Differential Item Functioning between **Pilot & Formal Administration** of the Norwegian Language Test **"Norskprøven"**

Ga Young Yoon
University of Oslo

Chia-Wen Chen
University of Oslo

Tor Midtbø
Skills Norway

CONTEXT: *Norskprøven* (Norwegian language test):

- Four tests: Reading, listening, writing & oral test
- Multistage testing design for reading and listening comprehension test (see Figure 1)
- Formal test administered four times a year (e.g. official certification or application for Norwegian citizenship)
- Pilot test administered twice a year to allow practice moments for test-takers & for data to support item calibration for the formal test administration & test assembly

Pilot: Low Stakes vs Formal: High Stakes

- Items can function differently under different test administration circumstances
- Potentially due to differences in test engagement test (Ulitzsch et al., 2019) or to item position or context changes between test forms (Davey & Lee, 2011)
- Such differential item functioning (DIF) can
  ➔Impact reliability & validity of test scoring
  ➔Complicate fair comparison across different tests

METHOD

- *Data*: item response data on 56 items of the reading comprehension test that were in common across the test forms for the group of respondents that took the formal high-stakes test & the respondent group from the low-stakes pilot test
- *DIF analysis* (See Figure 2)*:* IRT-based Likelihood Ratio tests (e.g., Kim, 2001) with a two-step purification procedure of the matching Criterion (Holland & Thayer, 1988; Lee & Geisinger, 2016)

RESULTS

- More correct response for formal test group compared to pilot test group (see Figure 5)
- *10 DIF items with large effect (see Figures 3-6.)*
- *DIF items discriminate more strongly for formal than for pilot test group ➔ less information in Pilot group*
- *DIF items' difficulties change less systematically and to a lesser degree*

CONCLUSION

Our results corroborate findings from previous studies that stake differences can lead to potential shifts in item parameters. The lower discrimination parameters in the Pilot test indicate the presence of more noise, potentially connected to factors such as low motivation. A closer qualitative follow-up is suggested for the items that were shown to fall prone to DIF to identify item features that contribute to item drift from low-stakes to high-stakes test administration

Methodological Procedure and Results

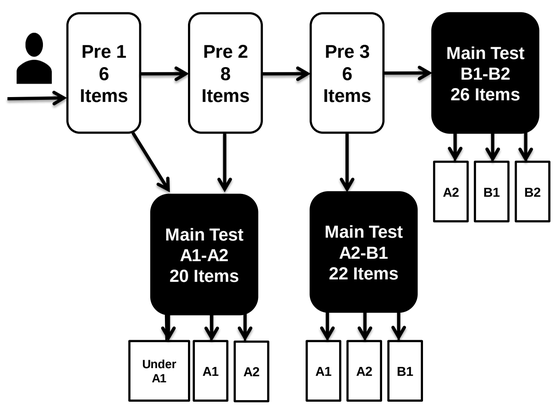

Figure 1. Multistage Routing Structure for Reading Comprehension Test in Norskprøven

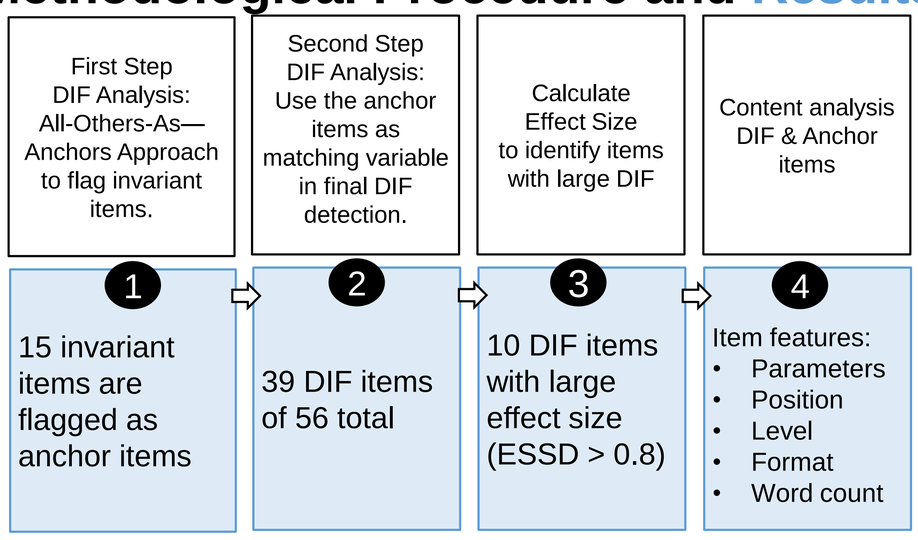

Figure 2: Overview of Main Analysis – Method and results

DIF Item Tendencies: More discriminating under high-stakes

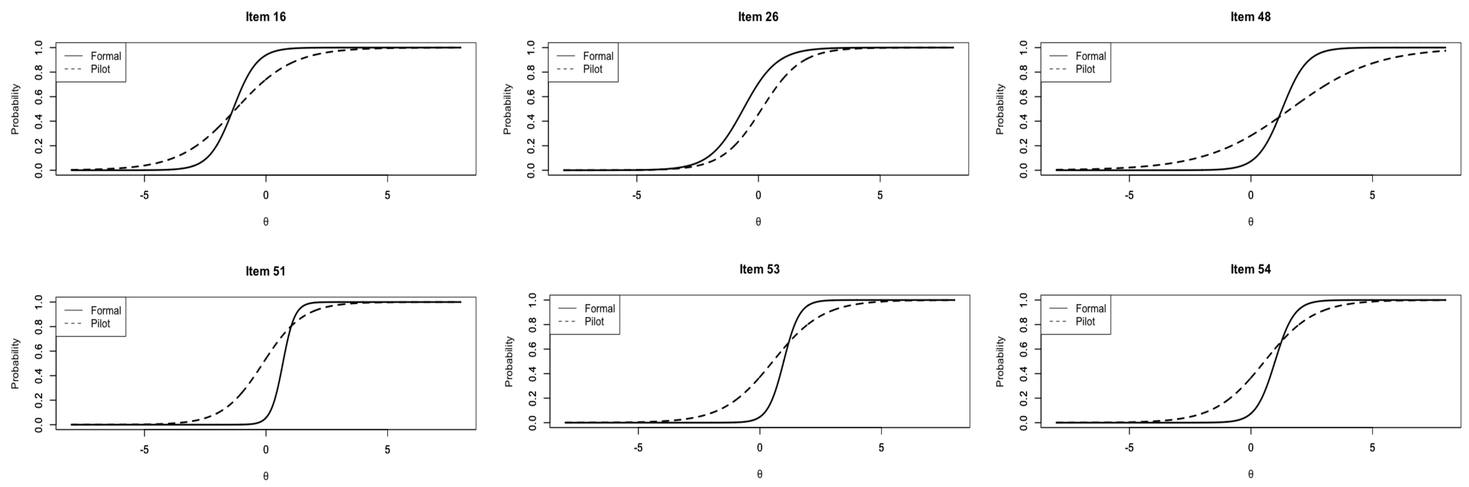

Figure 3. Some Examples of Item Characteristic Curves (ICC) on DIF items

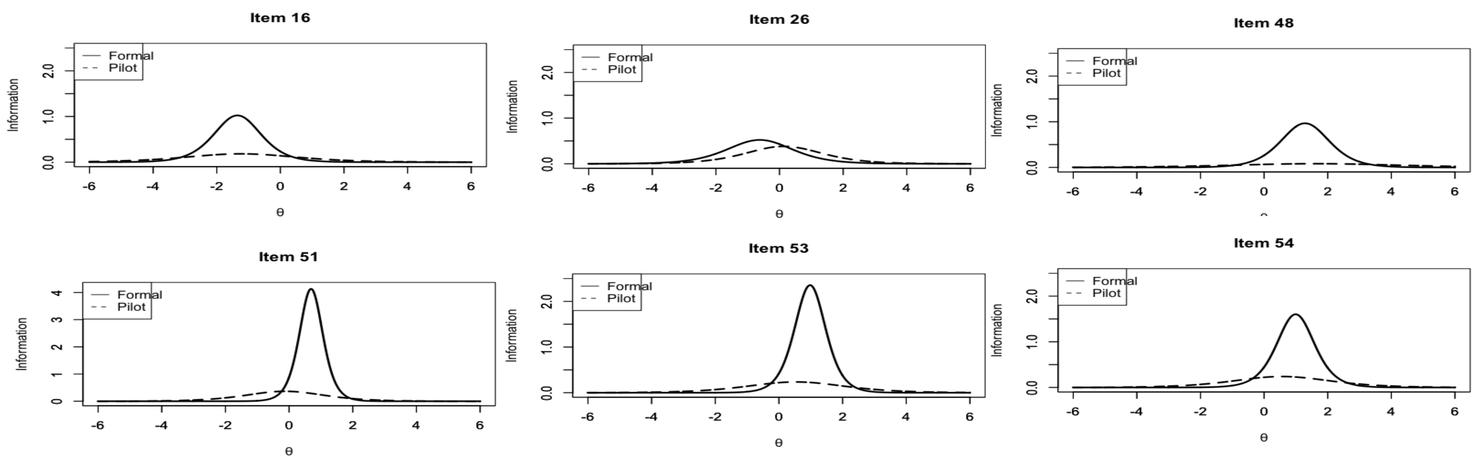

Figure 4 . Some Examples of Item Information Function on DIF items

Test Tendency: Easier & More informative under high-stakes

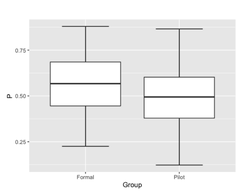

Figure 5. Proportion of Correct Responses per Item in Formal & Pilot Tests

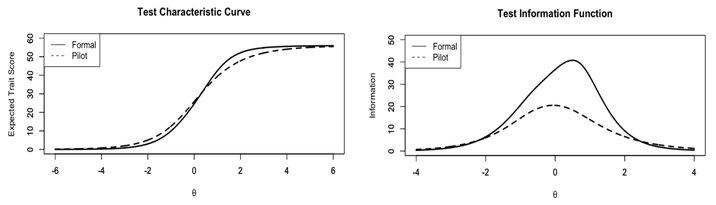

Figure 6. Test Characteristic Curve & Test Information Function

Item Content

Item	Format	Word count	Pilot test form	Formal test form	P (Formal/Pilot)	a (Formal/Pilot)	b (Formal/Pilot)
2	MC short	49	A1	Pretest 1	0.81/0.55	1.96/0.66	-1.11/-0.33
7	MC short	45	A2	Pretest 2	0.69/0.33	2.13/1.28	-0.25/0.65
8	Click word	53	A2	Pretest 2	0.77/0.45	1.32/1.70	-0.77/0.04
16	Choose Text	-	A2	A1/A2-1	0.69/0.66	2.02/0.85	-1.35/-1.23
26	Calender	26	A1	A1/A2-2	0.44/0.44	1.45/1.24	-0.62/0.13
48	Voice of opinion	296	B1	B1/B2-1	0.46/0.37	1.97/0.57	1.29/1.63
51	Which Person?	543	B2	B1/B2-2	0.73/0.66	4.07/1.21	0.69/-0.13
53	Which Person?	543	B2	B1/B2-2	0.57/0.54	3.07/0.97	0.98/0.53
54	Which person?	543	B2	B1/B2-2	0.56/0.50	2.53/0.98	0.99/0.56
56	Which Person?	543	B2	B1/B2-2	0.34/0.36	2.02/0.63	1.53/1.68

Table 1. Content Characteristics for 10 DIF items

For any comments or further information

g.y.yoon@uv.uio.no

REFERENCES

Davey, T., & Lee, Y.-H. (2011). Potential Impact of Context Effects on the Scoring and Equating of the Multistage Gre® Revised General Test. *ETS Research Report Series*, *2011*(2), 1–44. https://doi.org/10.1002/j.2333-8504.2011.tb02262.x

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*(1), 89–114.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, *1986*(2), i-24.

Lee, H., & Geisinger, K. F. (2016). The Matching Criterion Purification for Differential Item Functioning Analyses in a Large-Scale Assessment. *Educational and Psychological Measurement*, *76*(1), 141–163. https://doi.org/10.1177/0013164415585166

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728–743. https://doi.org/10.1037/a0019986

Ulitzsch, E., Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12188