A two-step procedure for scaling multilevel data using Mokken's scalability coefficients

Letty Koopman, Bonne J. H. Zijlstra, L. Andries van der Ark

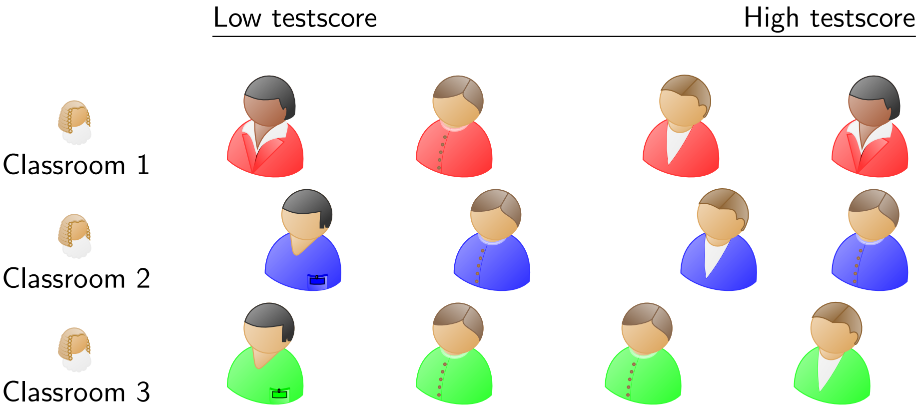
University of Amsterdam

V.E.C.Koopman@UvA.nl

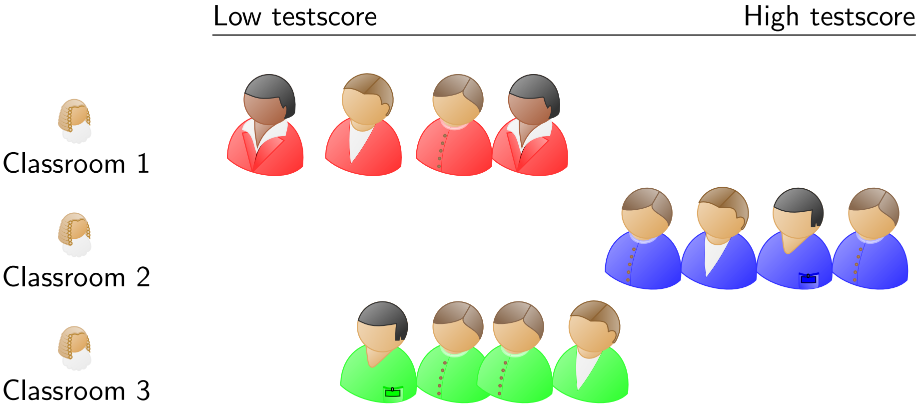
February 27, 2020

Multilevel Test Data

Item scores of respondents nested in groups:



Grouping Effect!



Multilevel Test Data

Scale *well-being with teachers*:

1. The teachers usually know how I feel
2. I can talk about problems with the teachers
3. If I feel unhappy, I can talk to the teachers about it
4. I feel at ease with the teachers
5. The teachers understand me
6. I have good contact with the teachers
7. I would prefer to have other teachers *(reversely scored)*

Scored 1 (*not true at all*) to 5 (*completely true*)

Multilevel Test Data

Scale *well-being with teachers*:

1. The teachers usually know how I feel
2. I can talk about problems with the teachers
3. If I feel unhappy, I can talk to the teachers about it
4. I feel at ease with the teachers
5. The teachers understand me
6. I have good contact with the teachers
7. I would prefer to have other teachers *(reversely scored)*

Scored 1 (*not true at all*) to 5 (*completely true*)

Original data: 16,297 students in 814 classes in 94 schools (COOL5-18)
Our subset: 651 students in 30 classes/schools

Scalability of Questionnaires

Scalability: Can we accurately order respondents on the latent concept *well-being with teachers*, using the test score?

Goal: Investigate the scalability of the items in multilevel test data using Mokken's scalability coefficients.

Mokken's Scalability Coefficients

Scalability coefficients for item-pairs (H_{ij}), items (H_i), and total scale (H)

- No relation between items i, j: $H_{ij} = 0$
- Perfect relation between items i, j: $H_{ij} = 1$

Mokken's Scalability Coefficients

Scalability coefficients for item-pairs (H_{ij}), items (H_i), and total scale (H)

- No relation between items i, j: $H_{ij} = 0$
- Perfect relation between items i, j: $H_{ij} = 1$

What is a Mokken scale?

- All $H_{ij} > 0$
- All $H_i \geq c$ (e.g., $c = 0.3$)

Mokken's Scalability Coefficients

Scalability coefficients for item-pairs (H_{ij}), items (H_i), and total scale (H)

- No relation between items i, j: $H_{ij} = 0$
- Perfect relation between items i, j: $H_{ij} = 1$

What is a Mokken scale?

- All $H_{ij} > 0$
- All $H_i \geq c$ (e.g., $c = 0.3$)

Strength of a Mokken scale

- $H \geq .3$ Weak scale
- $H \geq .4$ Medium scale
- $H \geq .5$ Strong scale

Stronger scale = more accurate ordering

Problem: Only traditional estimation methods available for H and SE

- Assumes simple random sample from the population
- Underestimated standard errors
- Confidence intervals too narrow

Possible concequences:

- Incorrectly admitting items to the final scale
- Overestimating the quality of the scale

Estimating Scalability Coefficients in Multilevel Test Data

Needed: Two-level estimation methods for H and SE

- Inspired by multi-rater data (scaling of groups)
- Within-rater scalability coefficient H^W similar to Mokken's H
- Negligible bias and good coverage of estimators
- Problem with unequal group sizes: $\widehat{SE}$ too large
    - Estimation used averaged proportions across groups
    - Adjustment: Use proportions weighted for group size

Estimating Scalability Coefficients in Multilevel Test Data

Needed: Two-level estimation methods for H and SE

- Inspired by multi-rater data (scaling of groups)
- Within-rater scalability coefficient H^W similar to Mokken's H
- Negligible bias and good coverage of estimators
- Problem with unequal group sizes: $\widehat{SE}$ too large
  - Estimation used averaged proportions across groups
  - Adjustment: Use proportions weighted for group size

Solution: Use adjusted version of within-rater estimation method.

Leads to: identical $\widehat{H}$ as one-level method, but different $\widehat{SE}$ (and CI).

Performance of the Methods

Simulation design: ICC, number of groups, group size

Performance of the Methods

Simulation design: ICC, number of groups, group size

Point estimate H: Unbiased in all conditions

Performance of the Methods

Simulation design: ICC, number of groups, group size

Point estimate H: Unbiased in all conditions

Standard errors:

- One-level bias $-.013 \approx -35\%$
  Worse for larger groups and larger ICCs
- Two-level bias $.003 \approx 7\%$
  Unequal group size no longer affected two-level bias
  Conservative for small ICC and very small groups
  Slightly underestimated for only 10 groups and large ICC

Performance of the Methods

Simulation design: ICC, number of groups, group size

Point estimate H: Unbiased in all conditions

Standard errors:

- One-level bias $-.013 \approx -35\%$
  Worse for larger groups and larger ICCs
- Two-level bias $.003 \approx 7\%$
  Unequal group size no longer affected two-level bias
  Conservative for small ICC and very small groups
  Slightly underestimated for only 10 groups and large ICC

Coverage: Similar patterns as standard errors

- One-level coverage .744
- Two-level coverage .949

Two-Step Scaling Procedure for Multilevel Data

Step 1: Scalability investigation using two-level confidence intervals

- Automated item selection procedure (AISP)
- Investigate dimensionality item set: Create one or more Mokken scales
- Starts with highest $\widehat{H}_{ij}$ and subsequently adds items
- Compares $CI(H_{ij})$ to zero and $CI(H_i)$ to c
- Use $c = 0,\ 0.05,\ 0.1,\ \ldots,\ 0.55$
- Look for relevant outcome patterns to decide on final scale

Two-Step Scaling Procedure for Multilevel Data

Step 1: Scalability investigation using two-level confidence intervals

- Automated item selection procedure (AISP)
- Investigate dimensionality item set: Create one or more Mokken scales
- Starts with highest $\widehat{H}_{ij}$ and subsequently adds items
- Compares $CI(H_{ij})$ to zero and $CI(H_i)$ to c
- Use $c = 0,\ 0.05,\ 0.1,\ \ldots,\ 0.55$
- Look for relevant outcome patterns to decide on final scale

Step 2: Estimate and test the intraclass correlation

- Use the test score on the final scale
- Perform an F-test: Null hypothesis $ICC = 0$
- If F-test is not significant: Use one-level standard errors for final scale

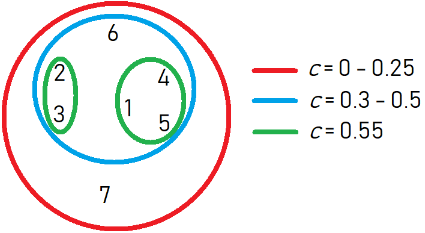
Perform Procedure in R Using mokken

Necessary functions will be implemented in R-package `mokken` very soon! (Version 3.0, if you already want to perform the analysis or get an update when it is updated, let me know!)

- `aisp(X, c = seq(0, .55, .05), two.level = TRUE, CI = TRUE)`: Performs AISP using a range of thresholds c and two-level confidence intervals
- `MLcoefH(X, se = TRUE, weigh.props = TRUE)`: Two-level method for point estimates and standard errors (use only within-rater coefficients)
- `ICC(X)`: Gives ICC estimates per item and for the total scale, with an F-test for the total scale ICC
- `coefH(X[, -1], se = TRUE)`: One-level method for point estimates and standard errors

Results for Scale Well-Being With Teachers

Step 1:



Conclusion: Use only first six items in final scale (For 7 items:)

Step 2:

- $ICC = .168$, $F(26, 621) = 5.08$, $p < .001$ ($ICC = .170$)

Conclusion: Retain two-level estimates

Resulting scale:

- All $H_i > .5$ (all $H_i > .25$)
- Strong scale $.563 \leq H \leq .663$ (Medium scale $.493 \leq H \leq .605$)

Discussion

- Don't use one-level standard errors for Mokken's coefficients in multilevel data!
- Use new (more accurate) two-level standard errors (but why overestimated for small ICC?)
- Perform a two-step procedure for scalability analysis in multilevel data
  1. Scalability analysis using two-level confidence intervals
  2. Investigate within-group dependency
- Scale investigation finished? No, not quite yet: Generalize methods to check nonparametric IRT model assumptions in multilevel data

University of Amsterdam

Thank you!

Letty Koopman
V.E.C.Koopman@UvA.nl

References

- Koopman, L., Zijlstra, B. J. H., De Rooij, M. & Van der Ark, L. A., (2019). Bias of Two-Level Scalability Coefficients and Their Standard Errors. *Applied Psychological Measurement*. Advance online publication. doi: 10.1177/0146621619843821
- Koopman, L., Zijlstra, B. J. H. & Van der Ark, L. A., (2019). Standard errors of two-level scalability coefficients. *British Journal of Statistical and Mathematical Psychology*. Advance online publication. doi: 10.1111/bmsp.12174
- Koopman, L. Zijlstra, B. J. H, & Van der Ark, L. A. (2020). *A two-step procedure for scaling multilevel data using Mokken's scalability coefficients*. Manuscript in preparation.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands: Mouton. https://doi.org/10.1515/9783110813203
- Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, 70, 137158. http://doi.org/10.1111/bmsp.12078
- Snijders, T. A. B. (2001). Two-level non-parametric scaling for dichotomous data. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 319-338). New York, NY: Springer. doi:10.1007/978-1-4613-0169-1_17
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Intraclass Correlation

σ^2 = population within-group variance
τ^2 = population between-group variance

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2}$$

Snijders & Bosker (2012) p. 18

Standard Errors - Two-Level Method

Let $\mathbf{g}(\mathbf{n})$ be the transformation of the frequencies of item-score patterns $\mathbf{n}$, resulting in a vector with all the scalability coefficients $\mathbf{H}$.

Assumption:

- Probabilities $\mathbf{p}$ of item-score patterns $\mathbf{n}$ differ per group
- Results in multinomial distribution per subject
- $\text{V}(\mathbf{n})^*$ is the multinomial covariance matrix of vector $\mathbf{n}$
- $\text{V}(\mathbf{n}) = \text{V}(\mathbf{n})^* + sr(r-1)\text{V}(\mathbf{p})$

$\mathbf{G}$ = The Jacobian of $\mathbf{g}(\mathbf{n})$ (i.e., matrix of first-order partial derivatives)

Delta method: $\text{V}(\mathbf{H}) \approx \mathbf{G}\ \text{V}(\mathbf{n})\ \mathbf{G}^T$

Koopman, Zijlstra, & Van der Ark (2019)

Simulation Study

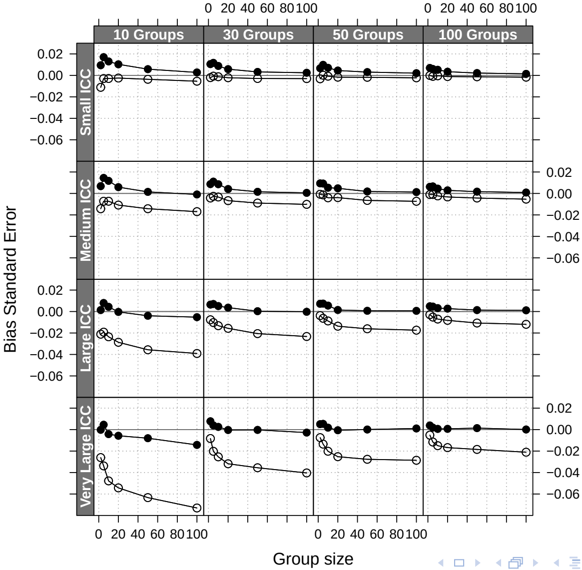
Outcome measures:

- Bias of point estimate
- Bias of one- and two-level standard errors
- Coverage of one- and two-level 95% confidence interval

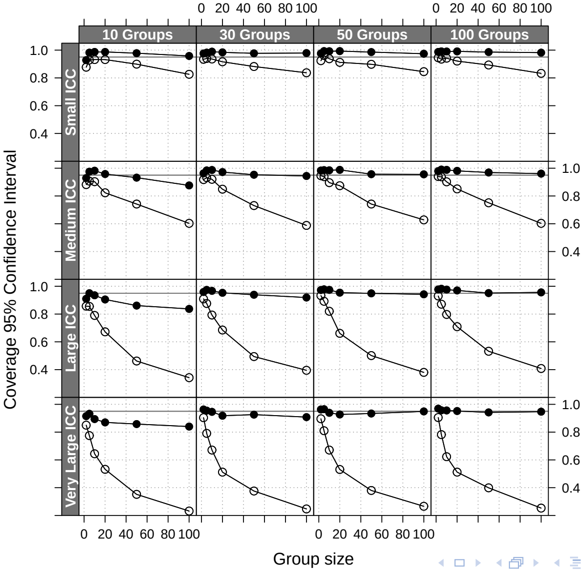
Simulation Design:

- Within-group dependency (small, medium, large, very large)
- Number of groups (10, 30, 50, 100)
- Group size
  - Equal group sizes (2, 5, 10, 20, 50, 100)
  - Unequal group sizes ([10; 30], related or unrelated to latent trait)

Bias of the Standard Error



Coverage of the 95% Confidence Interval



Simulation Study - Unequal Groups

Condition	Bias $\widehat{H}$	Bias $\widehat{SE}$		Coverage CI	
		One-level	Two-level	One-level	Two-level
Equal, 20 respondents	$-.006$	$-.014$	.003	.743	.956
Unequal, independent	$-.006$	$-.015$	.002	.735	.930
Unequal, dependent	$-.013$	$-.013$	.001	.716	.925

Scalability Coefficients SWMD

	Seven Items			Six Items			
Item	$\widehat{H}$	$\widehat{SE}$	95% CI	$\widehat{H}$	$\widehat{SE}$	95% CI	ICC
1	.570	.033	[.506; .634]	.606	.032	[.544; .669]	.126
2	.592	.029	[.535; .649]	.635	.027	[.581; .689]	.111
3	.563	.030	[.504; .622]	.611	.029	[.555; .668]	.103
4	.606	.033	[.540; .671]	.632	.033	[.567; .696]	.142
5	.590	.030	[.532; .648]	.632	.030	[.574; .690]	.074
6	.537	.030	[.478; .596]	.561	.030	[.502; .620]	.120
7	.387	.053	[.284; .490]				
Total	.549	.029	[.493; .605]	.613	.025	[.563; .663]	.168