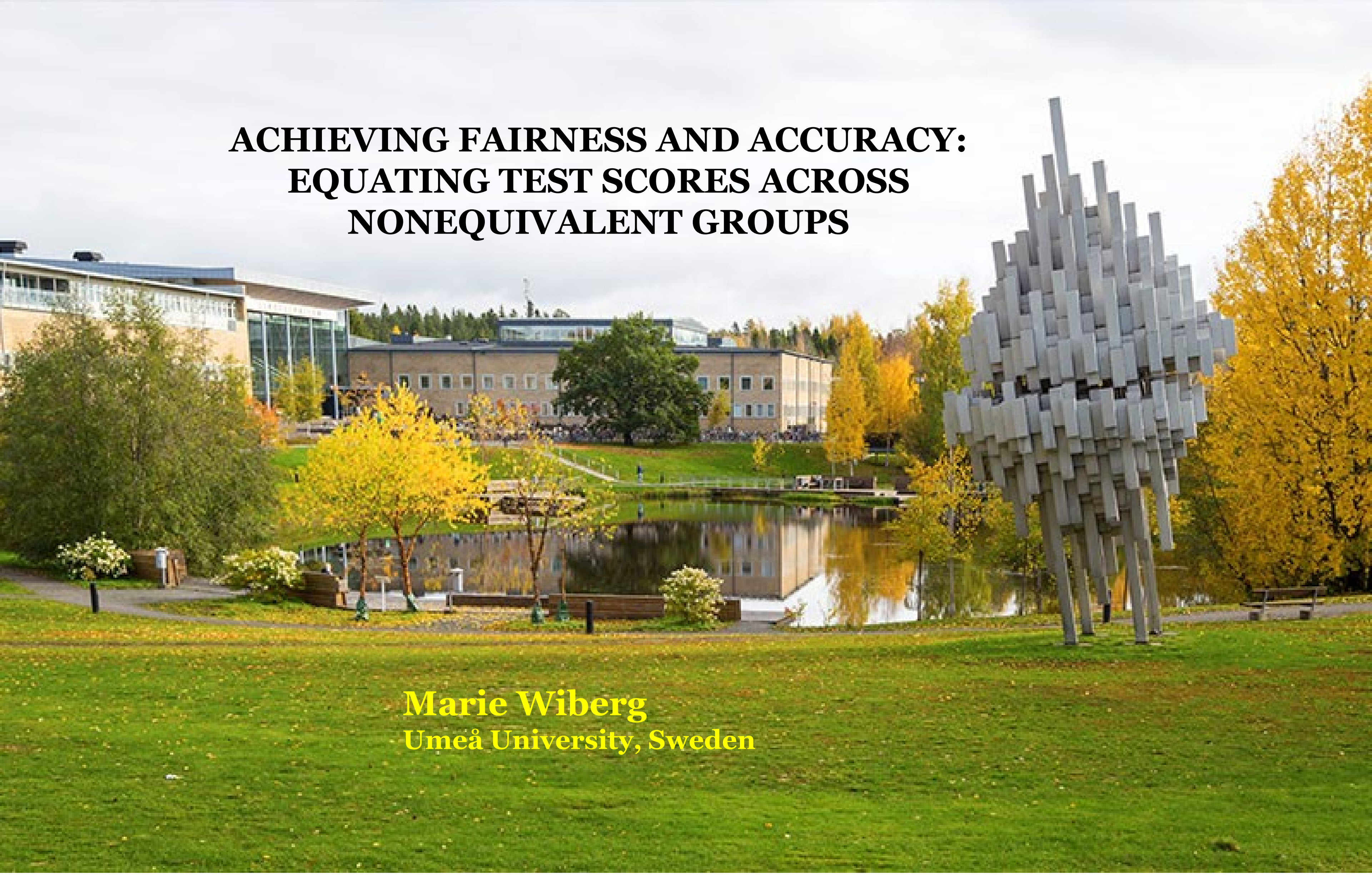


ACHIEVING FAIRNESS AND ACCURACY: EQUATING TEST SCORES ACROSS NONEQUIVALENT GROUPS

Marie Wiberg
Umeå University, Sweden

UMEÅ UNIVERSITY

DIFFERENT TEST SCORES

National tests

SAT®

ACT®

PISA Programme for International Student Assessment

IEA TIMSS

FAIRNESS – IN WHAT SENSE?

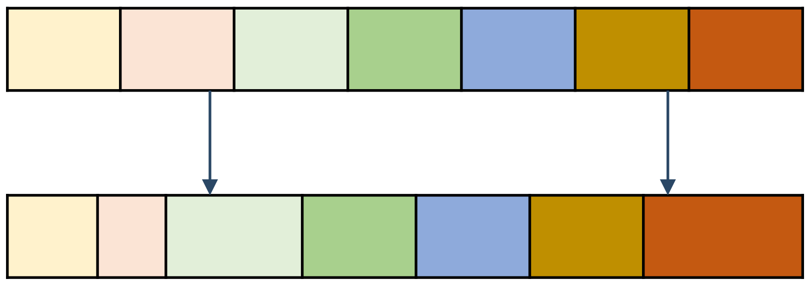
ACCURACY

Content

Methods

Now and over time

EQUATING TEST SCORES



Form X

$$\hat{\varphi}_Y(x) = G_Y^{-1}\left(F_X(x)\right)$$

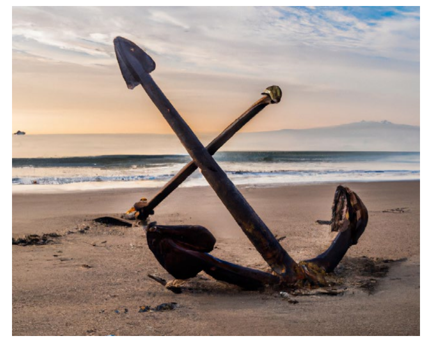
Form Y

Equating as a family of statistical models and methods that are used to make test scores comparable among two or more versions of a test, so that scores on these different test forms, may be used interchangeably (González & Wiberg, 2017).

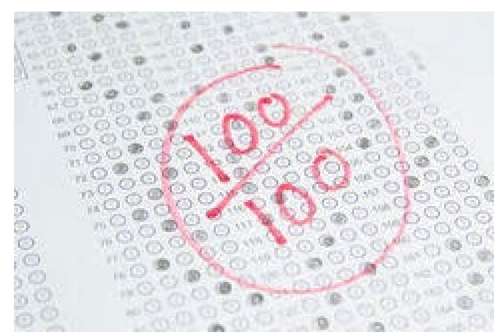
COMMON OBJECTS

- Test takers

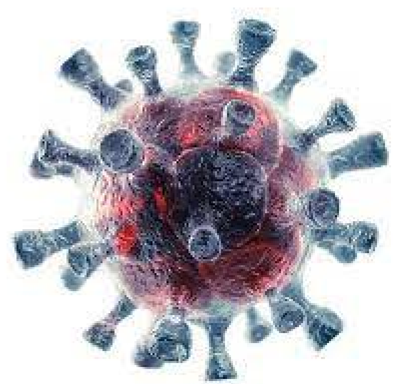
- Common (anchor) items



- Covariates

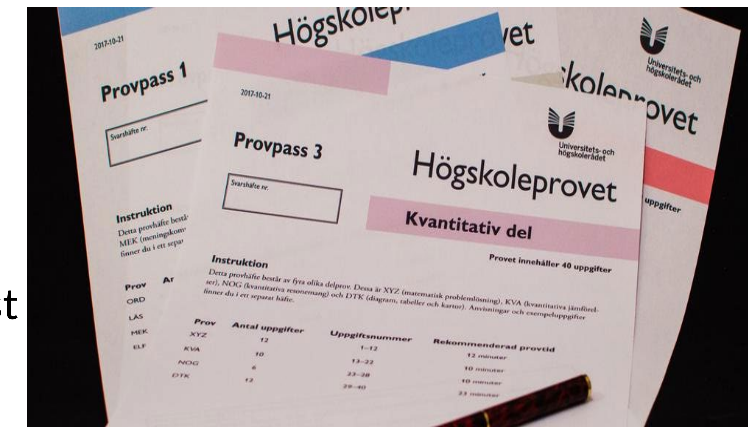


NONEQUIVALENT GROUPS



SWEDISH SCHOLASTIC APTITUDE TEST (SWESAT)

- High stake college admissions test administered twice a year
- Two subtests:
  - **Verbal**, which emphasizes word and reading comprehension.
  - **Quantitative**, which emphasizes mathematical knowledge and the ability to interpret and understand graphic information.
- 160 multiple-choice questions, binary scored
- Five test parts, each containing 40 questions:
  - two verbal parts,
  - two quantitative parts,
  - one with try-out items or an external anchor test

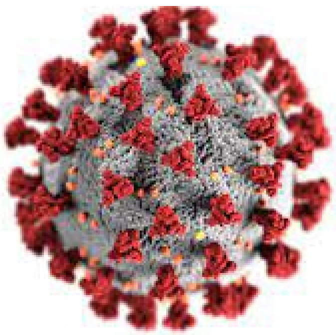


- Test result is valid for 8 years
- After each administration, tests are equated, and the test score is transferred to a standardized scale (0.0-2.0).
- About 60,000 test takers/administration

COVID EFFECTS: SWESAT

- Longer time SweSAT is valid (8 years).
- Only test taker without valid test results could take the test during COVID.
- Limited number of seats for test takers.

- Other test taking groups – compared with previous years.
- More people wanted to study.
- More unemployed people.



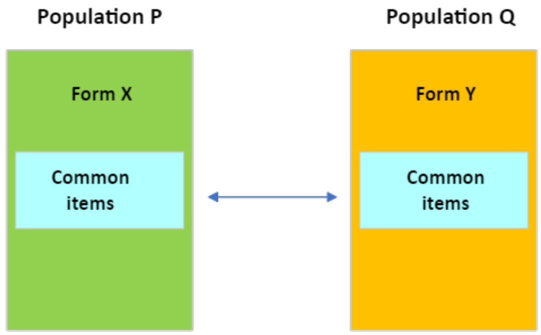
How to handle this to preserve fairness and accuracy?

TWO USEFUL APPROACHES

1. Anchor test

The **N**on **E**quivalent groups with **A**nchor **T**est (**NEAT**) design is used to disentangle **test form differences** from **group differences**.

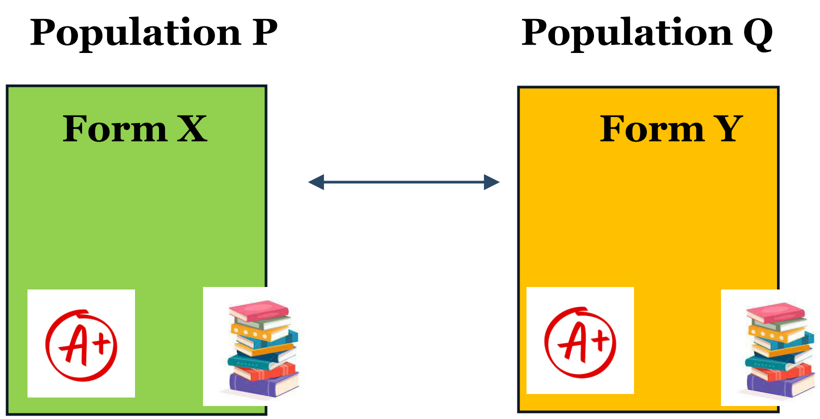
- To which group?
- Behaviour of items over time



2. Covariates

The **N**on **E**quivalent groups with **C**ovariates (**NEC)** design is used to disentangle **test form differences** from **group differences**.

- Which covariates?
- How to use them?



KERNEL EQUATING

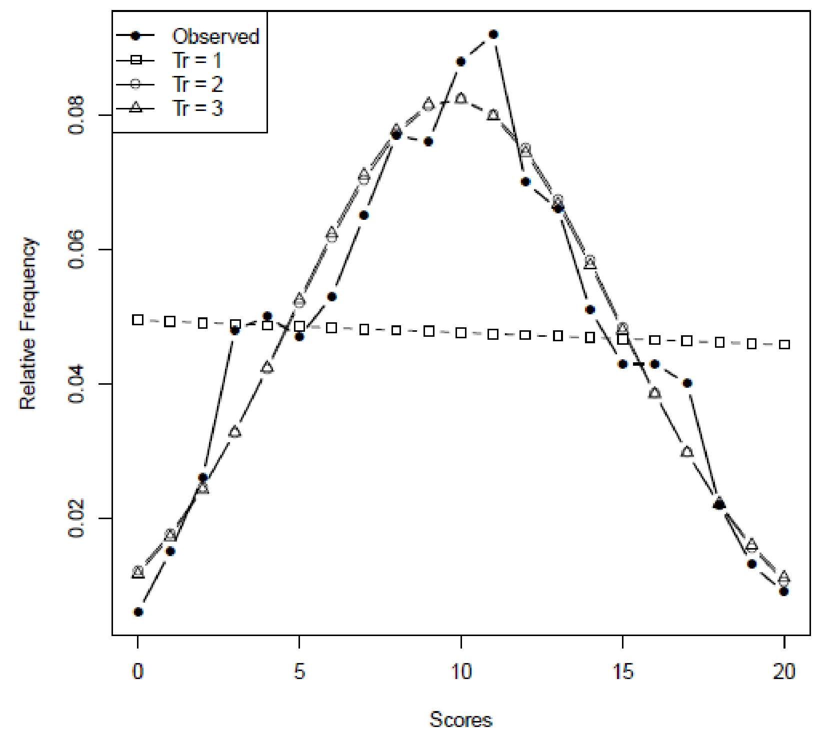
1. PRESMOOTHING (e.g. with loglinear models)

2. ESTIMATING SCORE PROBABILITIES

3. CONTINUIZATION (most test scores are discrete)

4. EQUATING

5. EVALUATION MEASURES
(e.g. standard errors and bias)



UMEÅ UNIVERSITY

KERNEL EQUATING

1. PRESMOOTHING

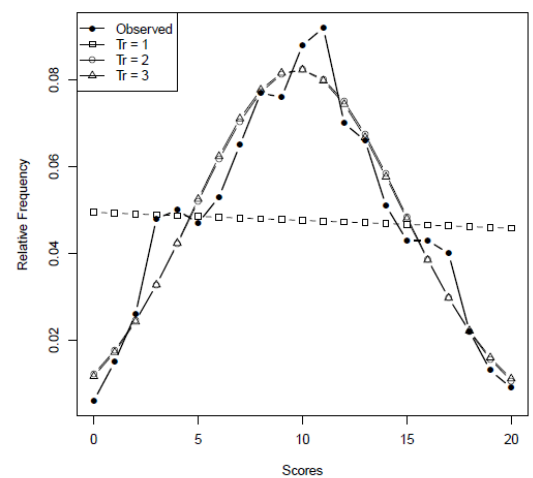
- **E**quivalent **G**roups (**EG**) design $\log(p_{jk}) = \beta_0 + \sum_{i=1}^{T_r} \beta_i^X (x_j)^i$

- **N**on**E**quivalent groups with **A**nchor **T**est (**NEAT**) design

$$\log(p_{jk}) = \beta_0 + \sum_{i=1}^{T_r} \beta_i^X (x_j)^i + \sum_{i=1}^{T_s} \beta_i^A (a_l)^i + \beta_{il}^{XA} x_j^i a_l^l$$

- **N**on**E**quivalent groups with **C**ovariates (**NEC**) design

$$\log(p_{jk}) = \beta_0 + \sum_{i=1}^{T_r} \beta_i^X (x_j)^i + \sum_{i=1}^{T_s} \beta_i^{Grade} (grade_l)^i + \beta_{il}^{XGrade} x_j^i grade_l^i$$



4. KERNEL EQUATING: ANCHOR TEST

NEAT design: Assume that the conditional distribution of X or Y given anchor test A, are the same in both populations *P* and *Q*:

$$P\left(X = x_j \middle| A = a_m, P\right) = P\left(X = x_j \middle| A = a_m, Q\right)$$

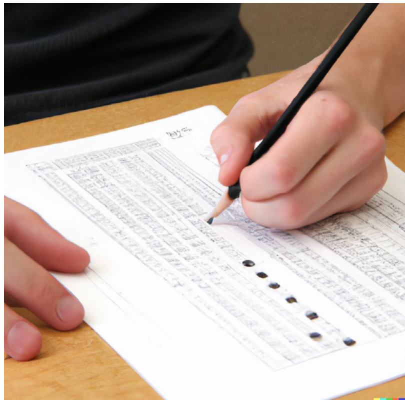
$$P\left(Y = y_k \middle| A = a_m, P\right) = P\left(Y = y_k \middle| A = a_m, Q\right)$$

Poststratification Equating (PSE) $T = wP + (1-w)Q$

$$\hat{\varphi}_{Y(PSE)}\left(x\right) = \hat{G}^{-1}_{Th_Y}\left(\hat{F}_{Th_X}\left(x\right)\right)$$

Chained Equating (CE)

$$\hat{\varphi}_{Y(CE)}\left(x\right) = \hat{G}^{-1}_{Qh_Y}(\hat{H}_{Qh_A}(\hat{H}^{-1}_{Ph_A}(\hat{F}_{Ph_X}(x))))$$



UMEÅ UNIVERSITY

KERNEL EQUATING: RAW COVARIATES

NEC design: Assume that the conditional distribution of X or Y given covariates Z, are the same in both populations **P** and *Q*:

$$P\left(X = x_j \middle| Z = z_l, P\right) = P\left(X = x_j \middle| Z = z_l, Q\right)$$

$$P\left(Y = y_k \middle| Z = z_l, P\right) = P\left(Y = y_k \middle| Z = z_l, Q\right)$$

Postratification Equating (PSE) $T = wP + (1-w)Q$

$$\hat{\varphi}_{Y(PSE)}\left(x\right) = \hat{G}_{Th_Y}^{-1}\left(\hat{F}_{Th_X}\left(x\right)\right)$$

Chained Equating (CE)

$$\hat{\varphi}_{Y(CE)}\left(x\right) = \hat{G}_{Qh_Y}^{-1}(\hat{H}_{Qh_Z}(\hat{H}_{Ph_Z}^{-1}(\hat{F}_{Ph_X}(x))))$$

NEC DESIGN: PROPENSITY SCORES

The propensity score (PS) e(Z) is the conditional probability of being assigned to a particular test form given the covariate vector Z.

$$e(\mathbf{Z}) = \Pr(U = 1 \mid \mathbf{Z})$$

The PS are categorized based on their percentiles.

PRESMOOTHING WITH NEC PS DESIGN:

$$\log(p_{jk}) = \beta_0 + \sum_{i=1}^{T_r} \beta_i^X (x_j)^i + \sum_{i=1}^{T_s} \beta_i^{e(Z)} (e(Z)_l)^i + \beta_{il}^{Xe(Z)} x_j^i e(Z)_l^l$$

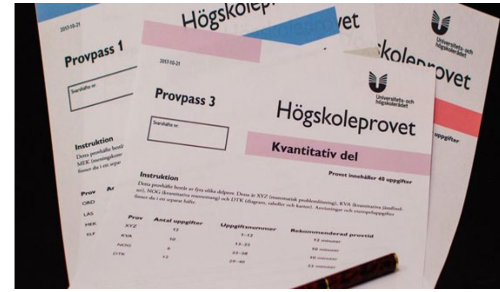
Poststratification (PSE) NEC PS $T = wP + (1-w)Q$

$$\hat{\varphi}_{Y(PSE)}\left(x\right) = \hat{G}_{Th_Y}^{-1}\left(\hat{F}_{Th_X}\left(x\right)\right)$$

Chained Equating (CE) NEC PS $\hat{\varphi}_{Y(CE)}\left(x\right) = \hat{G}_{Qh_Y}^{-1}(\hat{H}_{Qh_{Ye(Z)}}(\hat{H}_{Ph_{Xe(Z)}}^{-1}(\hat{F}_{Ph_X}(x))))$

UMEÅ UNIVERSITY

EMPIRICAL STUDY



- 14,644 test takers: 7,322 test takers from two SweSAT administrations.
- 24-item “anchor” test: 12 items from two different test administrations.
- Covariates

  - Verbal test scores (0–30, 31-40, 41-50, 51-80)

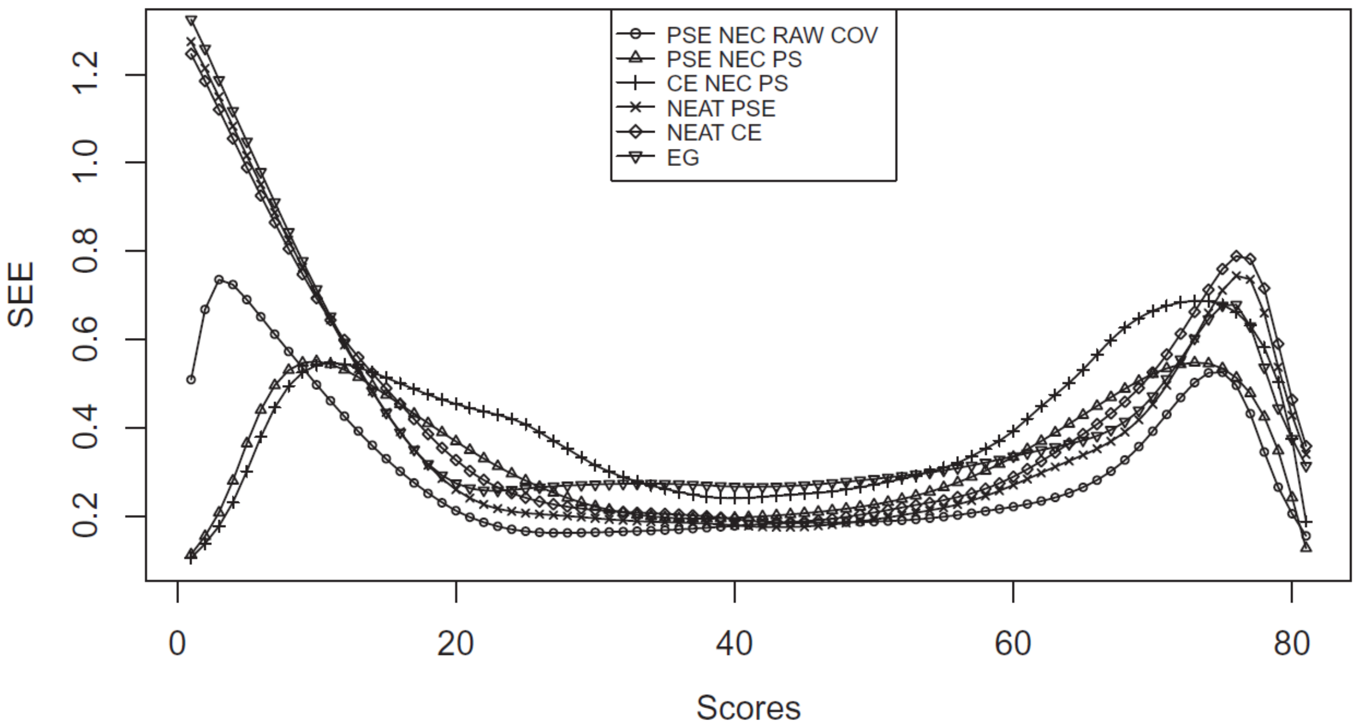
  - Gender (0 = female, 1 = male)

  - Age (0-20,21-24,25-29,30-39, 40-)

  - Propensity scores are divided into 20 categories.

	Verb	Age	Gender	Anchor
Correlation to Y	0.48	−0.14	0.26	0.81
Correlation to X	0.52	−0.13	0.28	0.81
Mean	43.91 (39.35)	1 (1)	0.42 (0.53)	12.17 (10.55)
Standard deviation	12.08 (11.56)	2 (2)	0.49 (0.50)	4.59 (4.64)

EMPIRICAL STUDY: SEE

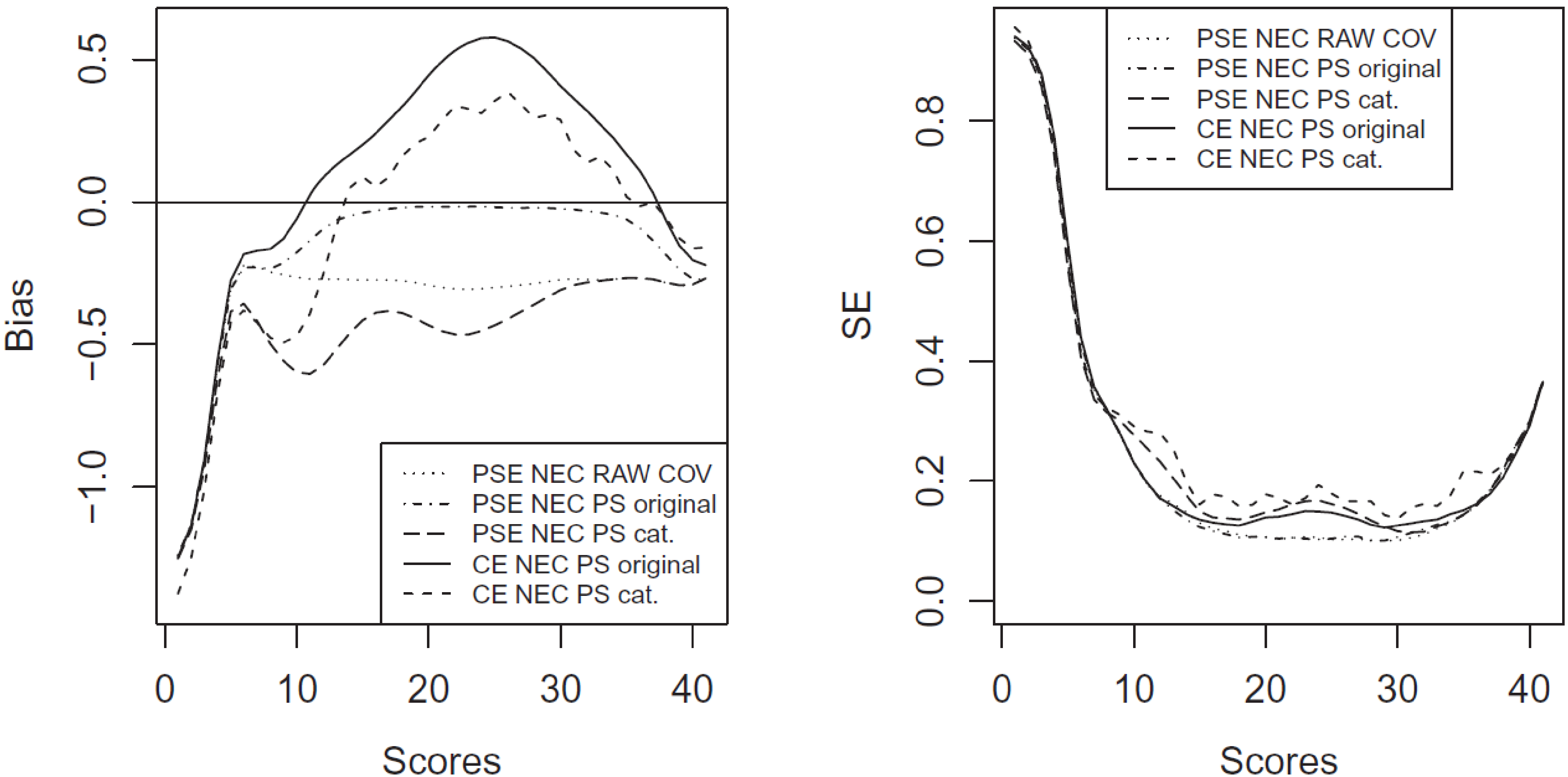


SIMULATION STUDY

- 10,000 test takers
- 1,000 replicates
- Two background variables generated following covariate distributions in SweSAT
- Propensity score as proxy for ability
- 20 propensity score categories
- Absolute standardized mean difference (ASMD) used to examine covariate balance.
- **Evaluation measures:**

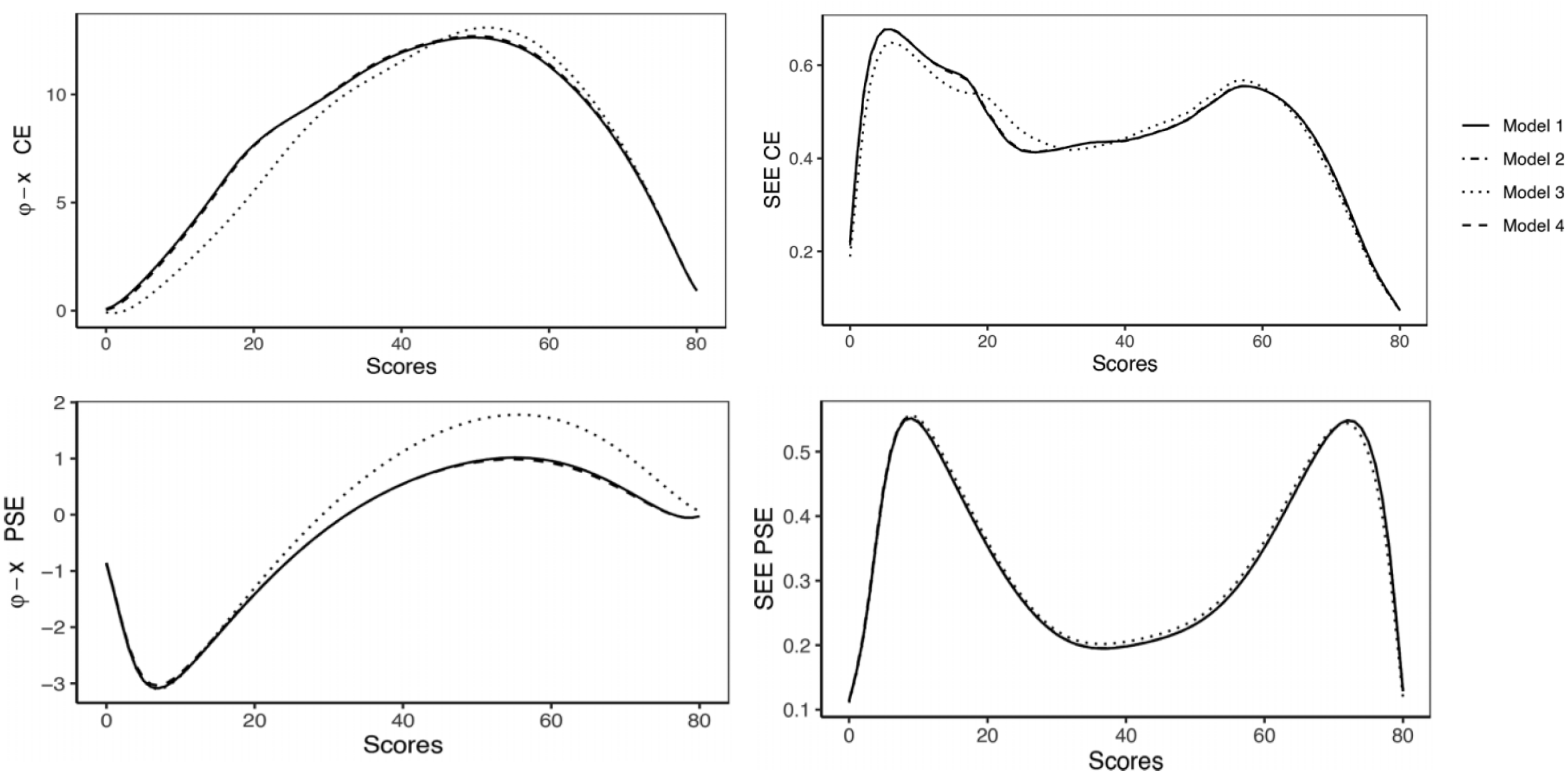
$$\text{Bias}\left(\hat{\varphi}_Y\left(x_i\right)\right)=\sqrt{\frac{1}{R}\sum_{g=1}^{R}\left(\hat{\varphi}_Y^{(g)}\left(x_i\right)-\varphi_Y\left(x_i\right)\right)} \qquad \text{SE}\left(\hat{\varphi}_Y\left(x_i\right)\right)=\sqrt{\frac{1}{R-1}\sum_{g=1}^{R}\left(\varphi_Y^{(g)}\left(x_i\right)-\overline{\varphi}_Y^{(g)}\right)^2}$$

RESULTS SIMULATION STUDY

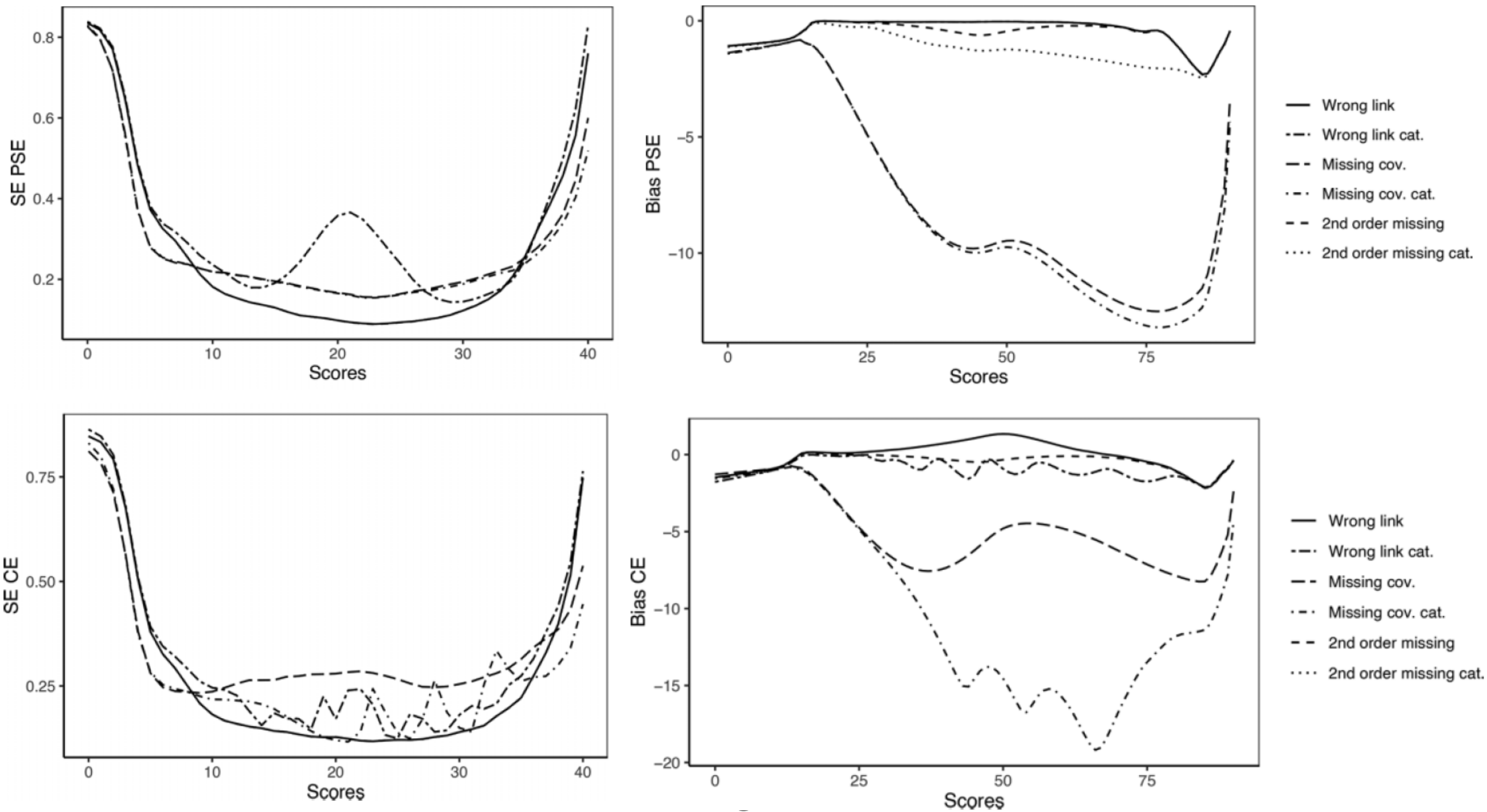


WHAT IF THE MODELS ARE MISSPECIFIED?

- **Same empirical SweSAT data**
  - Models: (1) Full model (2) Wrong link (probit/logit) (3) Missing a covariate (4) Including an interaction term
- **Similar simulation study**
  - Conditions 1) Wrong link 2) Omitting a covariate 3) Omitting a higher-order term.



MISSPECIFIED MODELS (SIMULATION STUDY)



HOW SHOULD WE CONSTRUCT THE ANCHOR TEST TO ADJUST FOR ABILITY DIFFERENCES?

UMEÅ UNIVERSITY

The anchor test is crucial to the accuracy of equating in the NEAT design.

What is a good anchor test?

What happens if the group ability differ a lot? Which groups should get the anchor test?

How does different anchor test form’s characteristics affect the equating transformation?

Approaches

- Empirical study
- Simulation study

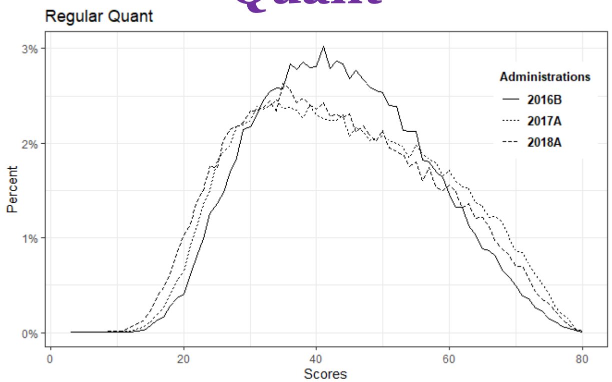
Equating methods

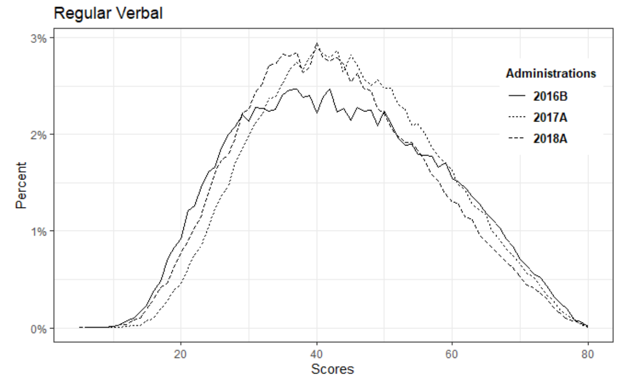
- Circle-arc equating
- Kernel Post-Stratification equating (KPSE)
- Kernel Chain equating (KCE)

Regular + Anchor (2016B, 2017A, 2018A)

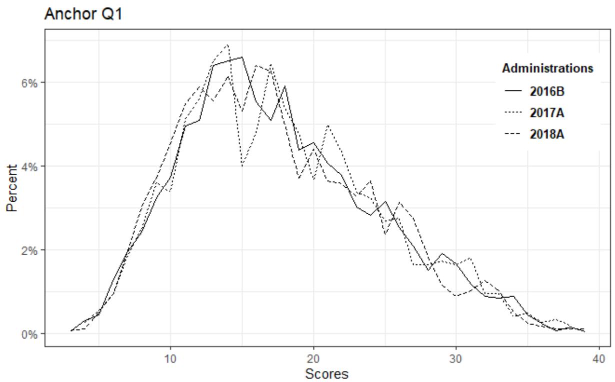
Quant

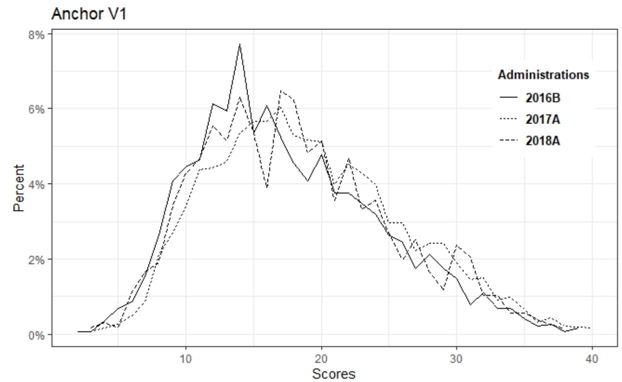
Verbal



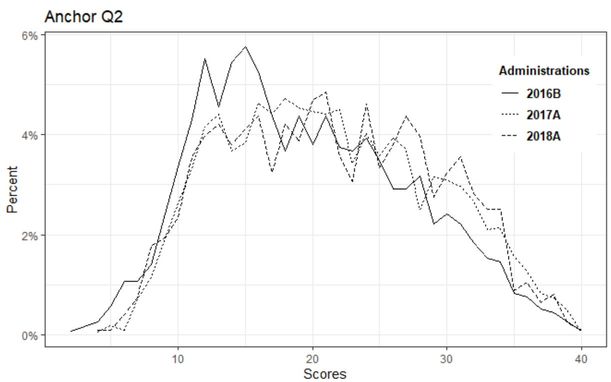


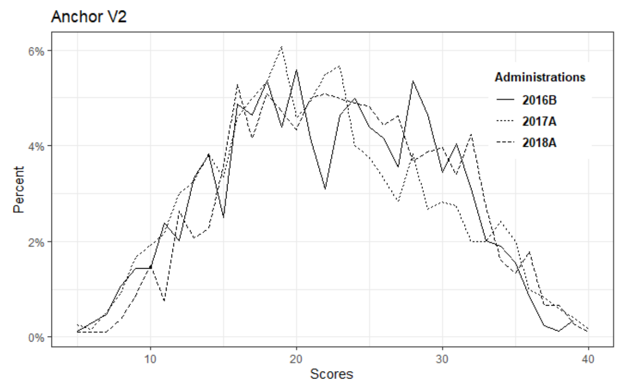
Regular





Anchor Q1, V1

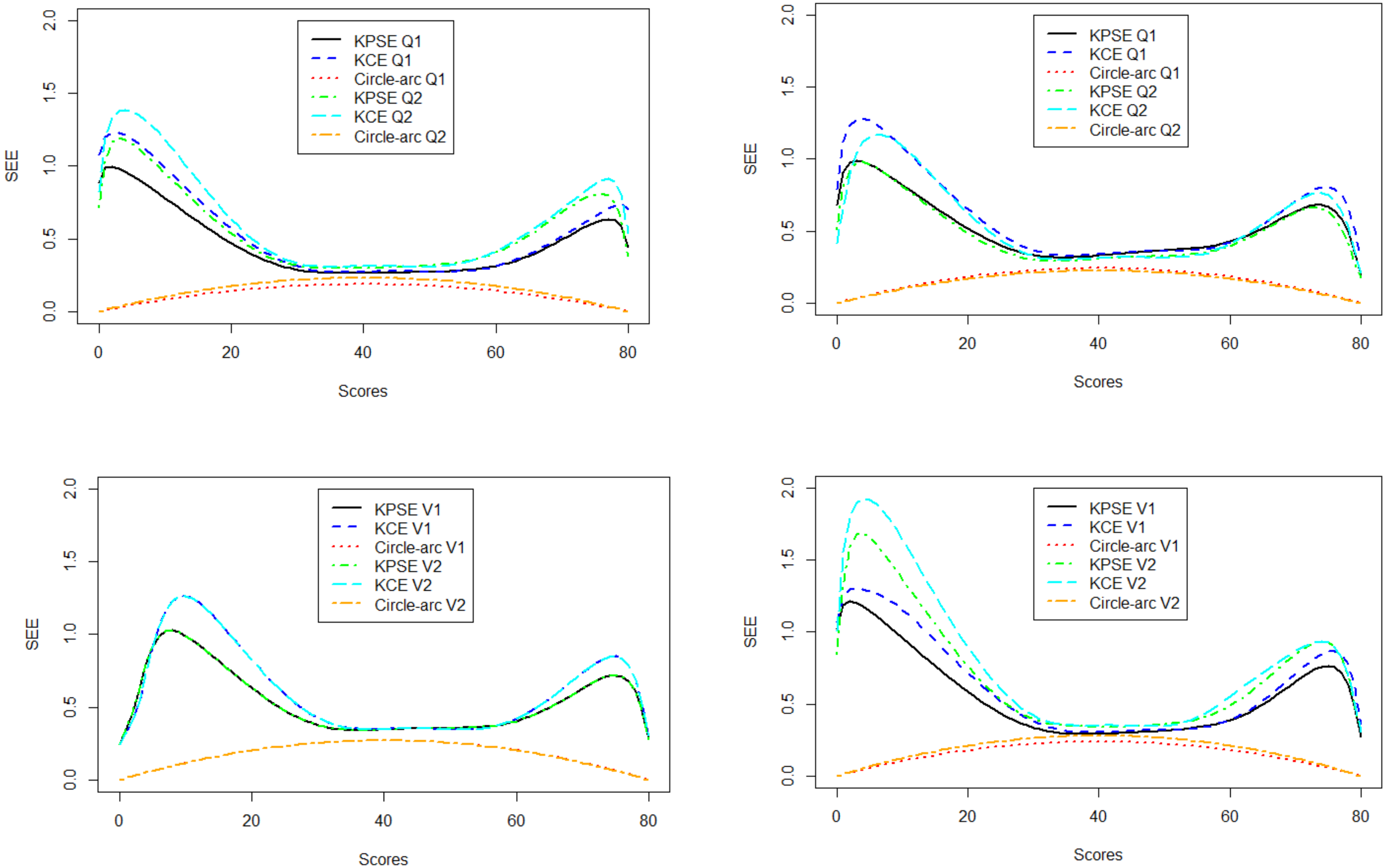




Anchor Q2, V2

UMEÅ UNIVERSITY

2018A ->2016B **2018A ->2017A**



UMEÅ UNIVERSITY

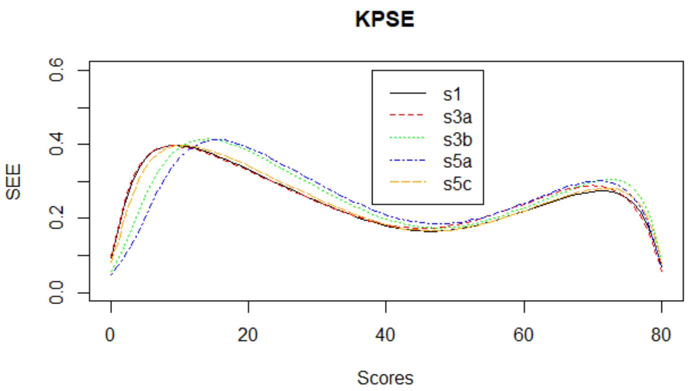
SIMULATION STUDY

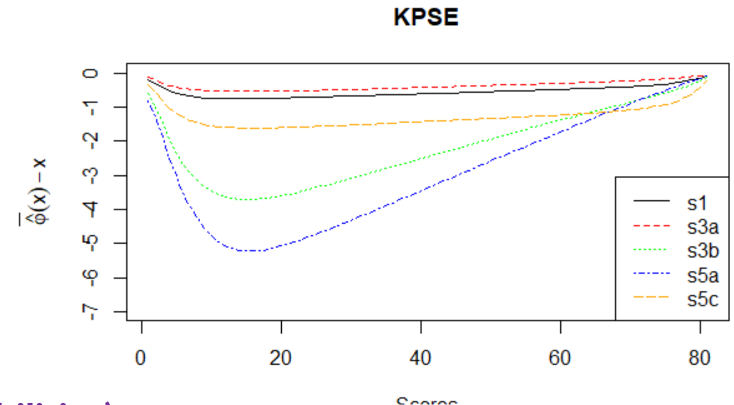
- Regular test with 80 multiple choice items and 40 items anchor test.
- 3PL IRT model
- The baseline case with the following item parameters:
  - discrimination: a~ LogNormal(0.3,0.4),
  - difficulty: b~ N(0.4,1), and
  - guessing: c~ Beta(1.6,6).
- Correlations (Regular test forms - Anchor tests): 0.78 - 0.82 (like real data).

In total, we examined 23 conditions by varying:

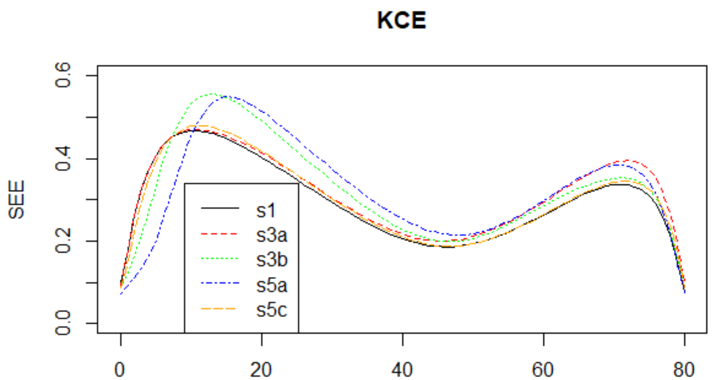
- item difficulty
- item discrimination
- the abilities of the different groups
- difficulties of both anchor and regular test forms

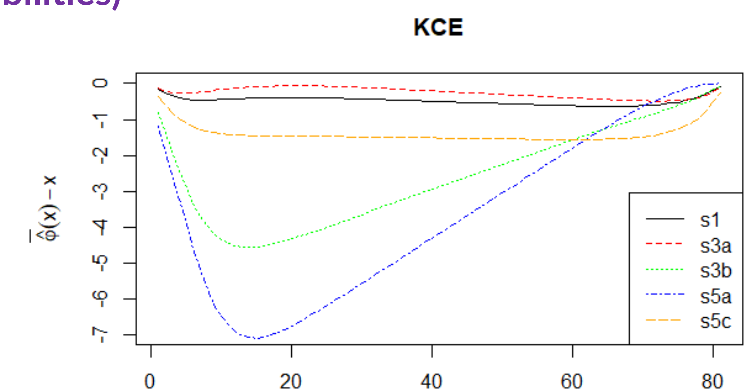
- SEE and Bias
- 500 replications.





Difficulty
(groups have similar abilities)





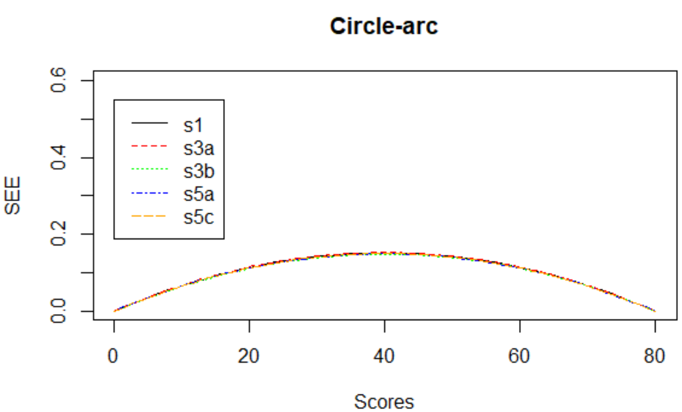
s1 - baseline case

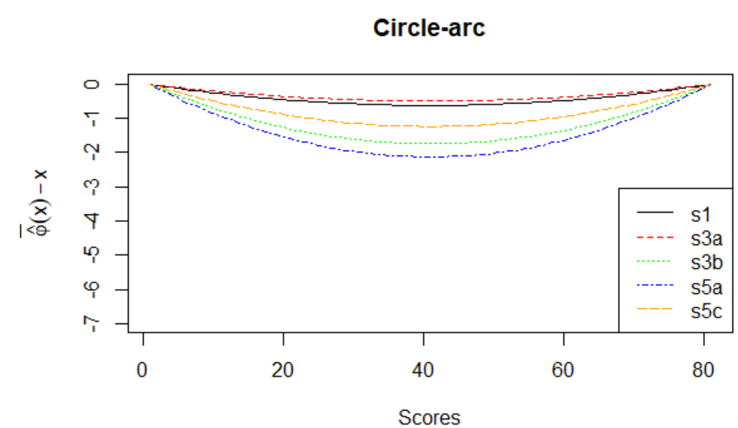
s3a – more difficult anchor than regular

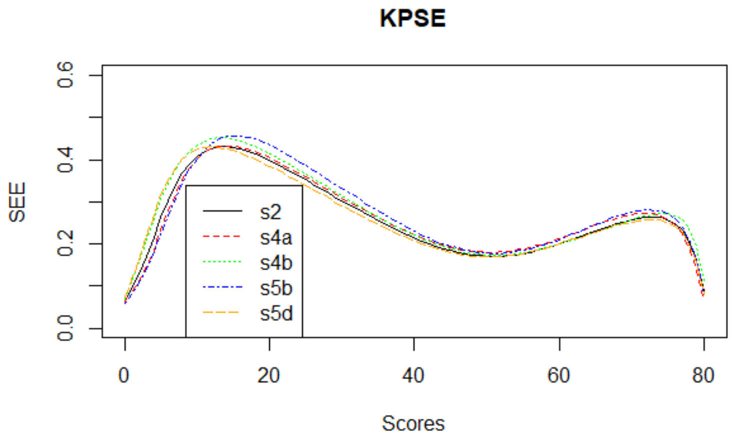
s3b – easier anchor

s5a – more spread difficulties in anchor

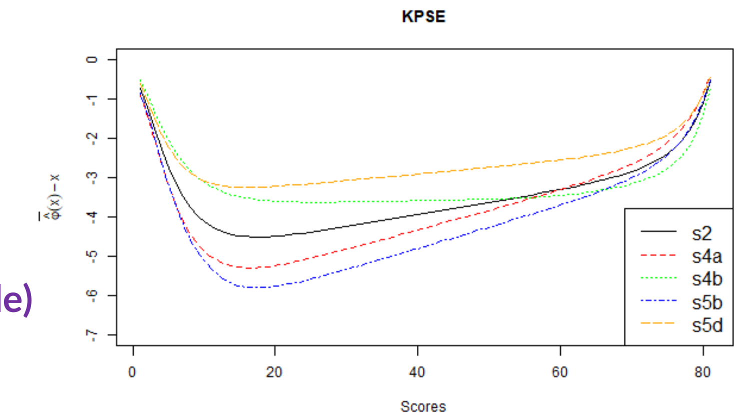
s5c – less spread difficulties in anchor







Difficulty
(one group more able)



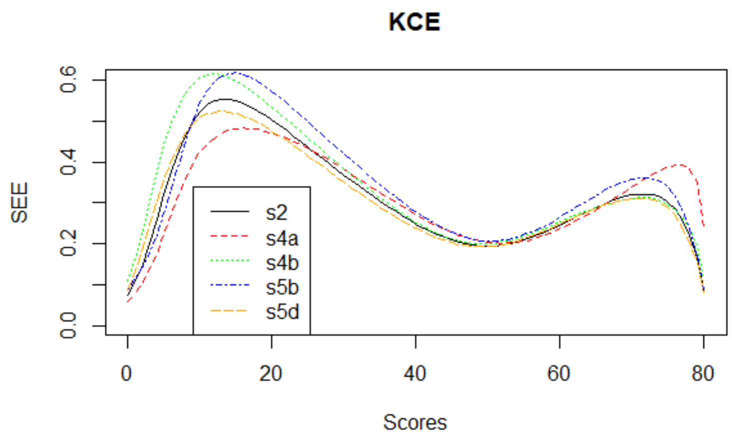
s2 - baseline case,

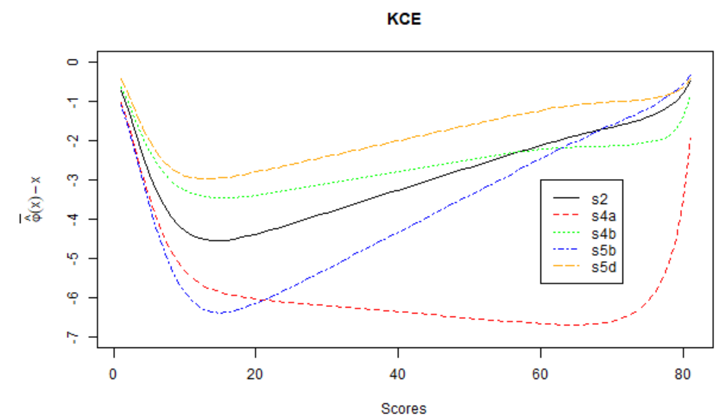
s4a – **more difficult** anchor than regular

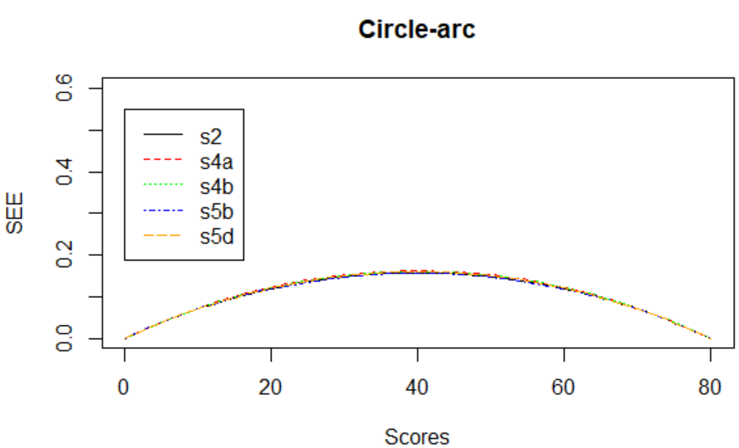
s4b – **easier anchor**

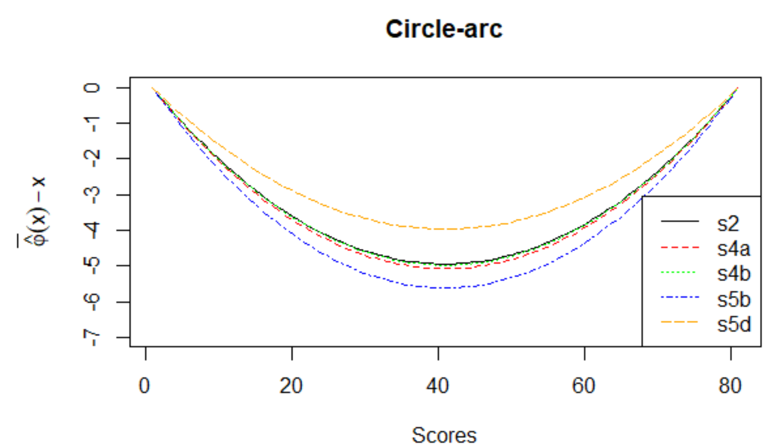
s5b – **more spread difficulties** in anchor

s5d – **less spread difficulties** in anchor

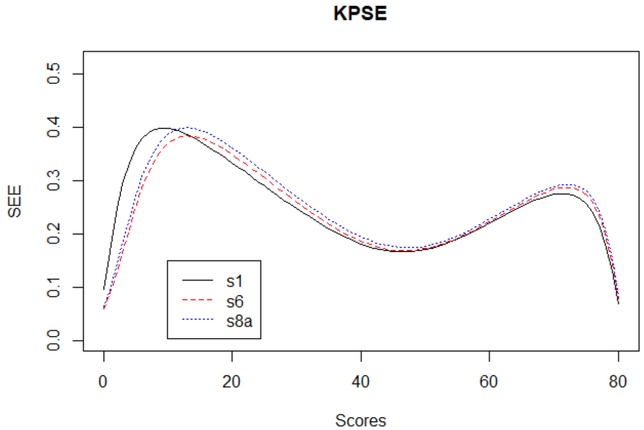


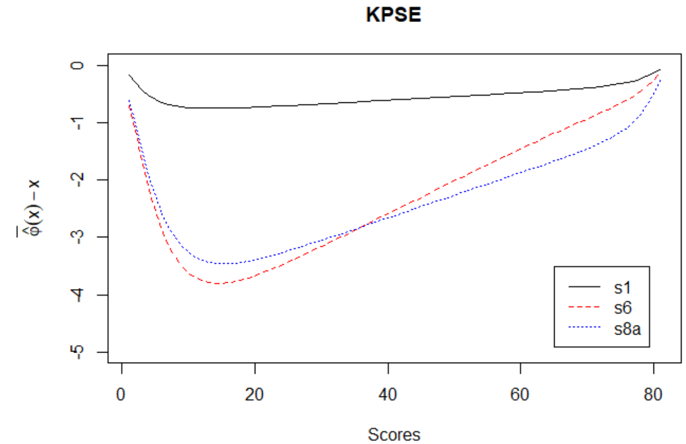




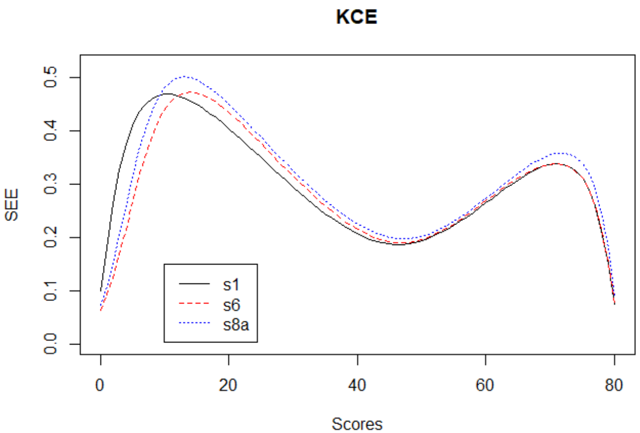


UMEÅ UNIVERSITY





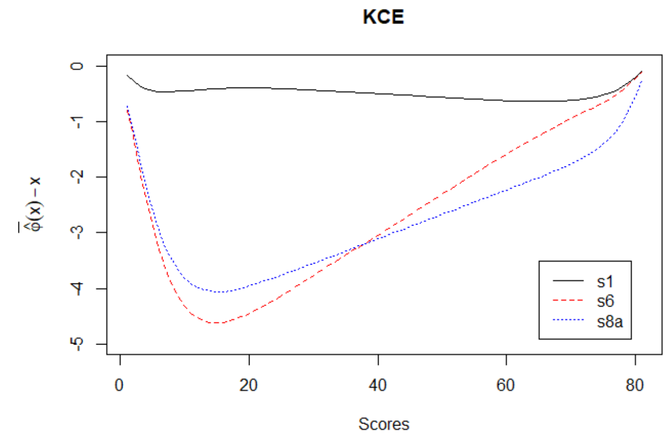
Discrimination
(groups have similar abilities)

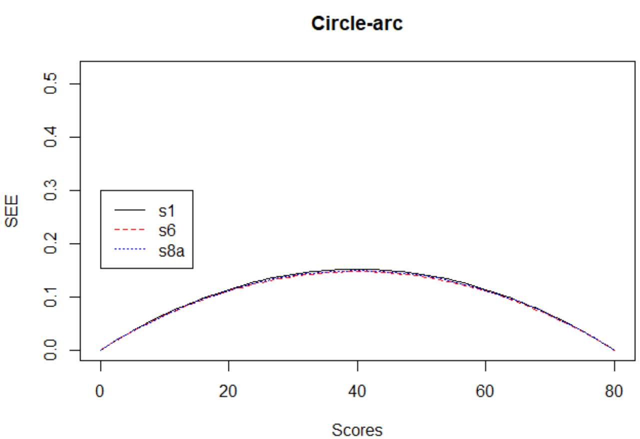


s1 - baseline case

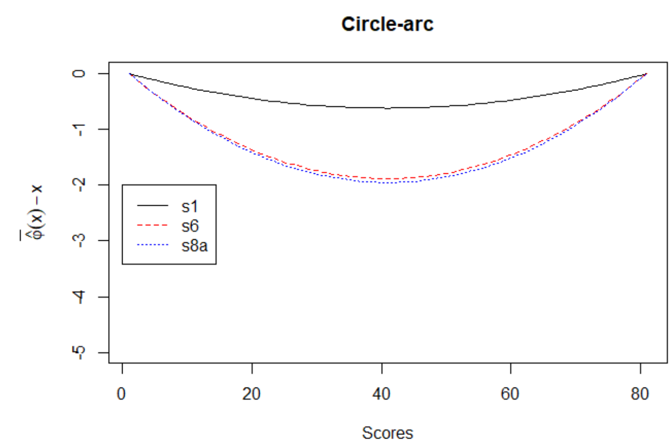
s6 – **more discriminating** anchor than regular

s8a – **less discriminating** anchor





UMEÅ UNIVERSITY



Abilities

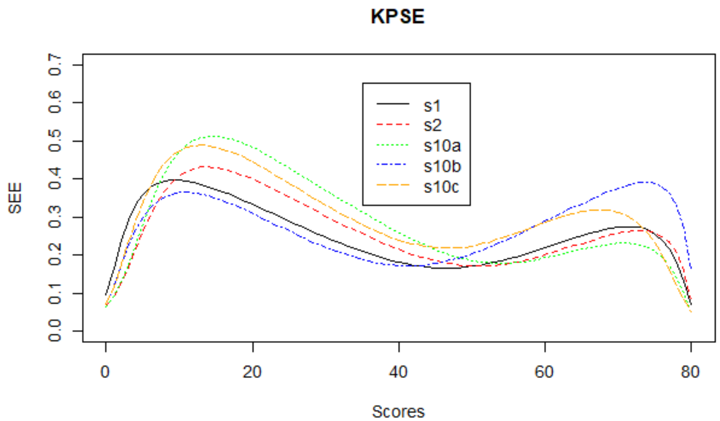
s1 - baseline case: groups are **similar**

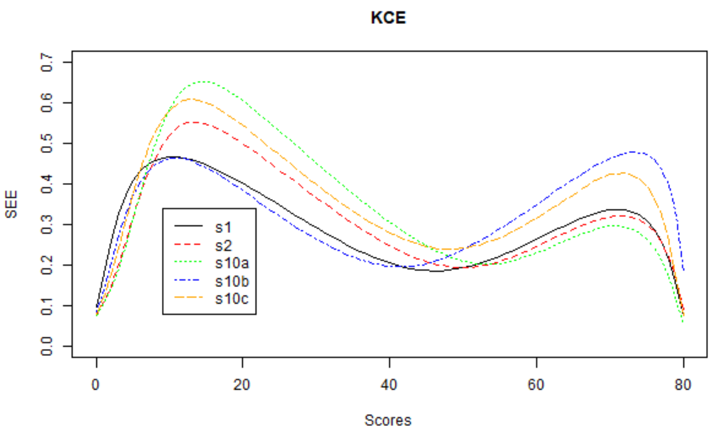
s2 – baseline case when one group is **more able**

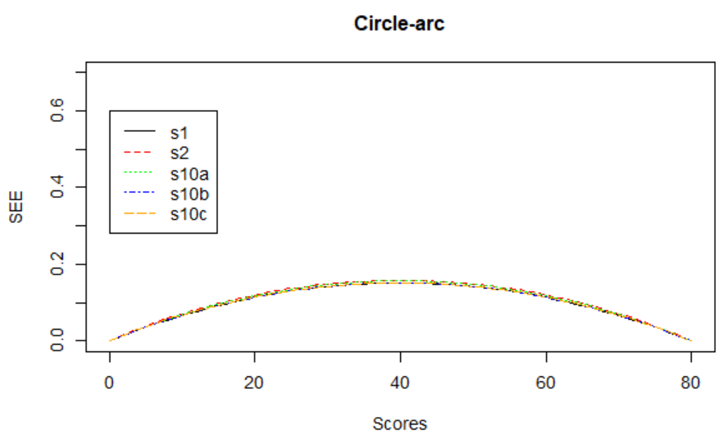
s10a – both groups have **high abilities**

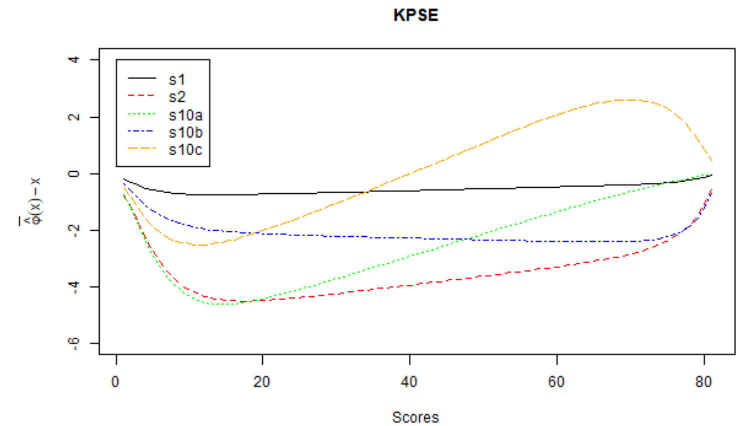
s10b – both groups have **low abilities**

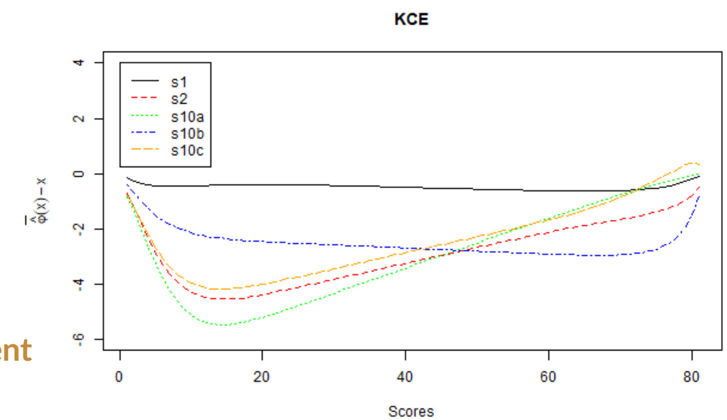
s10c –Groups are **different in ability.** One has low abilities and the other has high

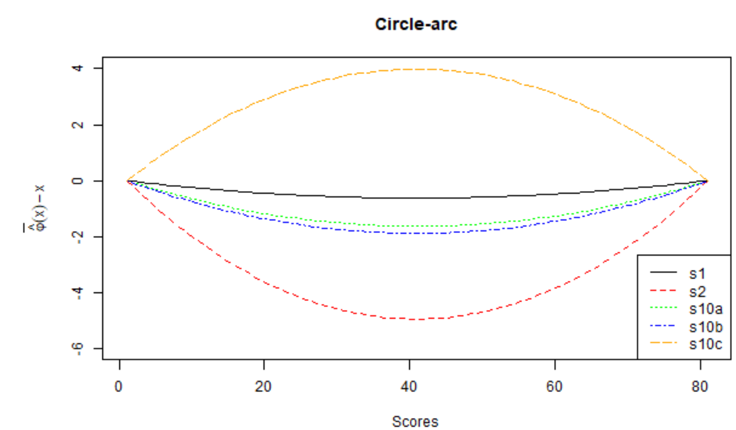






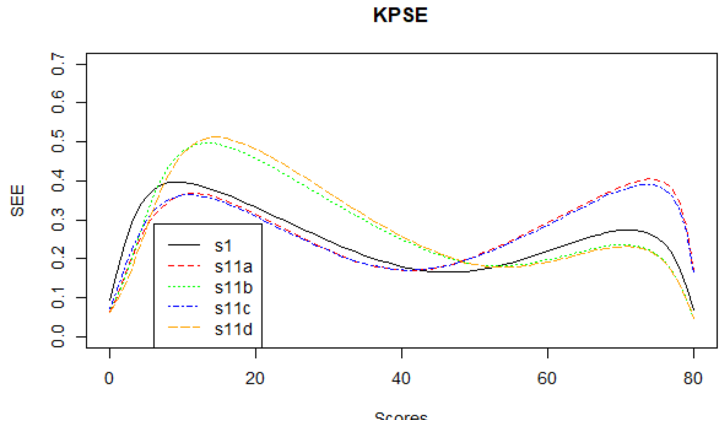


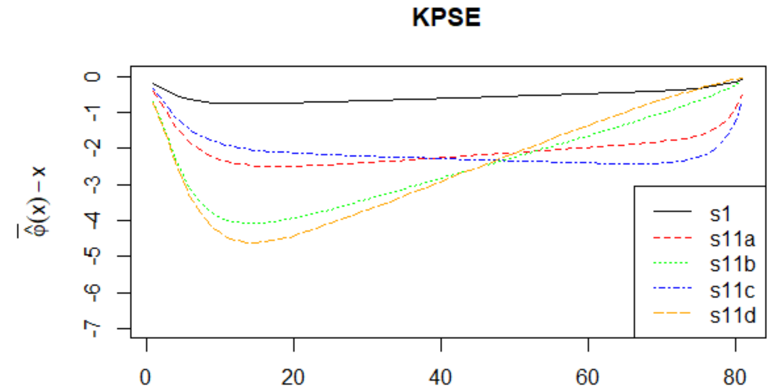




UMEÅ UNIVERSITY

REG difficulty





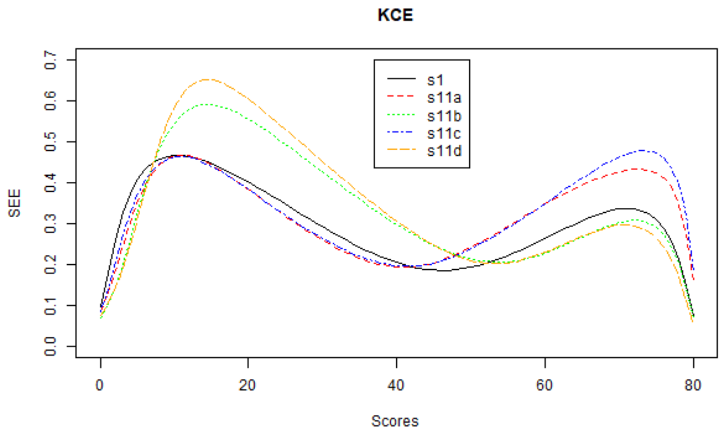
s1 - baseline case when groups are **similar**

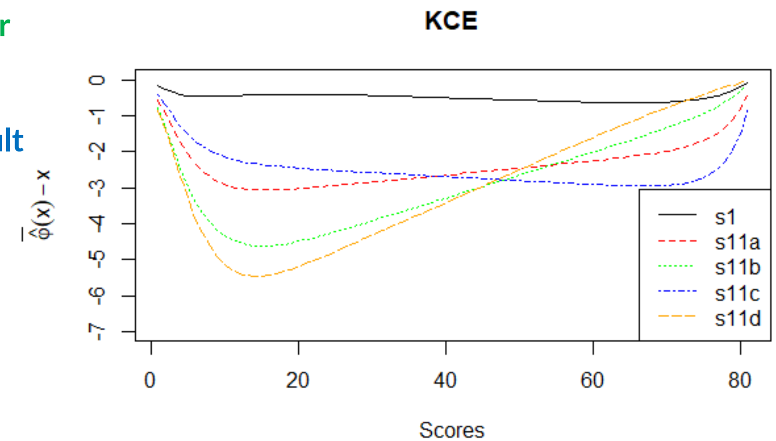
s11a – regular **is more difficult** than anchor

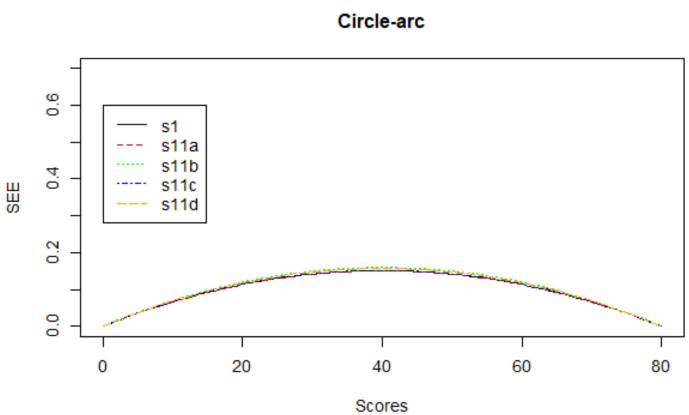
s11b – regular is **easier** than anchor

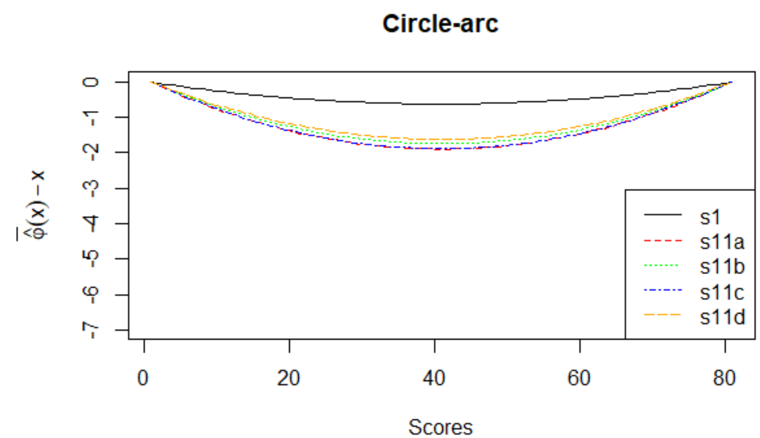
s11c – both are **difficult**

s11d – both are **easy**









UMEÅ UNIVERSITY

CONCLUSIONS

- We **must** adjust when the groups are nonequivalent.
- One possibility is to use the **NEC design with propensity scores**.
  - Careful in the selection of covariates.
  - Most important to include all covariates
- **Anchor test forms**
  - Which ability level the groups that receive the anchor test forms have impact equating results significantly, especially when one group is less able and the other is more abl .
  - The lowest SEE are achieved when the anchor test form and the regular test forms are of average difficulty.
  - If possible, give anchor test form to the average ability groups.
  - Easy anchor test forms and/or regular test forms, and anchor test forms with more spread difficulties affect equating negatively.

FUTURE RESEARCH

- Which covariates are useful for equating purposes?
- What is the best anchor test and who should it be given to?
- How should we handle unexpected problems in anchor tests (e.g. differential item functioning, parameter drift)

Some references

González, J. & Wiberg, M. (2017). *Applying test equating methods – using R*. Cham, Switzerland: Springer.

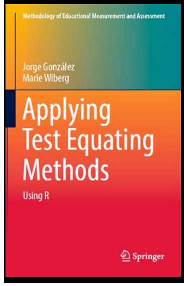
Laukaityte, I. & Wiberg, M. (2023). The impact of differences in group abilities and anchor test features on test score equating. Manuscript.

Wallin, G. & Wiberg, M. (2019). Propensity scores in kernel equating for non-equivalent groups. *Journal of Educational and Behavioral Statistics. 44*(4), 390-414.

Wallin, G. & Wiberg, M. (2023). Model misspecification and robustness of test score equating using propensity scores. *Journal of Educational and Behavioral Statistics,*

Wiberg, M. & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39*(5), 349-361.

Wiberg, M., González, J. & von Davier, A. A. (2024). *Generalized kernel equating*. Forthcoming book.



UMEÅ UNIVERSITY

Thank you for your attention!

marie.wiberg@umu.se

UMEÅ UNIVERSITY