doi: 10.1037/pas0000157. Epub 2015 May 25. PMID: 26011479.

If the originators of the MIREC study wanted to study fluoride, knowing that there would be systematic difference in fluoride exposure between cities, they should have designed the IQ assessment plan differently. They did not, which I believe makes MIREC an inappropriate database to study the association of fluoride with IQ.

I could find no details about the IQ assessment in any of the 3 papers that used the MIREC data set to study fluoride-IQ questions. Nor did the authors identify the use of different assessors in each city as a limitation. The only paper using this dataset we were able to find that discussed the way IQ of children was assessed was Etzel et al. (2018). "A single staff person from each study site administered in-person assessments." Etzel also noted that study staff from each participating study site completed a 3-day training session that was led by a PhD-level psychologist and focused on specialized training of these assessment tools.

I am not questioning the validity or reliability of the WPSSI-III scale of intelligence to assess a 3 year old Canadian child's cognitive abilities nor the intentions of those who were responsible for training staff to assess IQ. Nor I am criticizing the developers of MIREC who designed the study before they knew that anyone would try to assess the association of fluoride with IQ. However, intentions do not ensure that the score of one rater will be identical to that of another. In fact, we know that is never the case and why inter-rater reliability should be assessed when multiple examiners are used in a study. Moreover, no responsible researcher would ever use different examiners if the exposure of interest was known to be different in the cohorts to be assessed by each examiner without a plan to account for this study design in analysis.

In the absence of any acknowledgement of concern about the IQ outcome measure by the researchers who used MIREC data to explore the effects of fluoride nor any way to control for examiner variation, those studies cannot possibly be considered to be "probably low risk of bias". Their contributions, if any, to any conclusions reached by the NTP State of Science report should reflect that.

William R Maas, DDS, MPH, MS
April 27, 2023