NCME

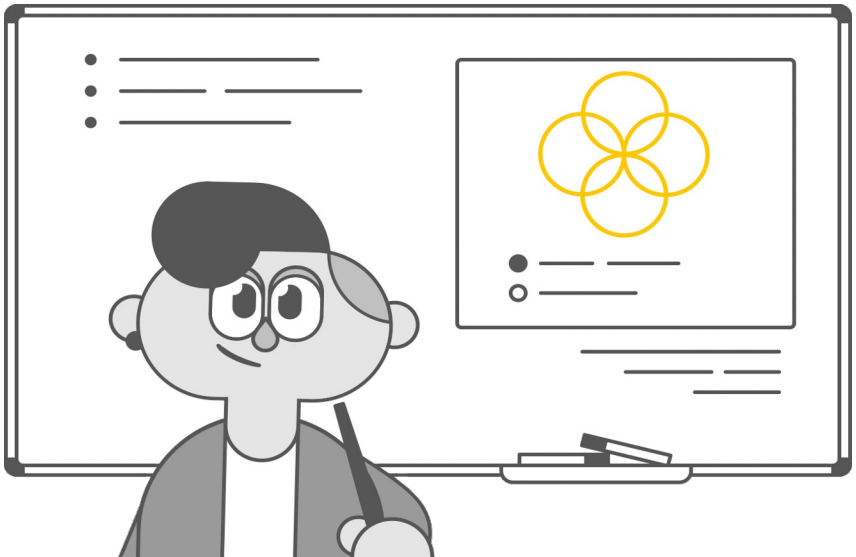
Sampling Considerations for Building Concordance Tables

Ramsey Cardwell
Steven Nydick
J.R. Lockwood
Alina von Davier

April 15, 2023

Agenda

- Background/Context
- Challenges
- (Possible) Solutions
- Results
- Wrap-up



Background

DET Concordance Update

1. The DET undertook an update to concordance with other test in 2022
    a. Many more institutions have adopted the DET since the previous concordance study in 2019
    b. The test-taker population has changed considerably in that time
2. The concordance study needed to address several methodological challenges
    a. Collecting a sufficiently large sample
    b. Combining data sources
    c. Ensuring generalizability of the results

Data Collection Process

1. All DET test-takers are asked to provide alternate assessment scores (self-report).
2. We solicited official score reports on alternate assessments from all relevant test takers (official).
3. We planned to combine official (1,643 for IELTS and 328 for TOEFL) with self-report (3,806 for IELTS and 1,276 for TOEFL) scores to determine equivalent TOEFL/IELTS score for specific DET scores.

Data Collection Process

1. **Self-report**: DET test-takers are asked about alternate assessment scores.
2. **Official**: We solicited official score reports from DET test takers.
3. We planned to combine official with self-report scores to determine equivalent TOEFL/IELTS score for specific DET scores.

	IELTS	TOEFL
Self-report	3,806	1,276
Official	1,643*	328*

*We continued the official data collection, repeated the concordance later with a larger sample, and got nearly equivalent results.

Challenges

Problem 1: Self-Report Bias

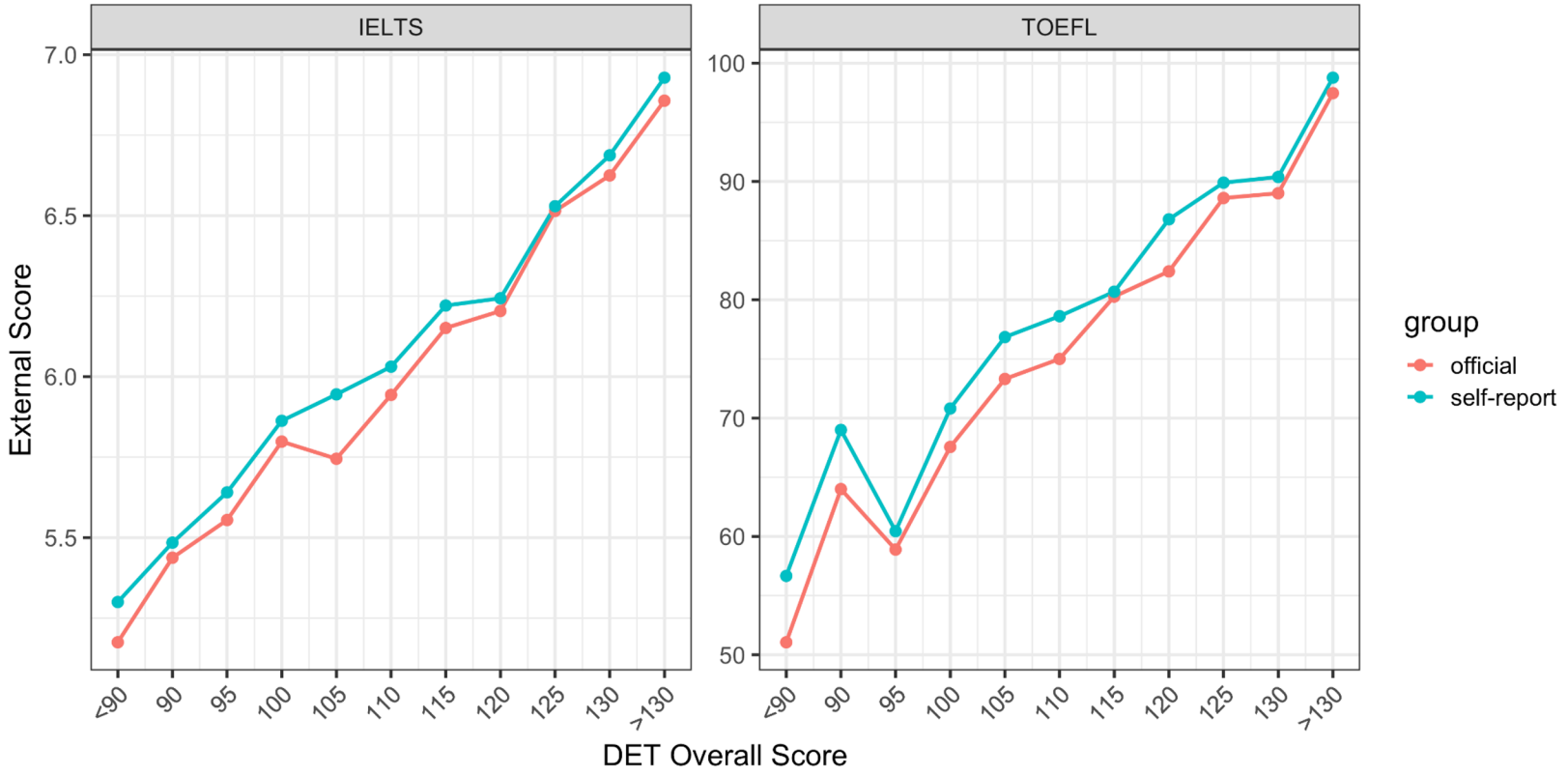
Self-report external scores for DET test-takers had bias:

- Collect official and self-report data from the same people.
- Calculate the average bias at each DET score point:

$$B(x) = S(x) - O(x)$$

  - B(x) estimates the mean reporting bias given DET score x.
  - B(x) is positive for TOEFL and IELTS scores.

Reporting Bias (Paired Data)

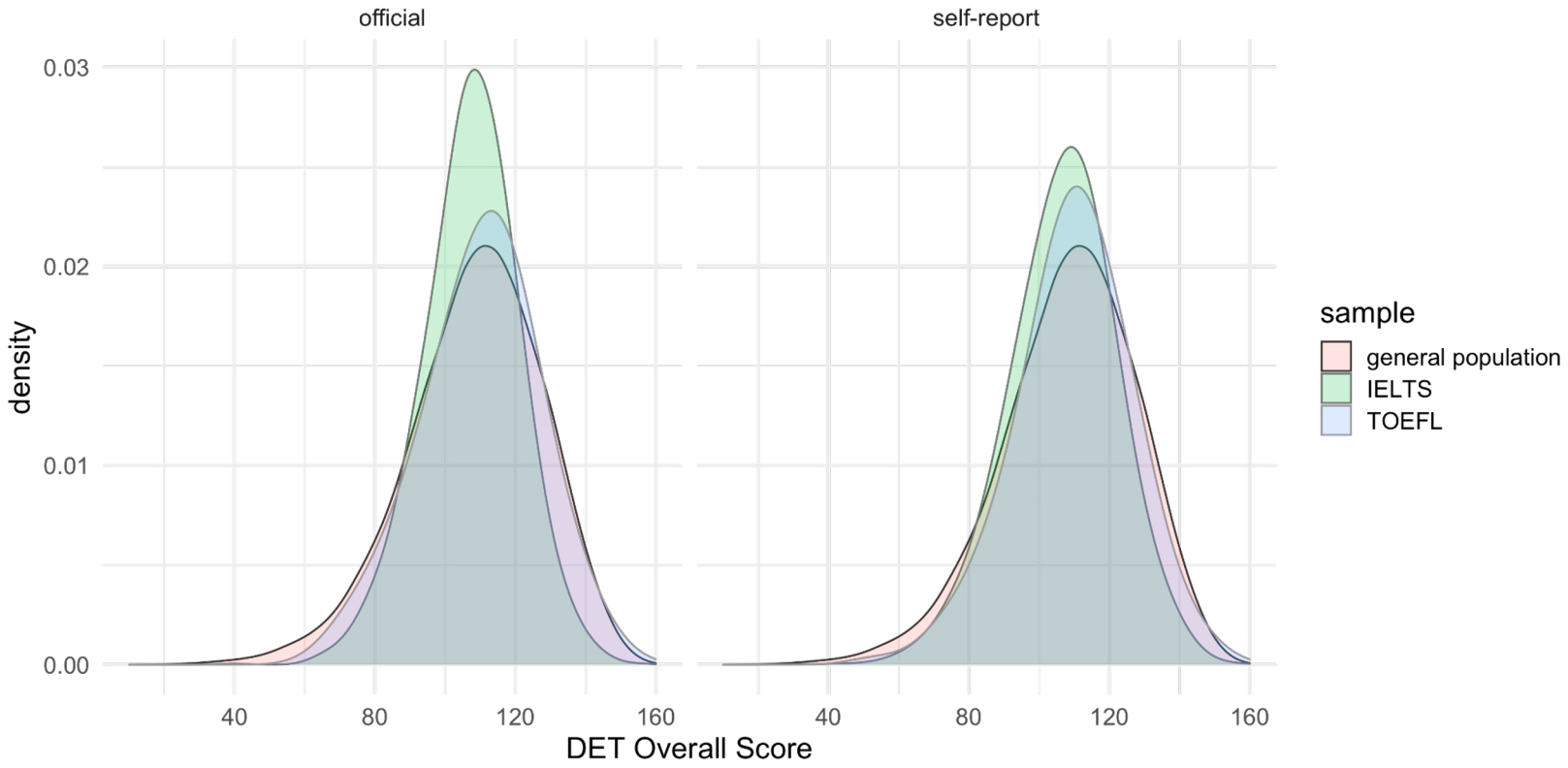


Problem 2: Self-Selection

Concordances should hold in a specified distribution:

- People who reported IELTS and TOEFL scores do not match the overall DET distribution.
- The concordance should hold in the population defined by the marginal DET distribution.

DET Score Distributions/Densities



Additional Issues

The following are additional practical issues that we had to address:

1. Some people took the DET and/or alternate assessments more than once.
2. Several test-takers had outlying and/or influential score pairs.
3. We wanted to quantify sampling variability of the entire process.

(Possible) Solutions

Solution 1: Bias Adjustment

Assumption:

Bias applies not just to paired score but to all self-report scores!

For TOEFL (also applies to IELTS):

- At each DET score point, there is a distribution of self-report TOEFL scores.
- This (conditional) distribution is positively biased based on the estimates from the paired data.
- We want to adjust this (conditional) distribution downward!

How do we do this? MDIA!

Solution 1: MDIA

We compute an adjusted distribution using MDIA (Haberman, 1984).

1. Calculate new N* at each TOEFL score for a given DET score.
2. This new N* assumes that:
    a. The support of the distribution is identical to the original distribution.
    b. The overall counts remains unchanged.
    c. The mean of the new distribution has the desired bias adjustment.
3. How do we choose our new distribution (e.g., N*)?
    a. Minimize KL-divergence between old and new distributions.
    b. Add Lagrange multipliers to match constraints in 2.

Solution 2: Weighting

How did we weight the data to match the marginal DET population?

- Let x be scores on the DET and y be scores on an external assessment
- Let p be a PMF in the entire DET population and p* be a PMF in the population who report scores on the external assessment.

DET score in population of all DET test takers

$$p(y) = \sum_x p(y \cap x) = \sum_x [p(y|x) \times p(x)] \approx \sum_x [p^\star(y|x) \times p(x)] = \left[\frac{p^\star(y \cap x)}{p^\star(x)} \times p(x)\right]$$

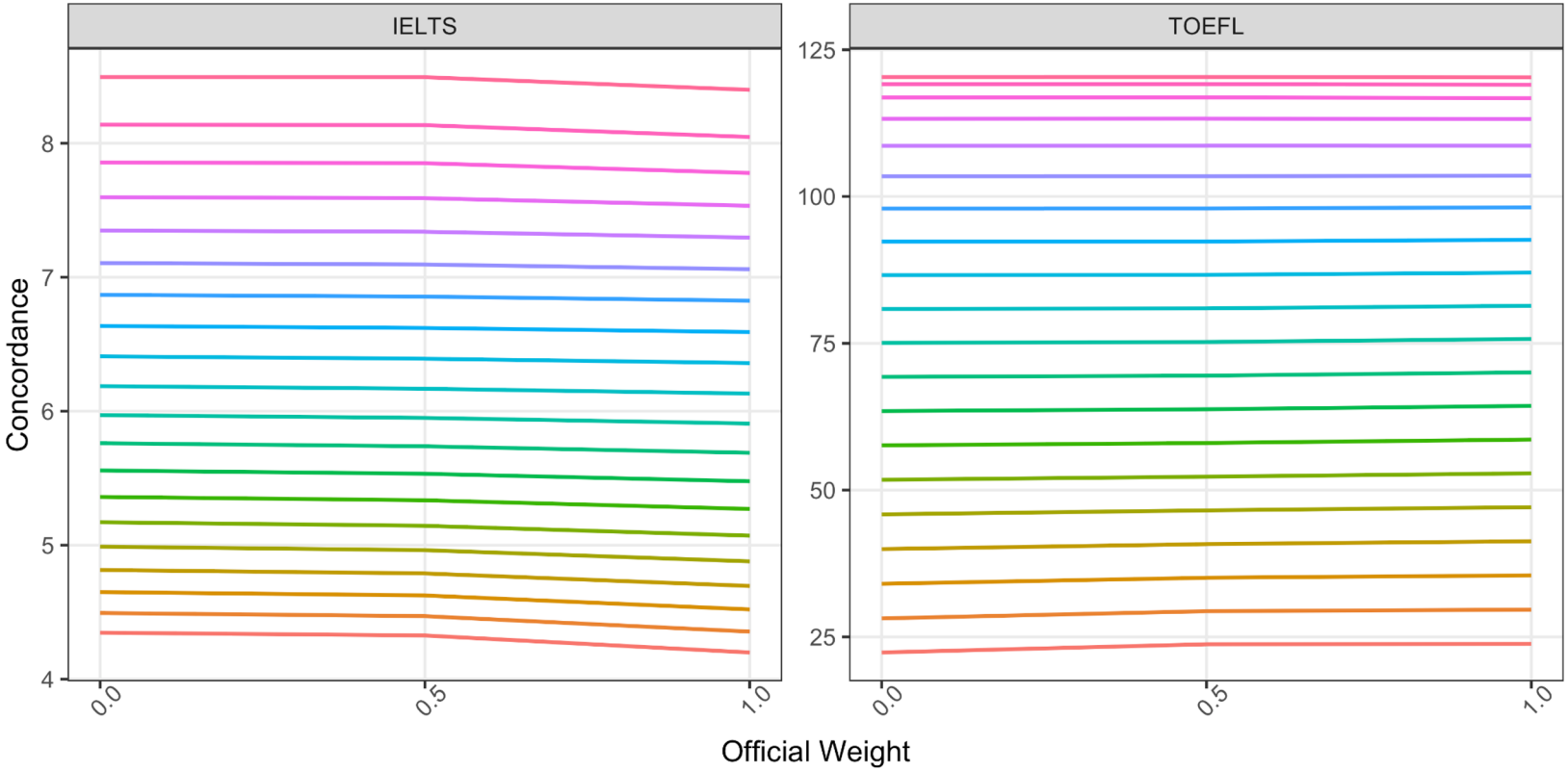
TOEFL/IELTS observed score

DET score in population who reports TOEFL/IELTS score

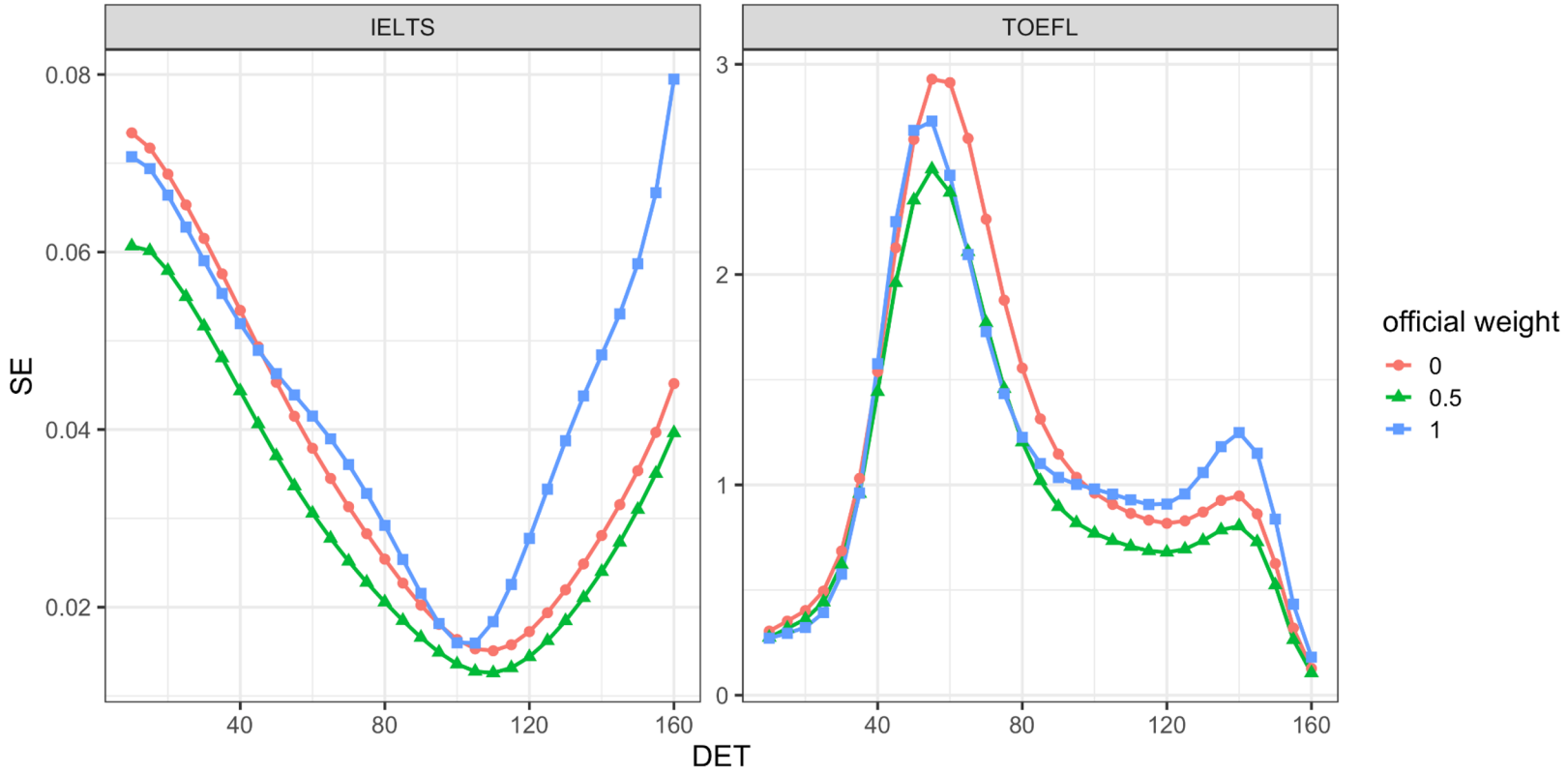
- We then obtain frequencies by multiplying by N_x .

Results

Combining Data Concordance



Combining Data Standard Errors



Model-Based DET Correlations at Each Step

	Official	Self	Combined	Weighted
TOEFL	.71	.69	.70	.74
IELTS	.65	.64	.64	.72

Final TOEFL Concordance

DET	TOEFL
0-60	0–23
65	24–29
70	30–34
75	35–40
80	41–46
85	47–52
90	53–58
95	59–64
100	65–69
105	70–75
110	76–81

DET	TOEFL
115	82–86
120	87–92
125	93–97
130	98–103
135	104–108
140	109–112
145	113–116
150	117–118
155	119
160	120

Final IELTS Concordance

IELTS	DET
0-4	10–60
4.5	65–75
5	80–90
5.5	95–100
6	105–115
6.5	120–125
7	130–135
7.5	140–145
8	150–155
8.5	160
9	160

Wrap-up

What We Wrestled With

The following questions we had to wrestle with:

1. How do you combine two sources of data where one has response bias and the other is very small?
2. How do you adjust the observed population to be similar to the desired population?

All methods returned nearly identical results after collecting more official data!