On cows and test construction

NIELS SMITS & KEES JAN KAN

Research Institute of Child Development and Education
University of Amsterdam, The Netherlands

SEM WORKING GROUP
AMSTERDAM 16/03/2018

Let's look at some cows

Let's look at some cows

Let's look at some cows

Let's look at some cows

- A dairy cow, a beef cow and a hybrid cow.
- Variability at two main characteristics: udder and muscles.
- Cow types bred (selection) on these two characteristics.
- Cow can excel on only one characteristic.
- Hence: trade-off between them.

Let's look at some cows

- A dairy cow, a beef cow and a hybrid cow.
- Variability at two main characteristics: udder and muscles.
- Cow types bred (selection) on these two characteristics.
- Cow can excel on only one characteristic.
- Hence: trade-off between them.

Let's look at some cows

- A dairy cow, a beef cow and a hybrid cow.
- Variability at two main characteristics: udder and muscles.
- Cow types bred (selection) on these two characteristics.
- Cow can excel on only one characteristic.
- Hence: trade-off between them.

Let's look at some cows

- A dairy cow, a beef cow and a hybrid cow.
- Variability at two main characteristics: udder and muscles.
- Cow types bred (selection) on these two characteristics.
- Cow can excel on only one characteristic.
- Hence: trade-off between them.

Let's look at some cows

- A dairy cow, a beef cow and a hybrid cow.
- Variability at two main characteristics: udder and muscles.
- Cow types bred (selection) on these two characteristics.
- Cow can excel on only one characteristic.
- Hence: trade-off between them.

Let's look at some cows

- A dairy cow, a beef cow and a hybrid cow.
- Variability at two main characteristics: udder and muscles.
- Cow types bred (selection) on these two characteristics.
- Cow can excel on only one characteristic.
- Hence: trade-off between them.

Why Cows?

- Breeding is based on selection.
- Offspring loses characteristics not used for selection.
- Analogy: breeding cows = test construction.
- Frame of reference for next slides.

Why Cows?

- Breeding is based on selection.
- Offspring loses characteristics not used for selection.
- Analogy: breeding cows = test construction.
- Frame of reference for next slides.

Why Cows?

- Breeding is based on selection.
- Offspring loses characteristics not used for selection.
- Analogy: breeding cows = test construction.
- Frame of reference for next slides.

Why Cows?

- Breeding is based on selection.
- Offspring loses characteristics not used for selection.
- Analogy: breeding cows $=$ test construction.
- Frame of reference for next slides.

Why Cows?

- Breeding is based on selection.
- Offspring loses characteristics not used for selection.
- Analogy: breeding cows $=$ test construction.
- Frame of reference for next slides.

Two test goals

In practice test results are used mostly for two reasons:

- **Measurement**: assign numerical values that accurately represent test takers' attribute.
  - *For example*: depression severity at intake.
  - Reliability is key.
- **Prediction**: give accurate forecasts of an external criterion.
  - *For example*: a high risk of a major depression diagnosis at clinical interview.
  - Predictive validity is key.

Two test goals

In practice test results are used mostly for two reasons:

- **Measurement**: assign numerical values that accurately represent test takers' attribute.
  - *For example*: depression severity at intake.
  - Reliability is key.
- **Prediction**: give accurate forecasts of an external criterion.
  - *For example*: a high risk of a major depression diagnosis at clinical interview.
  - Predictive validity is key.

Two test goals

In practice test results are used mostly for two reasons:

- **Measurement**: assign numerical values that accurately represent test takers' attribute.
  - *For example*: depression severity at intake.
  - Reliability is key.
- **Prediction**: give accurate forecasts of an external criterion.
  - *For example*: a high risk of a major depression diagnosis at clinical interview.
  - Predictive validity is key.

Two test goals

In practice test results are used mostly for two reasons:

- **Measurement**: assign numerical values that accurately represent test takers' attribute.
  - *For example*: depression severity at intake.
  - Reliability is key.
- **Prediction**: give accurate forecasts of an external criterion.
  - *For example*: a high risk of a major depression diagnosis at clinical interview.
  - Predictive validity is key.

Two test goals

In practice test results are used mostly for two reasons:

- **Measurement**: assign numerical values that accurately represent test takers' attribute.
  - *For example*: depression severity at intake.
  - Reliability is key.
- **Prediction**: give accurate forecasts of an external criterion.
  - *For example*: a high risk of a major depression diagnosis at clinical interview.
  - Predictive validity is key.

Two test goals

In practice test results are used mostly for two reasons:

- **Measurement**: assign numerical values that accurately represent test takers' attribute.
  - *For example*: depression severity at intake.
  - Reliability is key.
- **Prediction**: give accurate forecasts of an external criterion.
  - *For example*: a high risk of a major depression diagnosis at clinical interview.
  - Predictive validity is key.

Two test goals

In practice test results are used mostly for two reasons:

- **Measurement**: assign numerical values that accurately represent test takers' attribute.
  - *For example*: depression severity at intake.
  - Reliability is key.
- **Prediction**: give accurate forecasts of an external criterion.
  - *For example*: a high risk of a major depression diagnosis at clinical interview.
  - Predictive validity is key.

Consider this task

Construct two ‘5 item scales’ from pool of 10 items

1. Highest measurement precision.
2. Highest predictive accuracy.

Correlation matrix

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Rest-score	Criterion
Item 1										0.45	0.22
Item 2	0.32									0.59	0.21
Item 3	0.27	0.50								0.56	0.23
Item 4	0.41	0.31	0.31							0.51	0.28
Item 5	0.27	0.32	0.38	0.34						0.53	0.20
Item 6	0.15	0.42	0.36	0.15	0.30					0.40	0.22
Item 7	0.22	0.38	0.27	0.29	0.31	0.28				0.47	0.26
Item 8	0.36	0.42	0.44	0.42	0.49	0.28	0.44			0.67	0.25
Item 9	0.19	0.19	0.18	0.10	0.20	0.10	0.14	0.33		0.28	0.13
Item 10	0.42	0.46	0.44	0.57	0.41	0.27	0.35	0.55	0.26	0.68	0.35

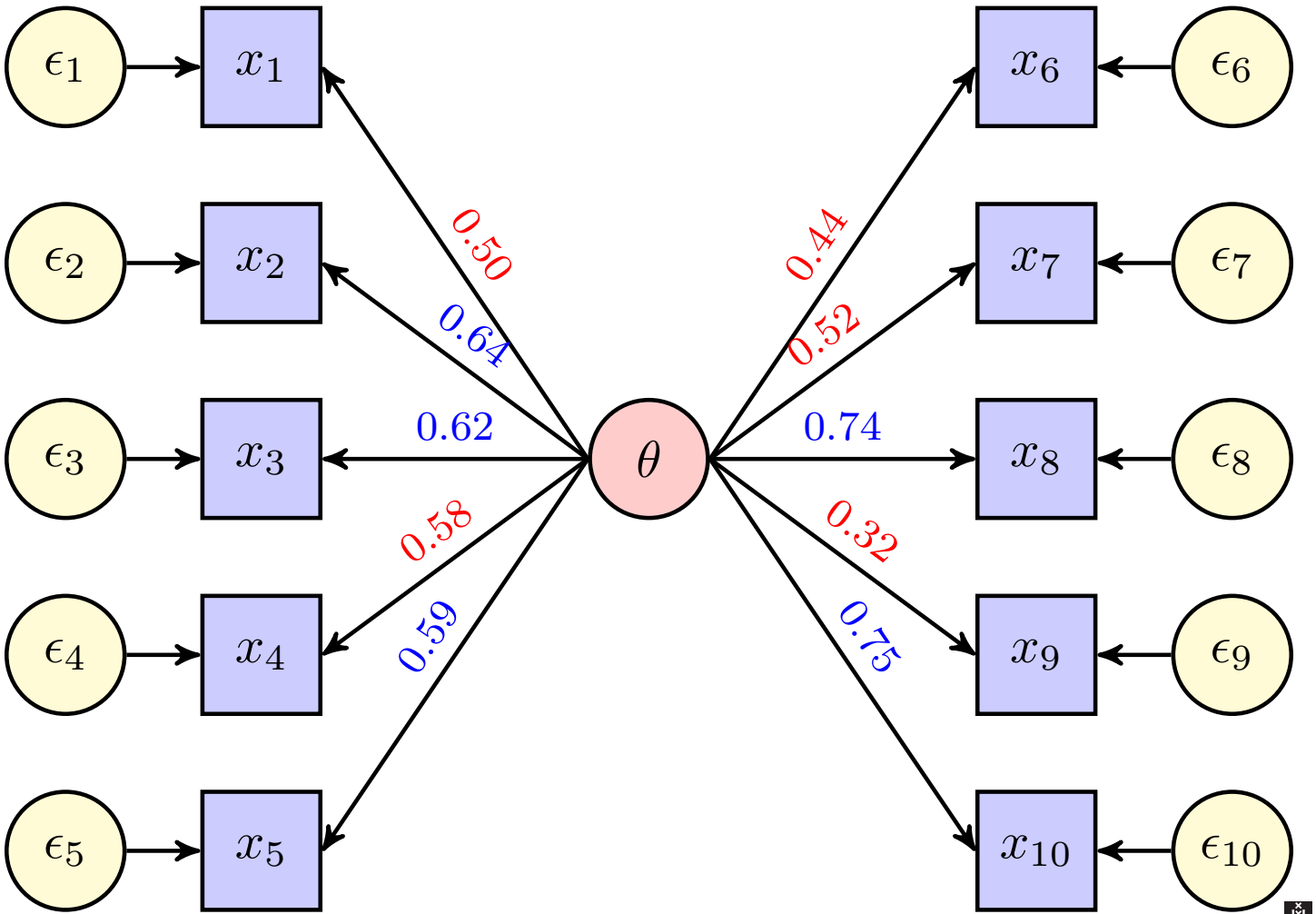
Correlation matrix

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Rest-score	Criterion
Item 1										0.45	0.22
Item 2	0.32									0.59	0.21
Item 3	0.27	0.50								0.56	0.23
Item 4	0.41	0.31	0.31							0.51	0.28
Item 5	0.27	0.32	0.38	0.34						0.53	0.20
Item 6	0.15	0.42	0.36	0.15	0.30					0.40	0.22
Item 7	0.22	0.38	0.27	0.29	0.31	0.28				0.47	0.26
Item 8	0.36	0.42	0.44	0.42	0.49	0.28	0.44			0.67	0.25
Item 9	0.19	0.19	0.18	0.10	0.20	0.10	0.14	0.33		0.28	0.13
Item 10	0.42	0.46	0.44	0.57	0.41	0.27	0.35	0.55	0.26	0.68	0.35

Correlation matrix

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Rest-score	Criterion
Item 1										0.45	0.22
Item 2	0.32									0.59	0.21
Item 3	0.27	0.50								0.56	0.23
Item 4	0.41	0.31	0.31							0.51	0.28
Item 5	0.27	0.32	0.38	0.34						0.53	0.20
Item 6	0.15	0.42	0.36	0.15	0.30					0.40	0.22
Item 7	0.22	0.38	0.27	0.29	0.31	0.28				0.47	0.26
Item 8	0.36	0.42	0.44	0.42	0.49	0.28	0.44			0.67	0.25
Item 9	0.19	0.19	0.18	0.10	0.20	0.10	0.14	0.33		0.28	0.13
Item 10	0.42	0.46	0.44	0.57	0.41	0.27	0.35	0.55	0.26	0.68	0.35

Select under one factor model



Select under predictive model

- Selection of items as ‘predictors’.
- Regression parameters fixed at unity.
- Stepwise, backward, forward or lasso.
- Here all possible subsets.

Scale	Items selected					Mean λ	Pred. validity
Measurement-based	Item 2	Item 3	Item 5	Item 8	Item 10	0.67	0.33
Prediction-based	Item 1	Item 4	Item 6	Item 7	Item 10	0.55	0.40

Scale	Items selected					Mean λ	Pred. validity
Measurement-based	Item 2	Item 3	Item 5	Item 8	Item 10	0.67	0.33
Prediction-based	Item 1	Item 4	Item 6	Item 7	Item 10	0.55	0.40

Scale	Items selected					Mean λ	Pred. validity
Measurement-based	Item 2	Item 3	Item 5	Item 8	Item 10	0.67	0.33
Prediction-based	Item 1	Item 4	Item 6	Item 7	Item 10	0.55	0.40

Scale	Items selected					Mean λ	Pred. validity
Measurement-based	Item 2	Item 3	Item 5	Item 8	Item 10	0.67	0.33
Prediction-based	Item 1	Item 4	Item 6	Item 7	Item 10	0.55	0.40

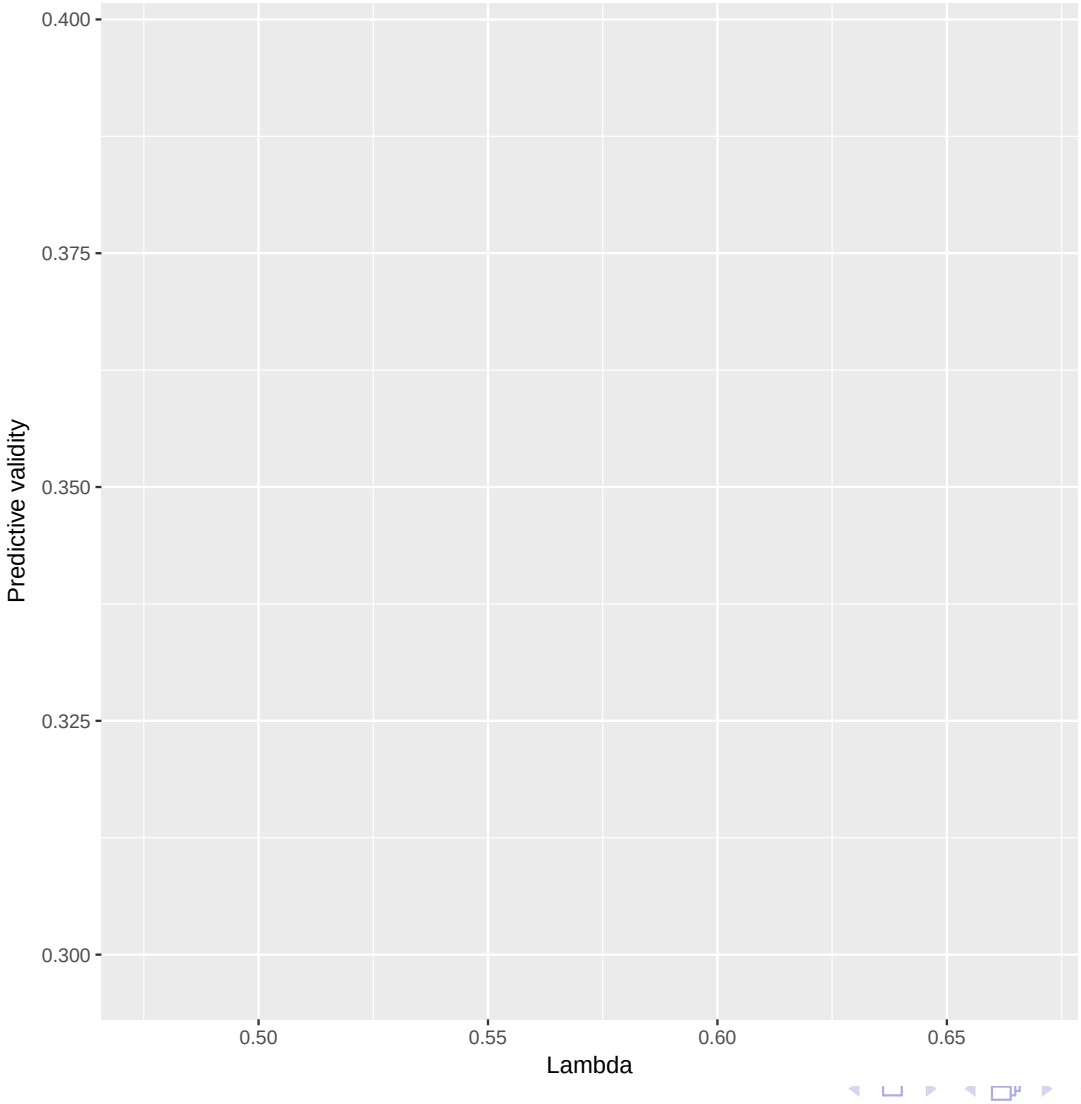
Optimize predictive validity

$$\rho_{XY} = \frac{\sum_{i=1}^{n} \sigma_i \rho_{iY}}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_i \sigma_j \rho_{ij}}}$$

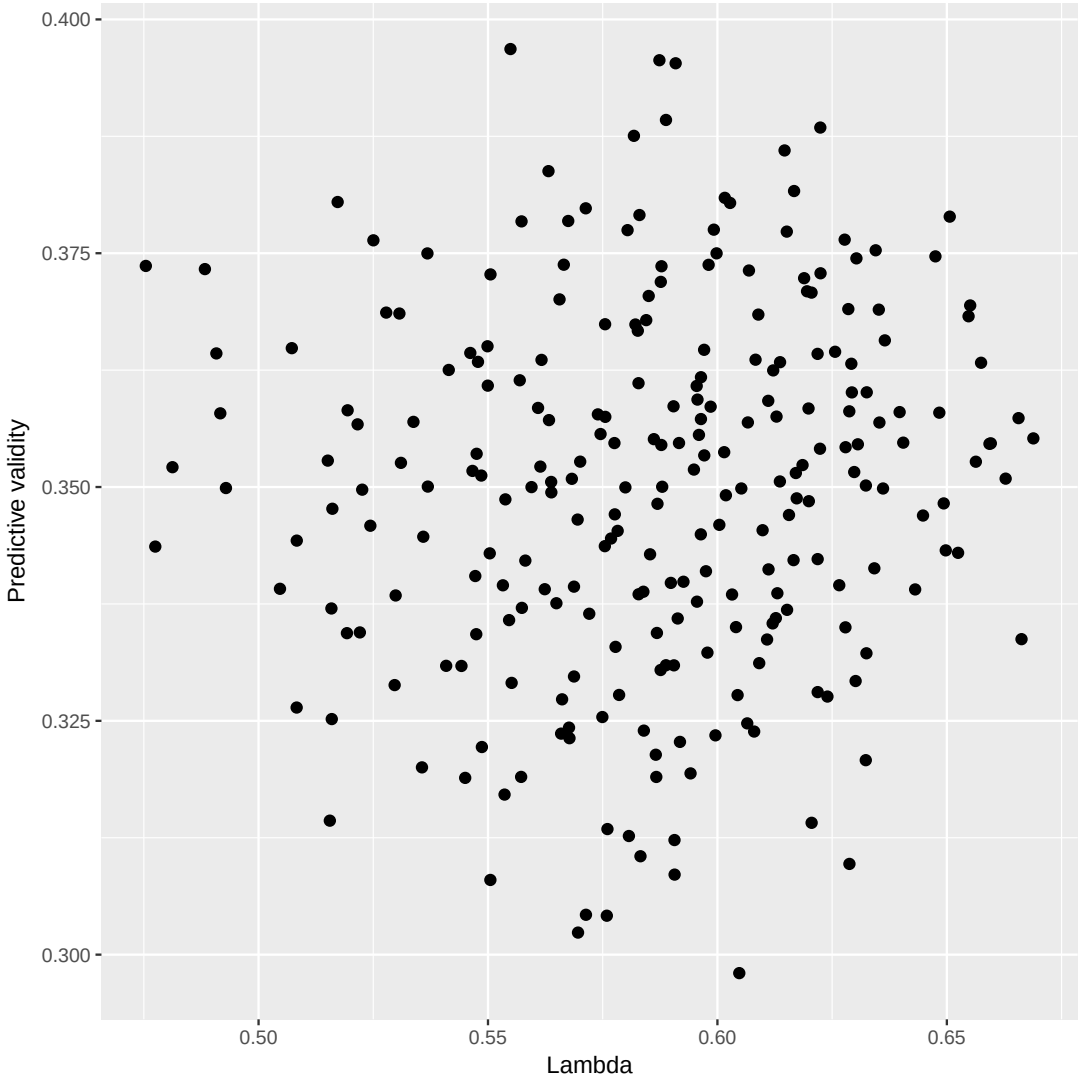
Optimize predictive validity

$$\rho_{XY} = \frac{\sum_{i=1}^{n} \sigma_i \rho_{iY}}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_i \sigma_j \rho_{ij}}}$$

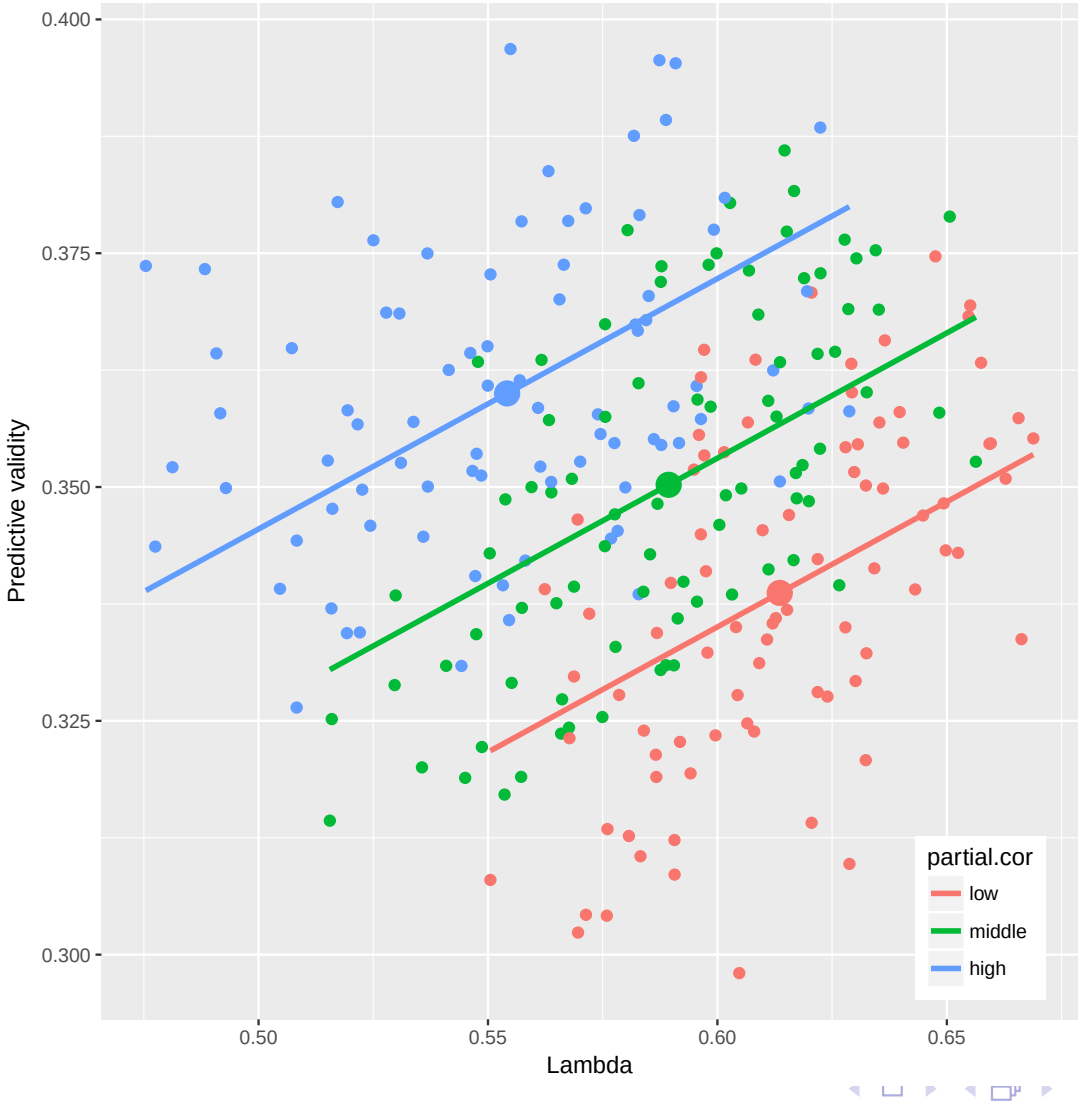
Looking at all combinations ($N = 252$)



Looking at all combinations ($N = 252$)



Looking at all combinations ($N = 252$)



Louis Guttman (1941)

"For a set of n quantitative variates, the derived variate with maximum internal consistency can be shown to be obtainable from the major principal axis of the matrix of the intercorrelations of the n variates (assuming the variates equally important). Now, this axis is the best single axis for approximating the matrix of intercorrelations (in the sense of least squares). As a consequence, it is the worst single axis for approximating the inverse of this matrix.
It is the inverse of the correlation matrix that occurs in linear prediction formulas for actual computations. Therefore, given no information about the criterion variate, we should expect that the single variate derived from the set which afforded the best approximation to the inverse matrix would tend to give better prediction. And conversely, we should expect that the single variate that afforded the worst approximation to the inverse would tend to give worse prediction.
Therefore, we should expect the variate most internally consistent for the set to tend to afford the worst prediction for a random variate outside the set."

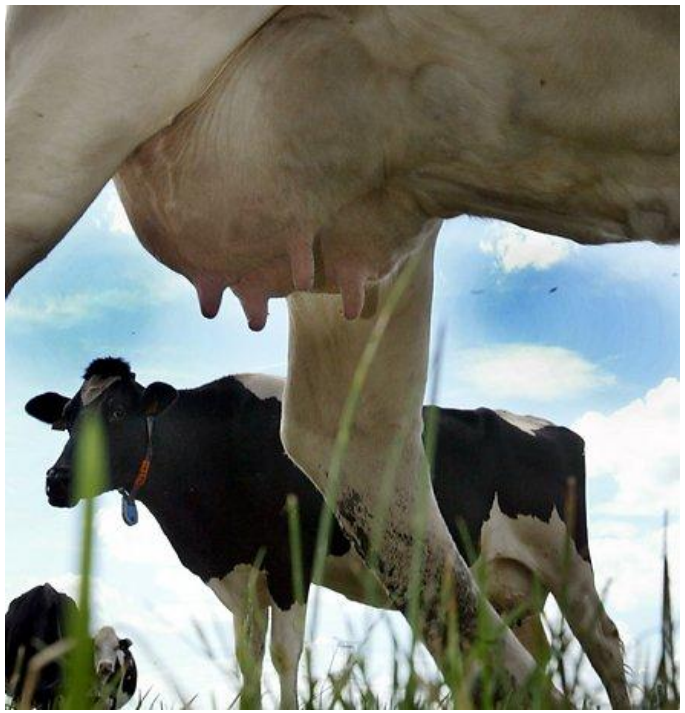
McDonald (1999)

"The apparent paradox is resolved, in effect, by denying that (7.23) is generally true. Here, X is a sum of item scores that fit a common factor model and their "error terms" are their unique parts. By definition, these m terms $E_1 \ldots, E_m$ are mutually uncorrelated, but the truth of (7.23) rests on the additional assumption that they are uncorrelated with all other variables. Such an assumption is extremely strong, generally falsifiable, and generally false. ... Thus, to make a good practical predictor, it is appropriate to choose an item set with good predictive utility, with no concern for its reliability/construct validity as measured by omega or bounded by alpha. ***In effect, this uses the relations of the unique parts of the items to the criterion to maximize predictive ability."***

$$\rho_{XY} = \rho_{TY}\sqrt{\rho_{XX'}}. \qquad (7.23)$$

Alternative explanation

You can't have your cake and eat it too!



Solutions

- Construct a single predictive scale, but accept that measurement model may be poor.
- Use a test battery (consisting of multiple scales with sound measurement precision) but be ready to pay for the extra costs.

Table 2.1 Questionnaire design methods

Class	Intuitive		Inductive		Deductive	
Method	Rational	Prototypical	Internal	External	Construct	Facet
Aspect	face validity	process validity	homo-geneity	criterion validity	construct validity	content validity
Step						
theoretical framework	work definition	–	–	–	nomological network	theoretical context
concept analysis	global description or typology	–	–	–	precise definitions, demarcation	facets and facet elements
item specifi-cation	informal criteria	–	homo-geneous	hetero-geneous	clear, homo-geneous, content saturated	mapping sentence
item production	diagnostic questions	act-nomination	no guidelines	no guidelines	based on definitions	based on definitions
item judgment	review by experts	redaction	no guidelines	no guidelines	item content and pilot results	item content and pilot results
scale construc-tion	face validity	proto-typicality ratings	dimensional analysis	item-criterion relation	convergent and discriminant item validities	dimen-sionality
validation	diagnostic comparison	reliability validity	cross validation test of mini theory	cross validation retest reliability	reliability convergent and discriminant validity	reliability validity model fit

Oosterveld, Vorst & Smits (2018)

Table 2.1 Questionnaire design methods

Class	Intuitive		Inductive		Deductive	
Method	Rational	Prototypical	Internal	External	Construct	Facet
Aspect	face validity	process validity	homo-geneity	criterion validity	construct validity	content validity
Step						
theoretical framework	work definition	–	–	–	nomological network	theoretical context
concept analysis	global description or typology	–	–	–	precise definitions, demarcation	facets and facet elements
item specifi-cation	informal criteria	–	homo-geneous	hetero-geneous	clear, homo-geneous, content saturated	mapping sentence
item production	diagnostic questions	act-nomination	no guidelines	no guidelines	based on definitions	based on definitions
item judgment	review by experts	redaction	no guidelines	no guidelines	item content and pilot results	item content and pilot results
scale construc-tion	face validity	proto-typicality ratings	dimensional analysis	item-criterion relation	convergent and discriminant item validities	dimen-sionality
validation	diagnostic comparison	reliability validity	cross validation test of mini theory	cross validation retest reliability	reliability convergent and discriminant validity	reliability validity model fit

Oosterveld, Vorst & Smits (2018)

Table 2.1 Questionnaire design methods

Class	Intuitive		Inductive		Deductive	
Method	Rational	Prototypical	Internal	External	Construct	Facet
Aspect	face validity	process validity	homo-geneity	criterion validity	construct validity	content validity
Step						
theoretical framework	work definition	–	–	–	nomological network	theoretical context
concept analysis	global description or typology	–	–	–	precise definitions, demarcation	facets and facet elements
item specifi-cation	informal criteria	–	homo-geneous	hetero-geneous	clear, homo-geneous, content saturated	mapping sentence
item production	diagnostic questions	act-nomination	no guidelines	no guidelines	based on definitions	based on definitions
item judgment	review by experts	redaction	no guidelines	no guidelines	item content and pilot results	item content and pilot results
scale construc-tion	face validity	proto-typicality ratings	dimensional analysis	item-criterion relation	convergent and discriminant item validities	dimen-sionality
validation	diagnostic comparison	reliability validity	cross validation test of mini theory	cross validation retest reliability	reliability convergent and discriminant validity	reliability validity model fit

Oosterveld, Vorst & Smits (2018)

Conclusions

- Many construction methods exist.
- Trade-offs are everywhere.
- Measurement vs prediction.
- But also construct validity vs homogeneity.
- No test can excel at all goals.
- Different tests for different goals.

Conclusions

- Many construction methods exist.
- Trade-offs are everywhere.
- Measurement vs prediction.
- But also construct validity vs homogeneity.
- No test can excel at all goals.
- Different tests for different goals.

Conclusions

- Many construction methods exist.
- Trade-offs are everywhere.
- Measurement vs prediction.
- But also construct validity vs homogeneity.
- No test can excel at all goals.
- Different tests for different goals.

Conclusions

- Many construction methods exist.
- Trade-offs are everywhere.
- Measurement vs prediction.
- But also construct validity vs homogeneity.
- No test can excel at all goals.
- Different tests for different goals.

Conclusions

- Many construction methods exist.
- Trade-offs are everywhere.
- Measurement vs prediction.
- But also construct validity vs homogeneity.
- No test can excel at all goals.
- Different tests for different goals.

Conclusions

- Many construction methods exist.
- Trade-offs are everywhere.
- Measurement vs prediction.
- But also construct validity vs homogeneity.
- No test can excel at all goals.
- Different tests for different goals.

Thanks for your attention!

n.smits@uva.nl

Guttman, L. (1941). An outline of the statistical theory of prediction. Supplementary study B-1. In Subcommittee on Prediction of Social Adjustment (Ed.), *The prediction of personal adjustment.* New York: Social Science Research Council.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Smits, N., van der Ark, L. A., & Conijn, J. M. (2017). Measurement versus prediction in the construction of patient-reported outcome questionnaires: Can we have our cake and eat it? *Quality of Life Research*. doi: 10.1007/s11136-017-1720-4