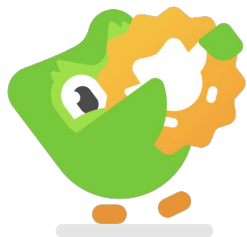


Considering inter-test relationships in high-stakes admissions testing: The case of English

Ramsey Cardwell, Steven Nydick, & J.R. Lockwood

Duolingo



Agenda

- Motivation of the study
- Correlational analyses
- Concordance analyses
- Conclusion / implications

What is the Duolingo English Test?

- Online, at-home, on-demand ELP test
- ~1 hour to complete, due to adaptive administration
- Mixture of item types & measurement approaches
 - Closed-ended: C-test, dictation, MCQ, highlight
 - Open-ended: picture description, extended speaking & writing
- Synchronous security measures & asynchronous human proctoring
- Accepted by >4,000 postsecondary institutions/programs worldwide

Why look at inter-test relationships?

- All three tests are being used for the same purpose(s)
 - Postsecondary admissions (primary), English course placement
- Criterion validity evidence for test score use (Standards, Kane)
 - e.g., Chapelle (2008), ETS (2010), Ihlenfeldt & Rios (2022)
- Convergent validity evidence for test construct
 - e.g., multitrait-multimethod matrix (Campbell & Fiske, 1959)

Why look at inter-test relationships?

- Limitations of correlations
 - Necessary but not sufficient evidence of construct similarity
 - Restriction of range, attenuation, non-linear relationships
- Limitations of concordances
 - Do not imply tests/scores are interchangeable
 - Are population-dependent

The 3 tests in this study

	DET	TOEFL iBT	IELTS Academic
<i>Primary use</i>	Postsecondary admissions		
<i>Time</i>	~1 hour	~3 hours	2 hours 45 mins
<i>Adaptive?</i>	Yes (partially)	No	
<i>Delivery</i>	Computer, at-home	Computer, center/home	Computer/paper, center
<i>Item types</i>	Selected response, gap-fill, extended monologic speaking & writing	MCQ, extended monologic speaking & writing	MCQ, speaking interview, extended writing
<i>Subscores</i>	Literacy, Conversation, Comprehension, Production	Speaking, Writing, Reading, Listening	

1:00



Type the statement that you hear



Your response

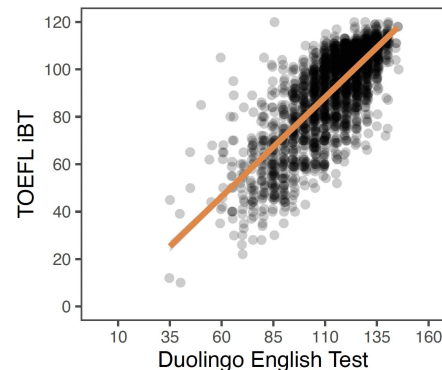
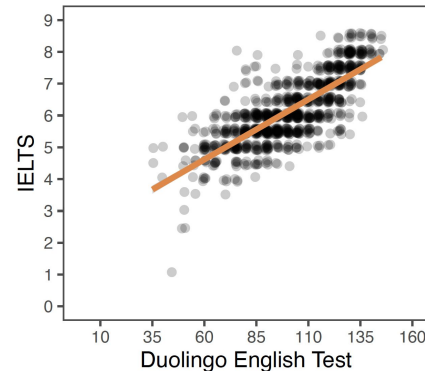
Number of replays left: 2

NEXT

Correlational analyses

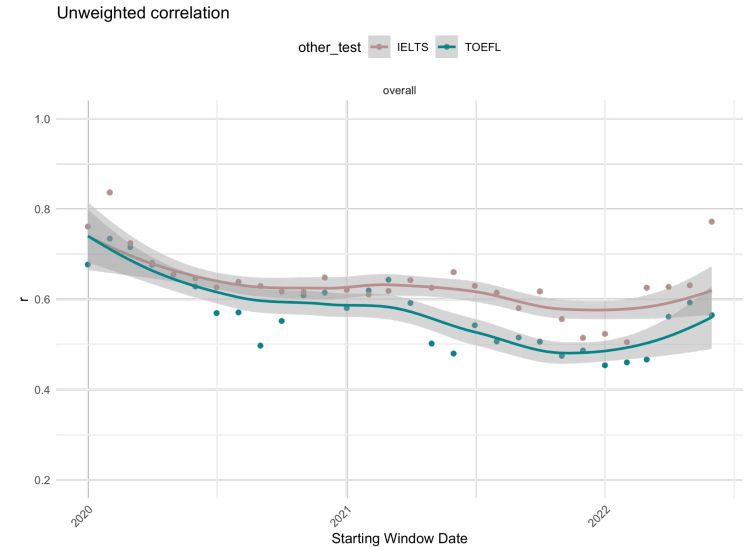
Previously reported total-score correlations

- IELTS–TOEFL $r = .73$ (ETS, 2010)
 - N = 1,153 (official score reports)
- DET–IELTS $r = 0.78$ (LaFlair & Settles, 2019)
 - N = 991 (self-report data)
- DET–TOEFL $r = 0.77$ (LaFlair & Settles, 2019)
 - N = 2,319 (self-report data)



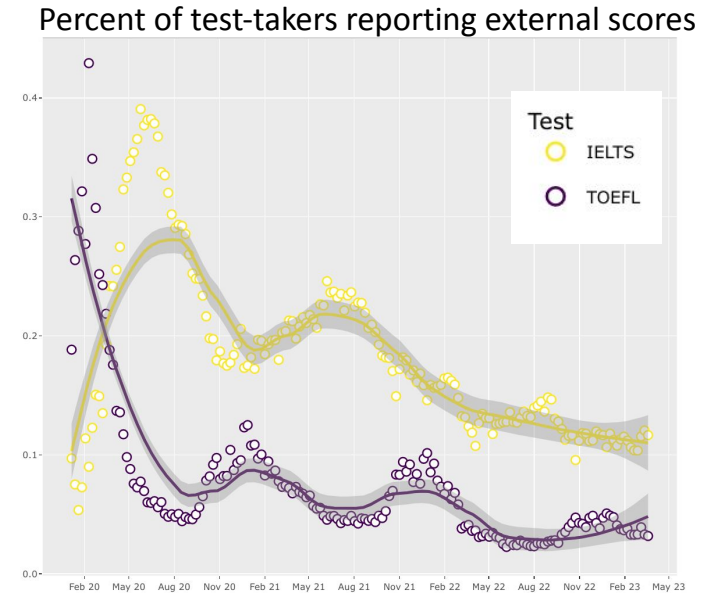
What are the observed inter-test correlations?

- DET test-takers can self-report TOEFL/IELTS scores at test end
 - Produces a continuous data stream
- Research questions:
 - Why do the observed correlations fluctuate over time?
 - How can the causes of fluctuation be accounted for?
 - What are the best estimates of DET–IELTS/–TOEFL correlation coefficients?



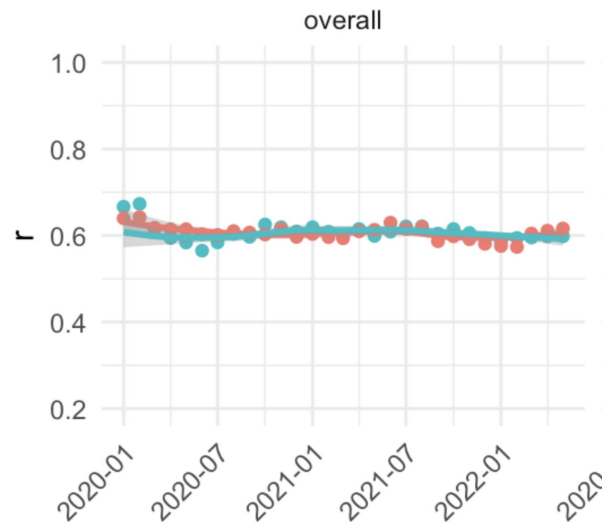
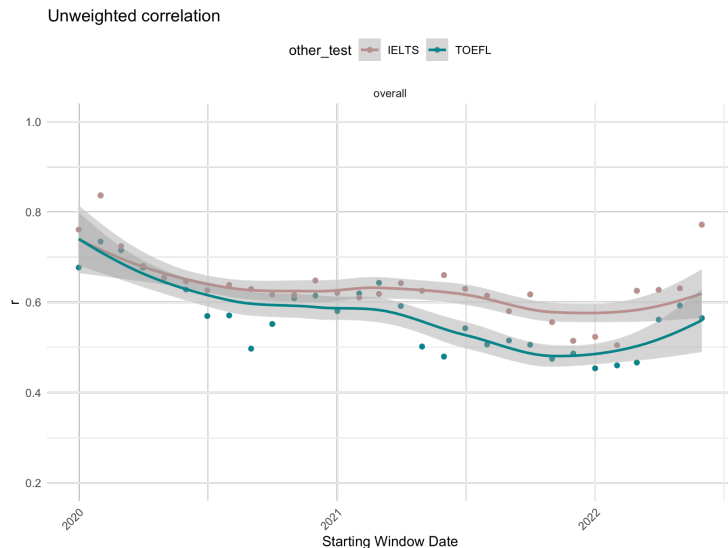
What has changed in recent years?

- Test-taker country of origin
 - China ↓, India ↑, ROW ↑
- Test-taker intent
 - ~15% postgrad (2020) ➡ ~38% (2022)
- Proportion of TTs reporting other scores
 - ~42% (Jan 2020) ➡ ~22% (Jan 2022)
- Population & self-reporting changes ➡ reduced variance



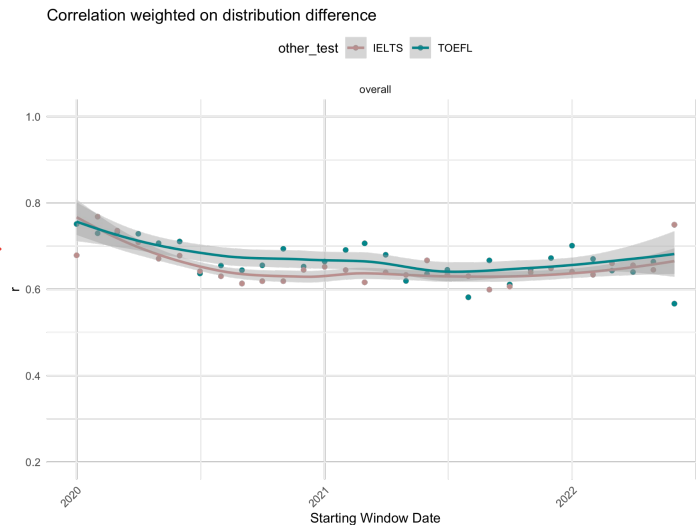
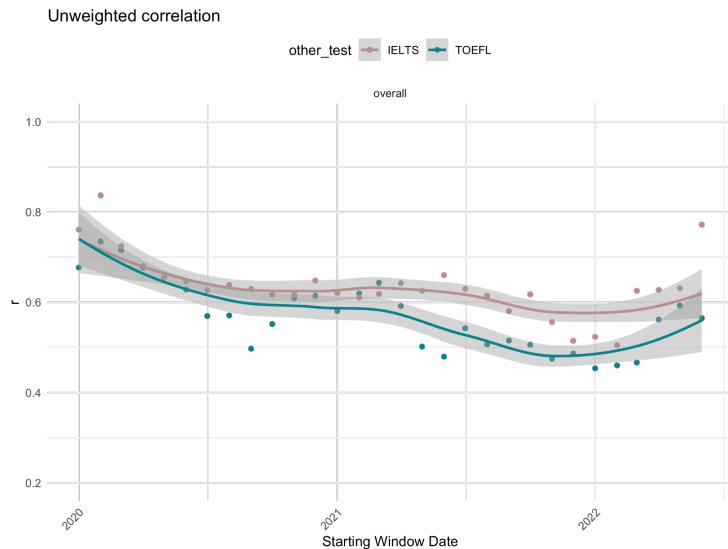
Adjusting by multiple imputation

- All certified tests Jan 2020 – May 2022
 - Impute IELTS/TOEFL scores based on DET score, country, language, & age



Adjusting by weighting

- All certified tests Jan 2020 – May 2022
 - Weight data to match population distributions of TOEFL/IELTS scores



Correlation results

- Imputation and weighting similarly reduced variation in the correlation coefficients over time
 - Supports sampling bias as a primary cause of the observed variation
- Overall adjusted correlations (without data cleaning)
 - $r^* = 0.63$ (IELTS); $r^* = 0.65$ (TOEFL)
- Limitation of only using self-report data
 - Next steps were using verified TOEFL/IELTS score data

Concordance analyses

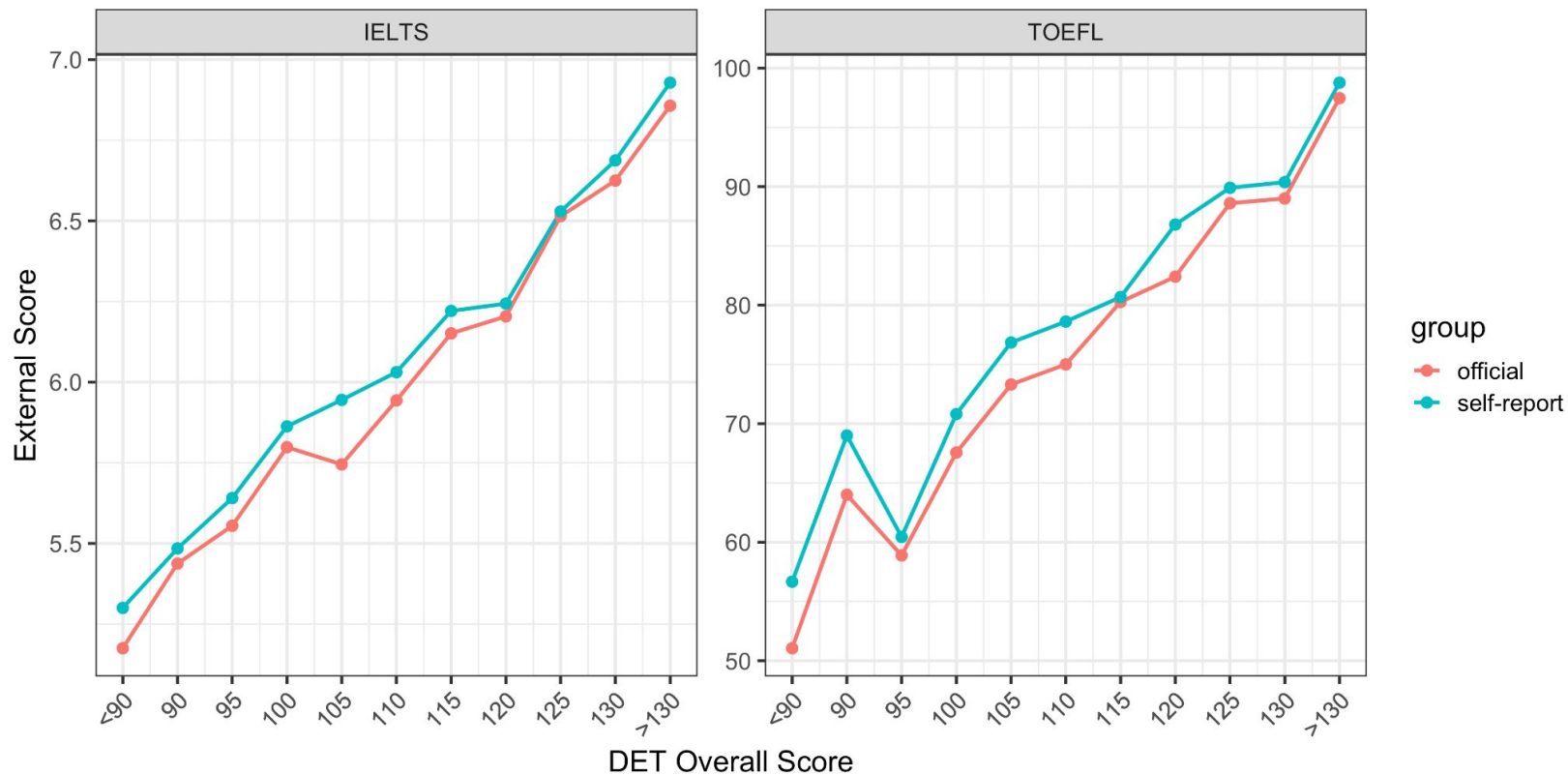
Data collection process

1. **Self-report:** DET test takers self-report scores at test end
2. **Official:** We asked for official score reports from DET test takers.
3. We planned to combine official with self-report scores to determine TOEFL/IELTS scores for a given DET score.

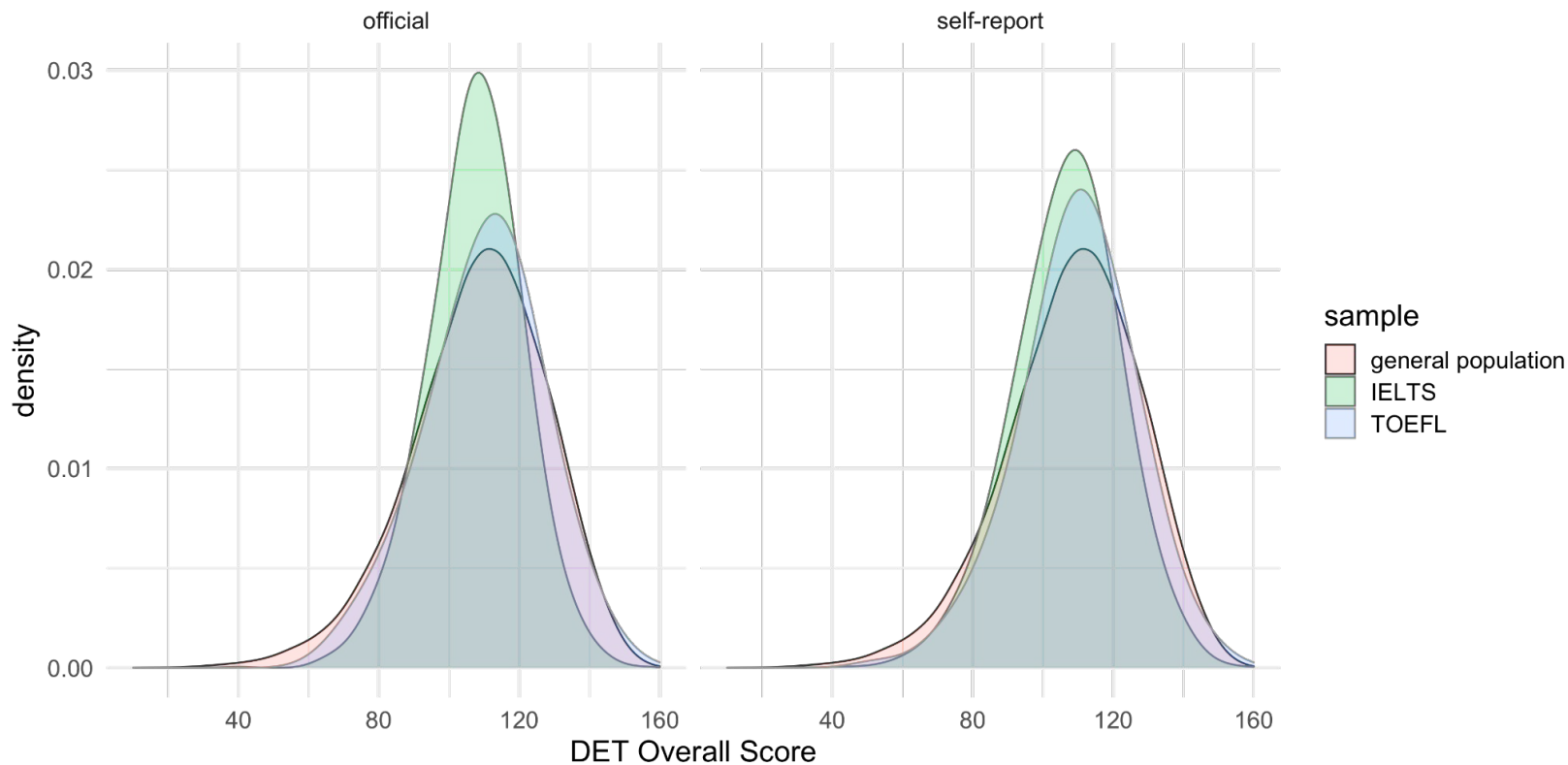
	IELTS	TOEFL
Self-report	3,806	1,276
Official	1,643*	328*

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

Challenge 1: Reporting bias



Challenge 2: Self-selection



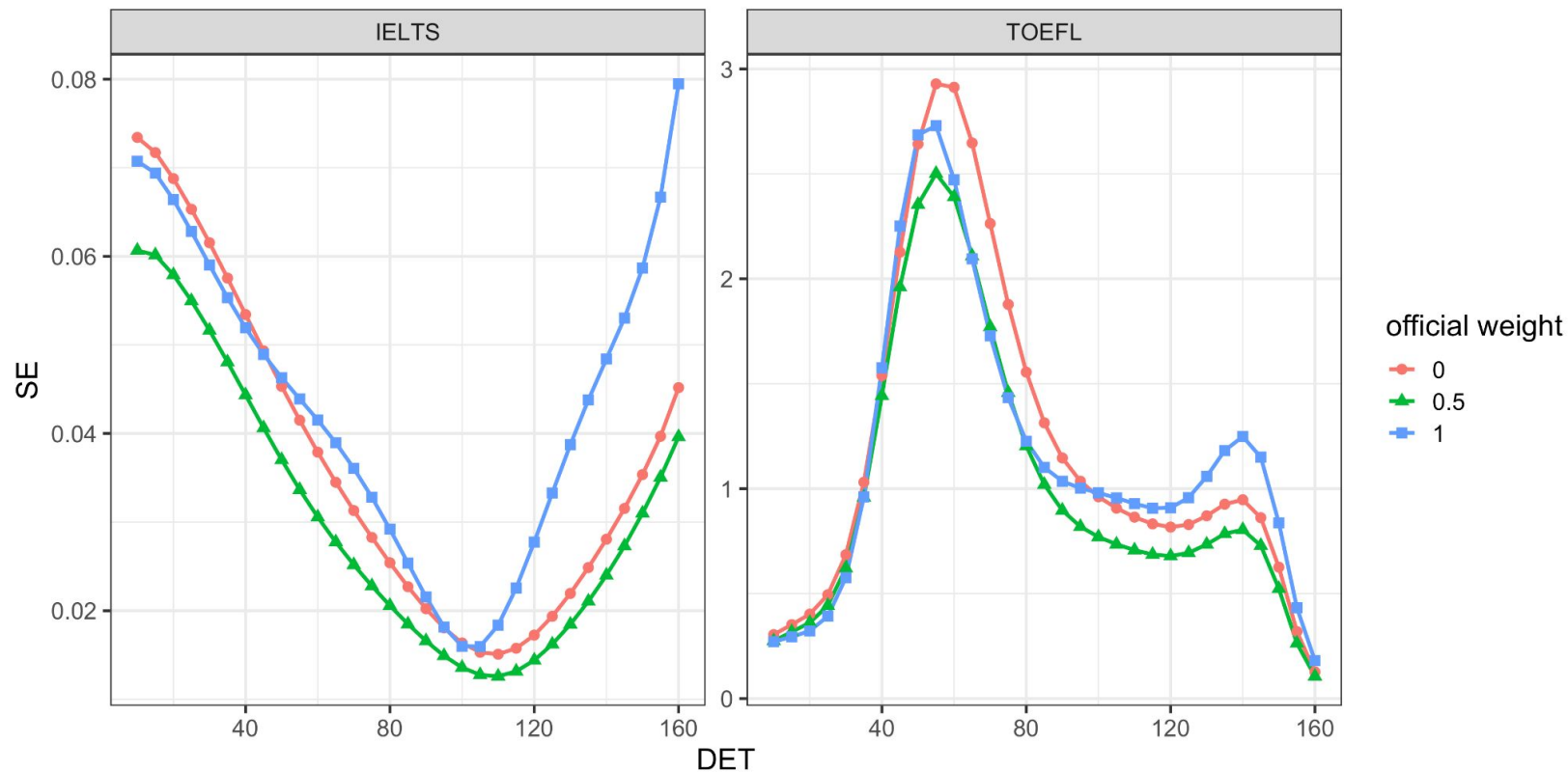
There were additional considerations

1. Several people took the DET and/or alternative assessments multiple times. We needed to figure out which scores we should keep.
2. Several test takers had outlying and/or influential score pairs.

We performed the following concordance

1. Disambiguate DET and external score pairs
 - a. Remove scores where DET and external score **>3~4 months apart**
 - b. Remove **outliers** (~1%)
 - c. Keep **most recent score** for each unique test-taker
2. Adjust **self-report data** for reporting bias (Haberman, 1984)
3. Combine **self-report and official** data
4. Weight data to match **overall DET score distribution**
5. Smooth data using **log-linear smoothing methods**
6. Equate data using **kernel** equating methods (von Davier et al., 2004)

Weighting led to smaller standard errors



Each step resulted in updated correlations

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

Which led to the final concordances

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

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

Conclusions & implications

Conclusions

- Correlations are stable over time after accounting for sampling variability and periodicity
- Concordances are stable across methodological choices and addition of new data

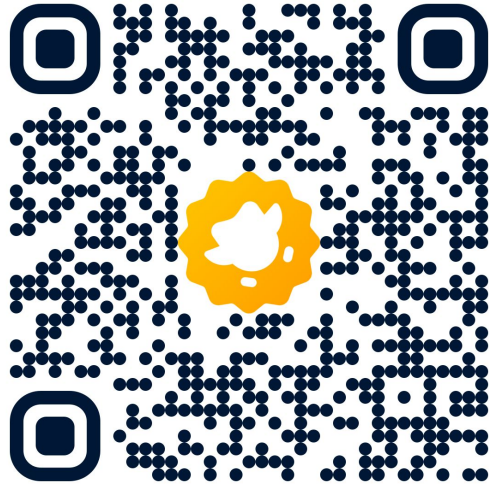
Implications for test developers

- Test developers should continue to monitor changes to language proficiency tests over time
- Population changes can cause variation in observed relationships

Implications for test score users

- English proficiency testing:
 - The DET shows strong, robust relationships to other ELP tests
- Language testing more broadly:
 - Score users should also be aware of factors affecting inter-test relationships when using multiple tests for the same purpose
 - Things that can affect inter-test relationships include population changes, sampling bias

¡Gracias!



References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, 12, 971-988.

Von Davier, A. A., Holland, P. W., and Thayer, D. T. (2004). *The Kernel Method of Test Equating*, New York, NY: Springer New York.