

Evaluating screening measures and implications for health equity

OSCAR GONZALEZ, PHD

TEST THEORY – PSYCHOMETRICS – MEASUREMENT

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

Research Interests

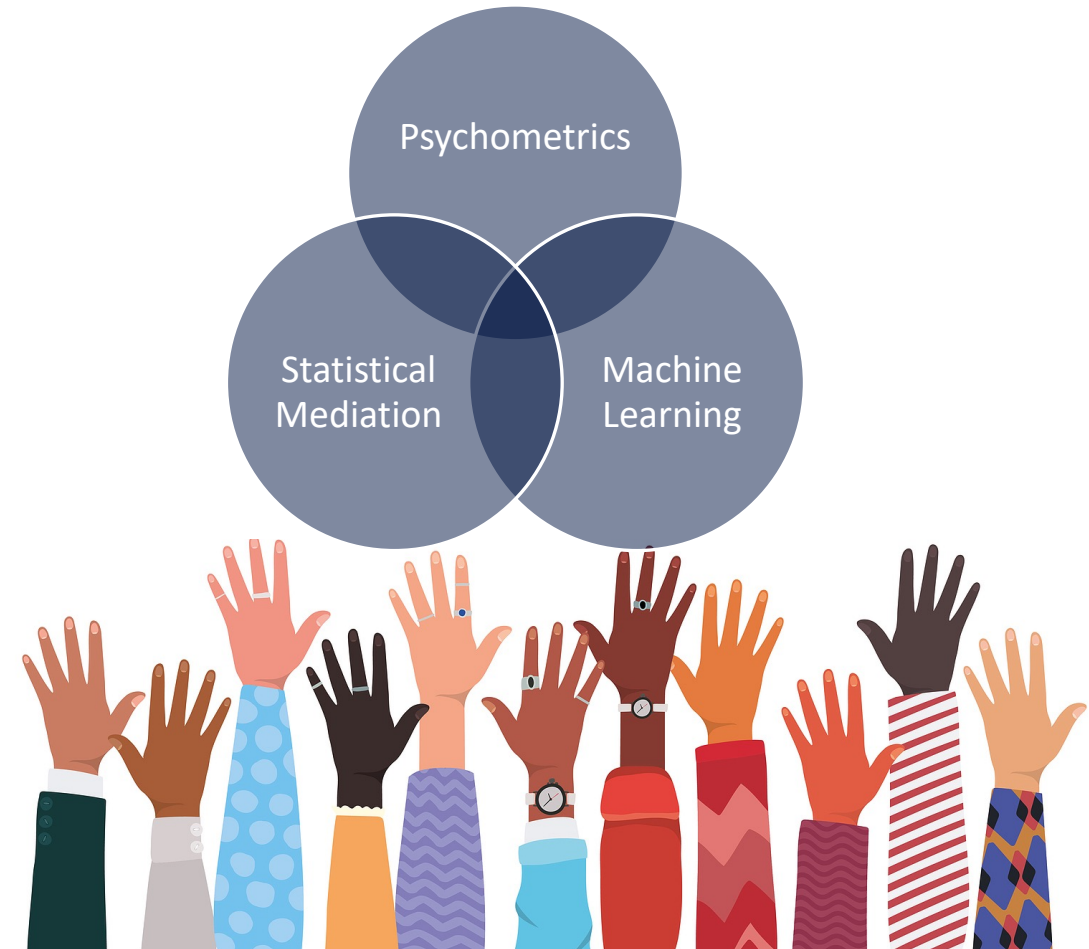
Psychological measurement

Plays a critical role in understanding human behavior and decision making

- Screening for disorders
- Evaluating prevention programs

Poor measurement = health inequities




Health equity: everyone has the opportunity to attain their highest level of health [APHA]



Measures for Selection

What are we are measuring?.. and are we doing it well?



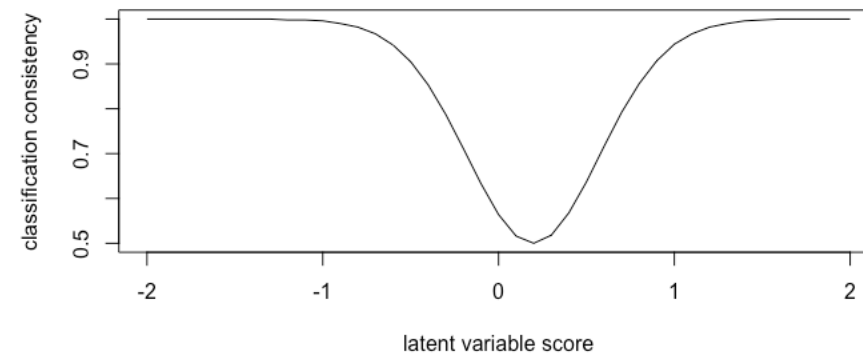
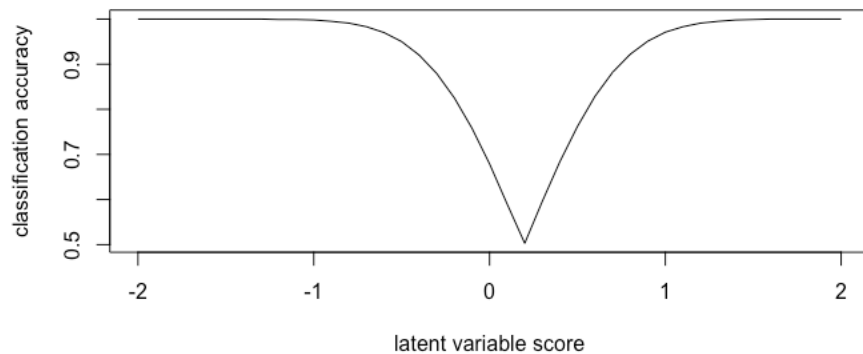
-  Accuracy and Consistency
-  Health Equity and Item Bias
-  Classification Uncertainty

I became very interested in...

Measures used for classifications/decisions

Plots...

- Describe the classification performance of a measure as a function of the latent variable
- Easy to communicate results to assessment specialists



Today's talk

Four projects on the classification performance of screening measures

Outline

- General background
- Current methods + problems
- New methods
- Extensions and future work



On Making Decisions

OSCAR GONZALEZ, PHD

TEST THEORY – PSYCHOMETRICS – MEASUREMENT

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL



A bit of history

Decision making has been an important methodological issue for social scientists

The war efforts were influential

Statistics and assessment

- Army Alpha and Army Beta in WWI
- Personnel Selection
- College Entrance Exams



Other Areas

Recruitment into studies

Accreditation and licensure

- CPA, Law, Medicine

Medical Diagnoses

- Structured Clinical Interview

Standard setting ["setting standards"]

- Educational achievement



Screening measures

Used in psychology and public health to select respondents

Common uses:

- support a diagnosis vs. reject a diagnosis
- provide services vs. withhold services
- conduct further assessment vs. leave alone

Screening process:

- Obtain item responses
- Aggregate into a score (typically, a sum score X^*)
- Compare X^* to a cut score



Geriatric Depression Scale (GDS-15)

Measure used in clinical and research settings to screen older adults for depression (Sheikh & Yesavage, 1986)

15 binary items rated *yes* or *no*

- *Are you basically satisfied with your life?*
- *Do you feel happy most of the time?*
- *Do you feel worthless on the way you are now?*

Sum responses, probable depression if score is ≥ 5



Desired properties of screening process

Accurate

- Select respondents who actually have the condition from the rest

Consistent

- Same decision across repeated testing

Equitable

- The decision depends only on the assessed construct, not on age, race, ethnicity...
- aka *no item bias*



Research

When items have bias, screeners may over or under flag individuals from specific groups

- Impacts **health equity** - who gets selected into intervention programs or who receives services

My research

- Methods to estimate classification accuracy and classification consistency
- Describe how item bias affects accuracy and consistency



Estimating accuracy and consistency

OSCAR GONZALEZ, PHD

TEST THEORY – PSYCHOMETRICS – MEASUREMENT

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

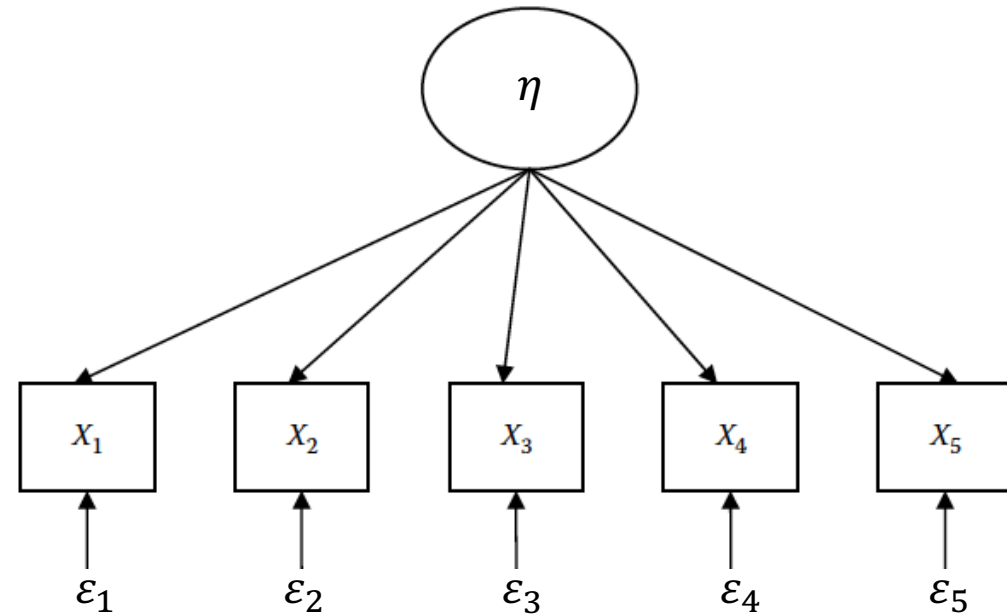
Psychometrics of Screening Measures

Most screening measures assess a construct, η

- Constructs affect the way we respond to items
- Reason why items correlate
- Catch: we cannot observe η

Factor models

- Describe the relation between the construct and the item
- There are many types of models



Diagram

- Circles are latent variables
- Squares are observed variables - items

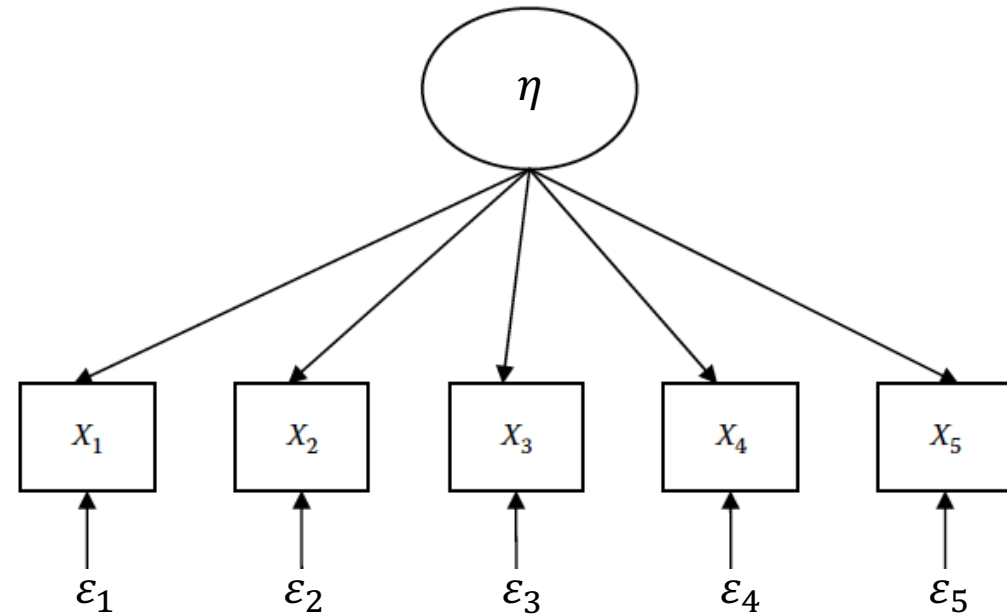
Psychometrics of Screening Measures

Most screening measures assess a construct, η

- Constructs affect the way we respond to items
- Reason why items correlate
- Catch: we cannot observe η

Linear factor model

- Describes the linear relation between η and X 's
- $X_{ij} = \tau_j + \lambda_j\eta_i + \varepsilon_{ij}$



Diagram

- Circles are latent variables
- Squares are observed variables - items

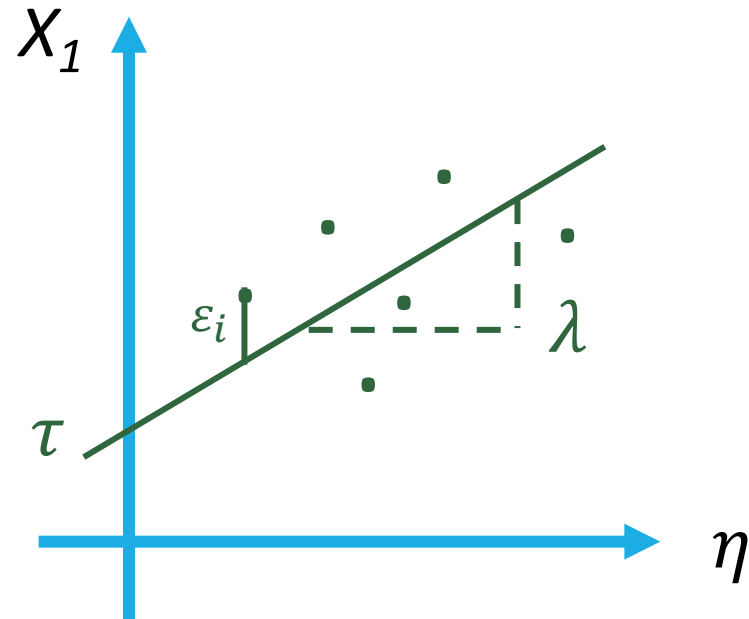
Psychometrics of Screening Measures

Most screening measures assess a construct, η

- Constructs affect the way we respond to items
- Reason why items correlate
- Catch: we cannot observe η

Linear factor model

- Describes the linear relation between η and X 's
- $X_{ij} = \tau_j + \lambda_j\eta_i + \varepsilon_{ij}$



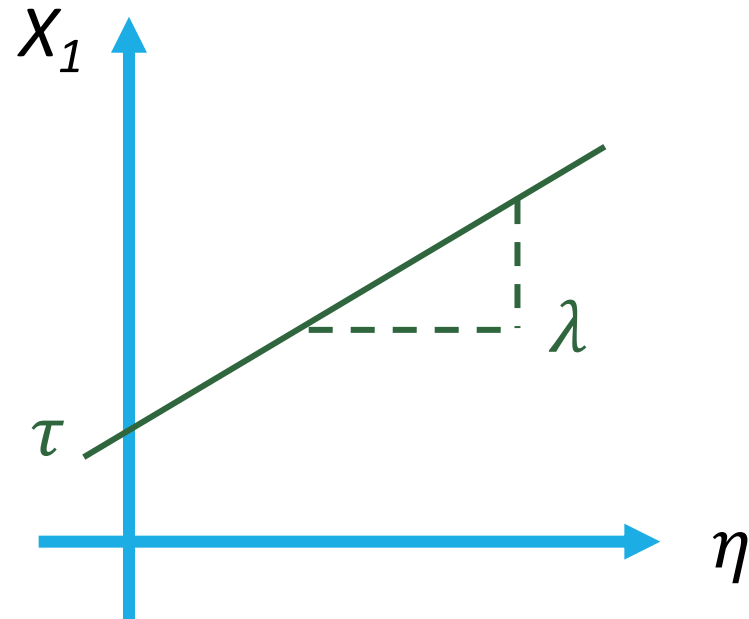
Psychometrics of Screening Measures

Most screening measures assess a construct, η

- Constructs affect the way we respond to items
- Reason why items correlate
- Catch: we cannot observe η

Linear factor model

- Describes the linear relation between η and X 's
- $X_{ij} = \tau_j + \lambda_j\eta_i + \varepsilon_{ij}$



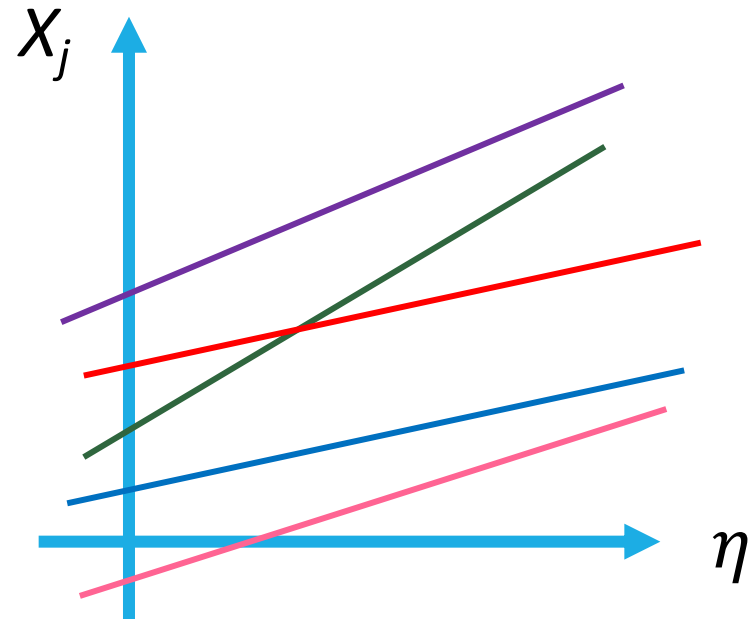
Psychometrics of Screening Measures

Most screening measures assess a construct, η

- Constructs affect the way we respond to items
- Reason why items correlate
- Catch: we cannot observe η

Linear factor model

- Describes the linear relation between η and X 's
- $X_{ij} = \tau_j + \lambda_j\eta_i + \varepsilon_{ij}$



Note: all the same predictor, η , but different vertical axes

Study Design

1. Recruit individuals with and without the condition
 - Administer gold standard if condition is unknown
2. Administer screener (perhaps 2x)
3. Score the responses
4. Determine a cutpoint that differentiates individuals with and without the condition



Typical method: 2 × 2 Table

Accuracy

- Screener prediction × True condition
- % of correct classifications

Consistency

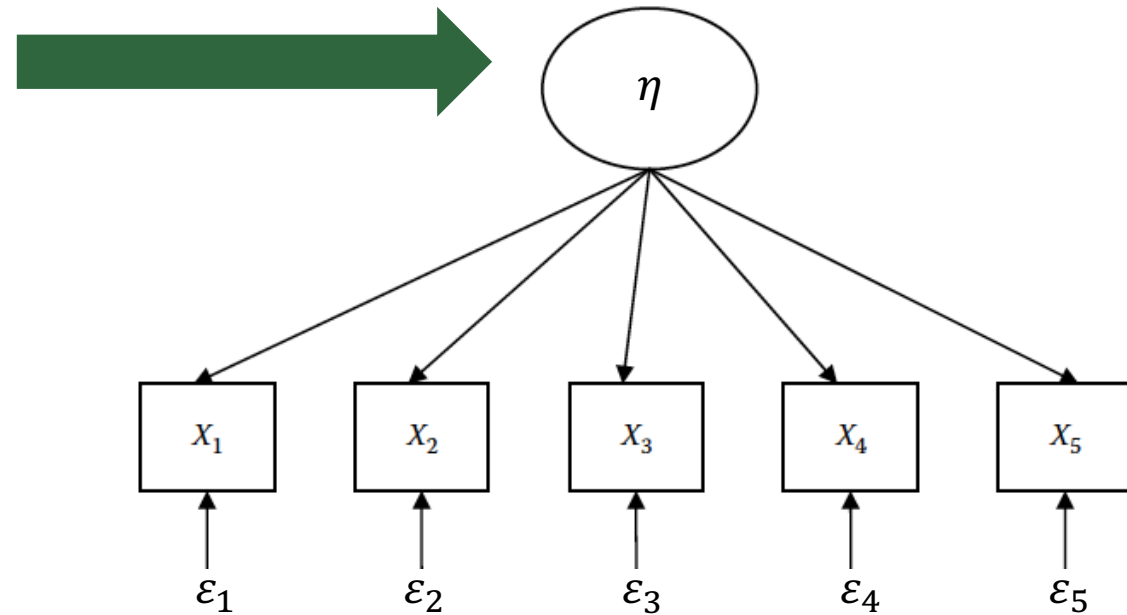
- Screener prediction T1 × prediction T2
- % of agreement

		True Diagnosis	
		No	Yes
Screener Predict T1	No	A	B
	Yes	C	D

		Screener Predict T2	
		No	Yes
Screener Predict T1	No	E	F
	Yes	G	H

Problems...

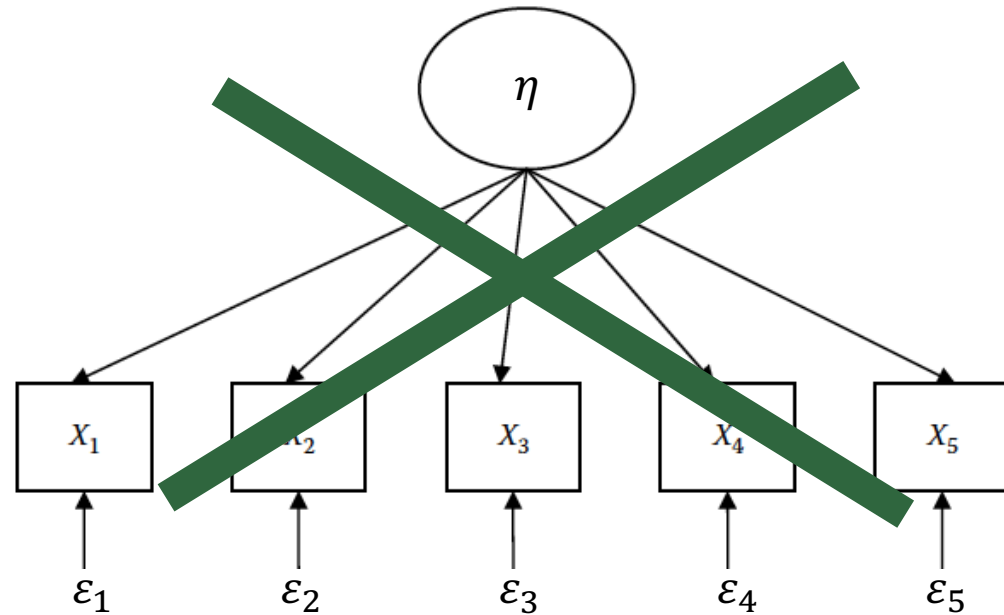
Ideally, we would make decisions based on latent variable, η , assessed by the items



Problems...

Ideally, we would make decisions based on latent variable, η , assessed by the items

- In applied settings, we use X^* , which correlates with η , but has measurement error
- Do the decisions based on X^* and η agree? If so, how much?



$$X^* = X_1 + X_2 + X_3 + X_4 + X_5$$

Problems...

Measurement error might not affect decisions on the screener for some individuals...

Problems...

Measurement error might not affect decisions on the screener for some individuals...

Consider a screener with a cutpoint of 20

- Time 1: $X_1^* = 10$
- Time 2: $X_2^* = 8$
- Time 3: $X_3^* = 11 \dots$

Problems...

Measurement error might not affect decisions on the screener for some individuals...

Consider a screener with a cutpoint of 20

- Time 1: $X_1^* = 10$
- Time 2: $X_2^* = 8$
- Time 3: $X_3^* = 11 \dots$

Scores might bounce around, but $X_t^* < 20$, so same decision...

- Can we identify these regions?
- People who are high or low on η

How about we leverage the properties of the linear factor model?

ASSUMING THE MODEL FITS...

How Accurate and Consistent Are Score-Based Assessment Decisions? A Procedure Using the Linear Factor Model

Oscar Gonzalez¹ , A. R. Georgeson²,
and William E. Pelham III³

Abstract

When scales or tests are used to make decisions about individuals (e.g., to identify which adults should be assessed for psychiatric disorders), it is crucial that these decisions be accurate and consistent. However, it is not obvious how to assess accuracy and consistency when the scale was administered only once to a given sample and the true condition based on the latent variable is unknown. This article describes a method based on the linear factor model for evaluating the accuracy and consistency of scale-based decisions using data from a single administration of the scale. We illustrate the procedure and provide R code that investigators can use to apply the method in their own data. Finally, in a simulation study, we evaluate how the method performs when applied to discrete (vs. continuous) items, a practice that is common in published literature. The results suggest that the method is generally robust when applied to discrete items.

Assessment

1–11

© The Author(s) 2022

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/10731911221113568

journals.sagepub.com/home/asm



Solution: use model-implied estimates of accuracy and consistency

Proposed Methods

OSCAR GONZALEZ, PHD

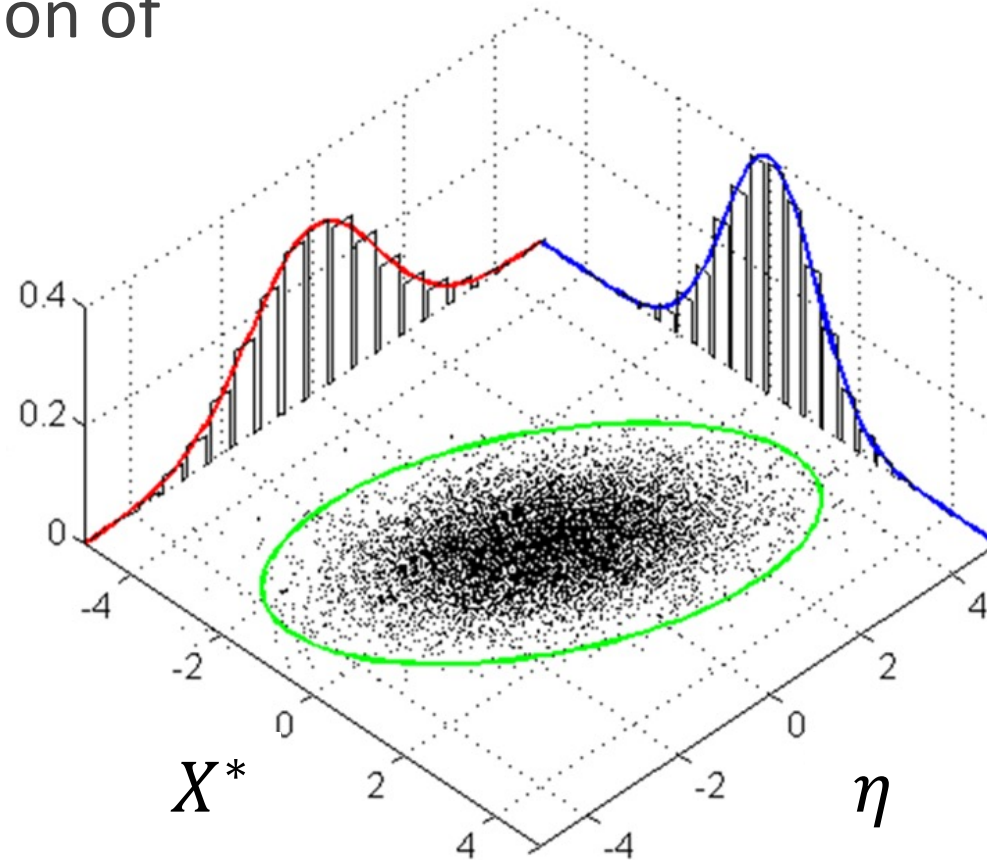
TEST THEORY – PSYCHOMETRICS – MEASUREMENT

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL



Bivariate (joint) distribution

Understand relation of X^* and η

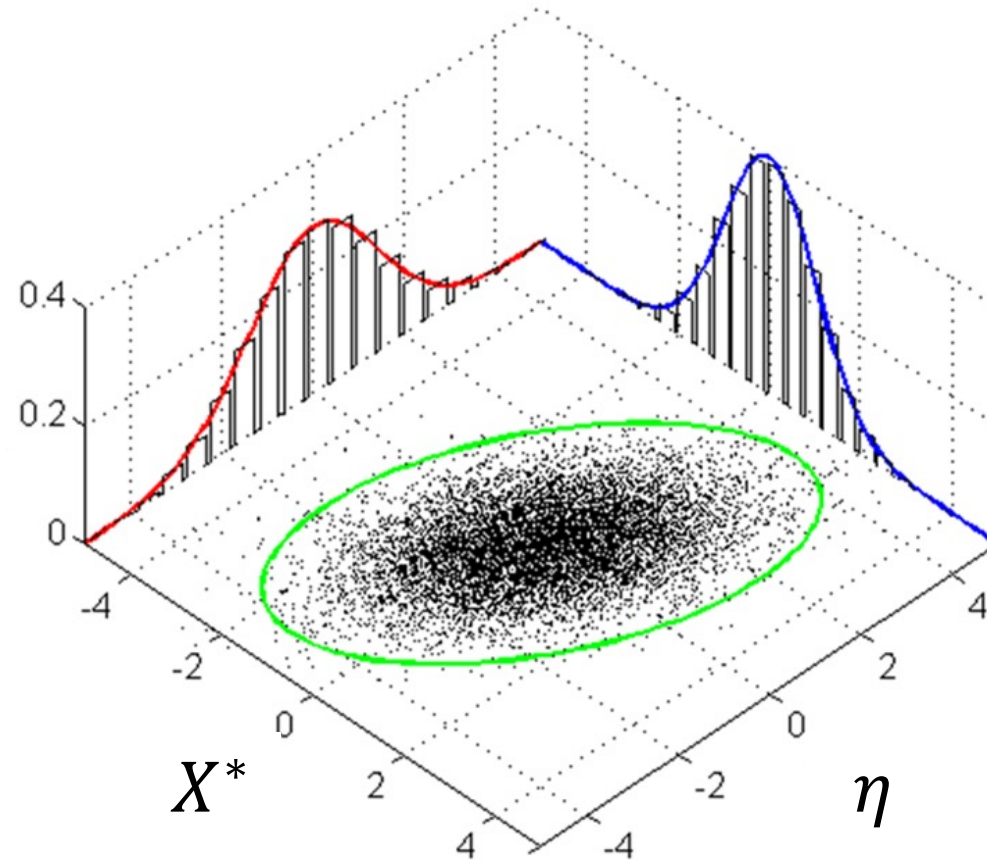


Gonzalez et al., 2023 Assessment

Bivariate (joint) distribution

We need to know:

- Mean of X^*
- Variance of X^*
- Correlation of X^* and η
- Correlation of X_1^* and X_2^*



We pre-specify:

- Mean of η , κ
- Variance of η , Φ

Gonzalez et al., 2023 Assessment

Relation between X^* and η

If:

$$X_j = \tau_j + \lambda_j \eta + \varepsilon_j$$

and...

$$X^* = X_1 + X_2 + X_3$$

Relation between X^* and η

If:

$$X_j = \tau_j + \lambda_j \eta + \varepsilon_j$$

and...

$$X^* = X_1 + X_2 + X_3$$

Then...

$$X^* = (\tau_1 + \lambda_1 \eta + \varepsilon_1) + (\tau_2 + \lambda_2 \eta + \varepsilon_2) + (\tau_3 + \lambda_3 \eta + \varepsilon_3)$$

Relation between X^* and η

If:

$$X_j = \tau_j + \lambda_j \eta + \varepsilon_j$$

and...

$$X^* = X_1 + X_2 + X_3$$

Then...

$$X^* = (\tau_1 + \lambda_1 \eta + \varepsilon_1) + (\tau_2 + \lambda_2 \eta + \varepsilon_2) + (\tau_3 + \lambda_3 \eta + \varepsilon_3)$$

$$X^* = (\tau_1 + \tau_2 + \tau_3) + (\lambda_1 \eta + \lambda_2 \eta + \lambda_3 \eta) + (\varepsilon_1 + \varepsilon_2 + \varepsilon_3)$$

Relation between X^* and η

If:

$$X_j = \tau_j + \lambda_j \eta + \varepsilon_j$$

and...

$$X^* = X_1 + X_2 + X_3$$

Then...

$$X^* = (\tau_1 + \lambda_1 \eta + \varepsilon_1) + (\tau_2 + \lambda_2 \eta + \varepsilon_2) + (\tau_3 + \lambda_3 \eta + \varepsilon_3)$$

$$X^* = (\tau_1 + \tau_2 + \tau_3) + (\lambda_1 \eta + \lambda_2 \eta + \lambda_3 \eta) + (\varepsilon_1 + \varepsilon_2 + \varepsilon_3)$$

$$X^* = (\tau_1 + \tau_2 + \tau_3) + (\lambda_1 + \lambda_2 + \lambda_3) \eta + (\varepsilon_1 + \varepsilon_2 + \varepsilon_3)$$

Factor out

Relation between X^* and η

If:

$$X_j = \tau_j + \lambda_j \eta + \varepsilon_j$$

and...

$$X^* = X_1 + X_2 + X_3$$

Then...

$$X^* = (\tau_1 + \lambda_1 \eta + \varepsilon_1) + (\tau_2 + \lambda_2 \eta + \varepsilon_2) + (\tau_3 + \lambda_3 \eta + \varepsilon_3)$$

$$X^* = (\tau_1 + \tau_2 + \tau_3) + (\lambda_1 \eta + \lambda_2 \eta + \lambda_3 \eta) + (\varepsilon_1 + \varepsilon_2 + \varepsilon_3)$$

$$X^* = (\tau_1 + \tau_2 + \tau_3) + (\lambda_1 + \lambda_2 + \lambda_3) \eta + (\varepsilon_1 + \varepsilon_2 + \varepsilon_3)$$

$$X^* = \tau^* + \lambda^* \eta + \varepsilon^*$$

Gonzalez et al., 2023 Assessment

Some relations [if the world works like FA]

Mean and variance of the summed score:

$$\mu_{X^*} = \tau^* + \lambda^* \kappa \quad \sigma_{X^*}^2 = \lambda^{*2} \Phi + \psi^*.$$

Correlation of X^* and η :

$$\text{Cor}(X^*, \eta) = \frac{\text{Cov}(X^*, \eta)}{\sqrt{\text{Var}(X^*)} \sqrt{\text{Var}(\eta)}} = \frac{\text{Cov}(\tau^* + \lambda^* \eta, \eta)}{(\lambda^{*2} \Phi + \psi^*)^{1/2} \Phi^{1/2}} = \frac{\lambda^* \Phi}{(\lambda^{*2} \Phi + \psi^*)^{1/2} \Phi^{1/2}} = \frac{\lambda^* \Phi^{1/2}}{(\lambda^{*2} \Phi + \psi^*)^{1/2}},$$

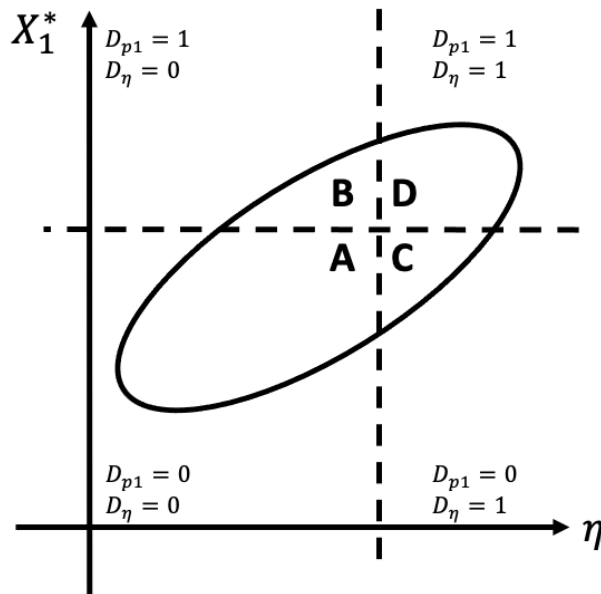
Correlation of X_1^* and X_2^* :

$$\text{Cor}(X_1^*, X_2^*) = \frac{\text{Cov}(X_1^*, X_2^*)}{\sqrt{\text{Var}(X_1^*)} \sqrt{\text{Var}(X_2^*)}} = \frac{\text{Cov}(\tau^* + \lambda^* \eta, \tau^* + \lambda^* \eta)}{(\lambda^{*2} \Phi + \psi^*)^{1/2} (\lambda^{*2} \Phi + \psi^*)^{1/2}} = \frac{\lambda^{*2} \Phi}{\lambda^{*2} \Phi + \psi^*}$$

Gonzalez et al., 2023 Assessment

General Procedure

Accuracy: Determine relation between X^* and η , impose cutpoints, and integrate quadrants

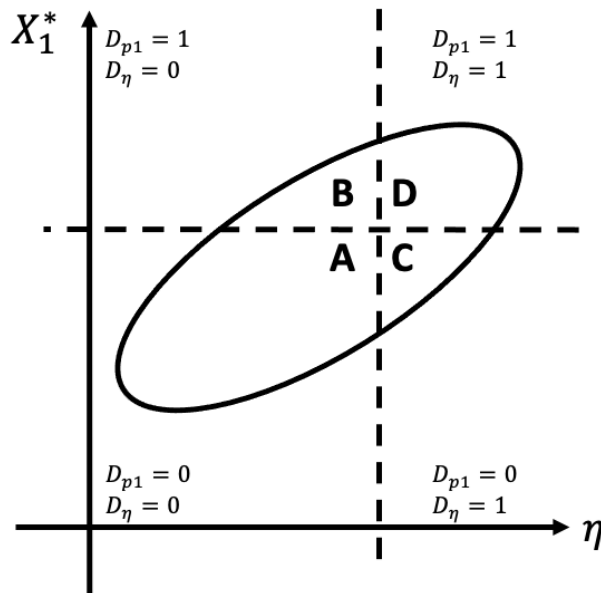


Classification Accuracy: $(A+D) / (A+B+C+D)$

Gonzalez et al., 2023 Assessment

General Procedure

Accuracy: Determine relation between X^* and η , impose cutpoints, and integrate quadrants



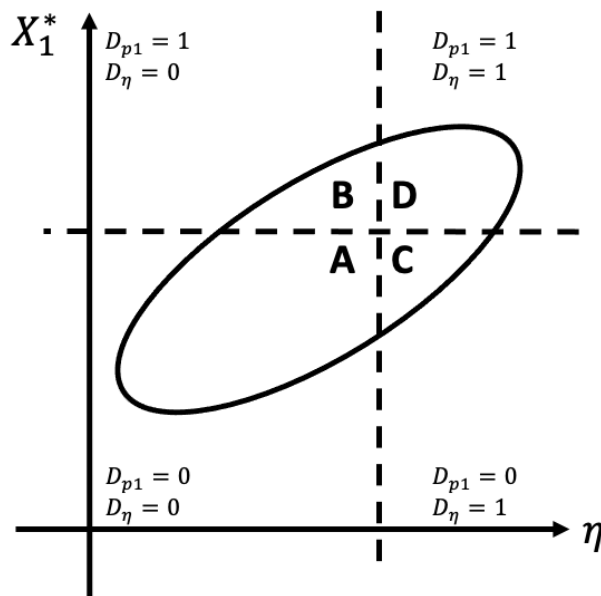
		True Diagnosis	
		No	Yes
Screeners	No	A	B
Predict T1	Yes	C	D

Classification Accuracy: $(A+D) / (A+B+C+D)$

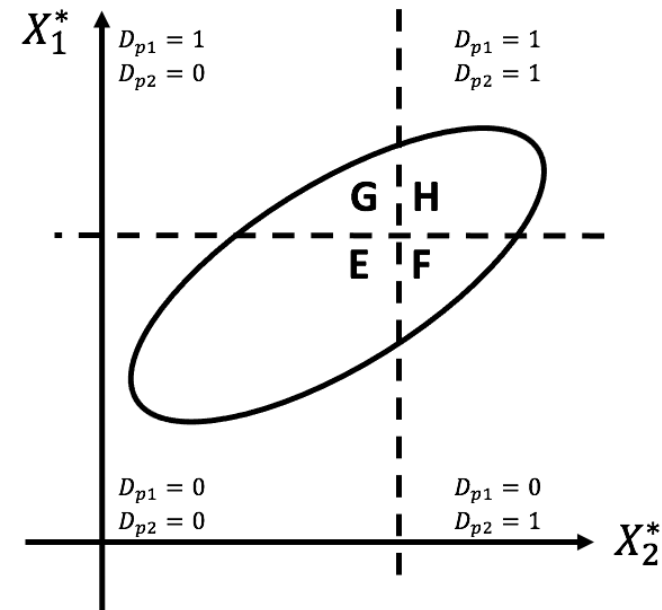
Gonzalez et al., 2023 Assessment

General Procedure

Consistency: Determine relation between X_1^* and X_2^* , impose cutpoints, and integrate quadrants



Classification Accuracy: $(A+D) / (A+B+C+D)$

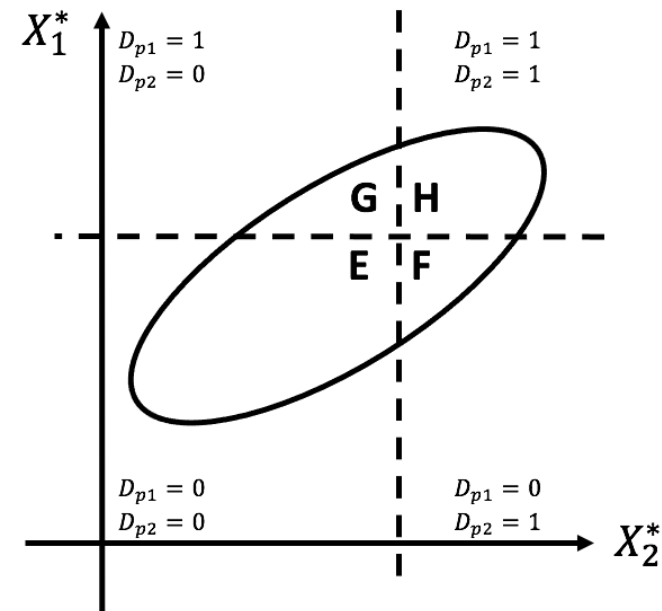


Classification Consistency: $(E+H) / (E+F+G+H)$

General Procedure

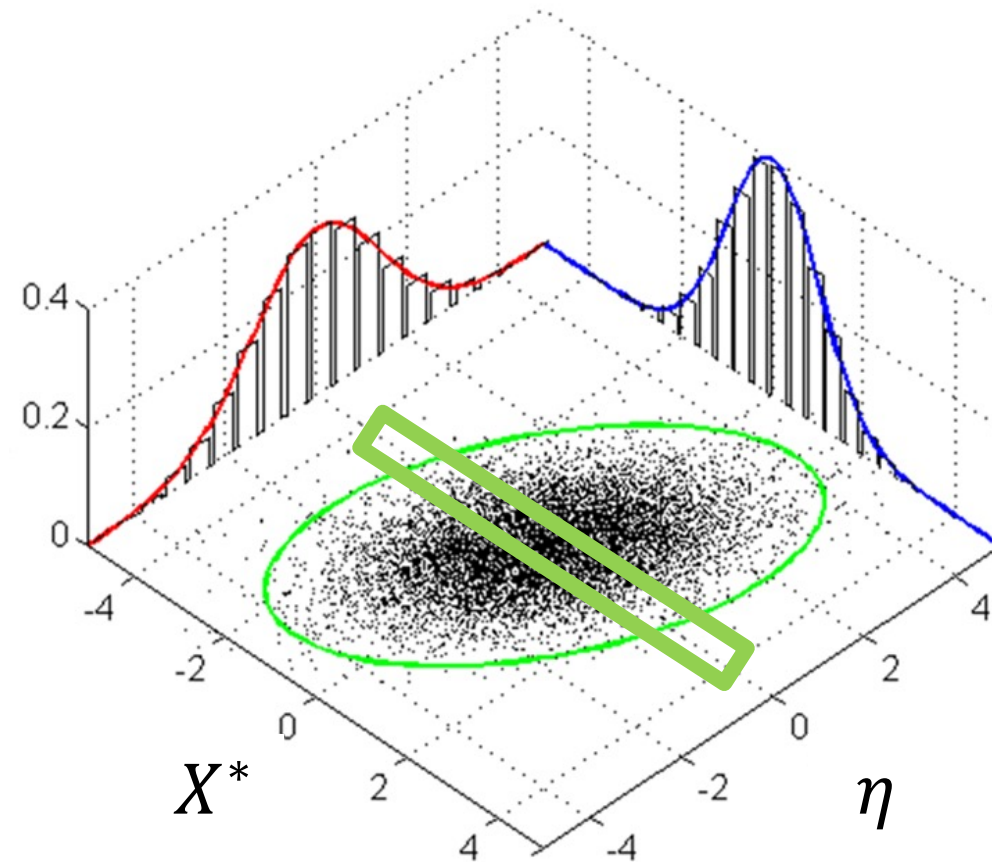
Consistency: Determine relation between X_1^* and X_2^* , impose cutpoints, and integrate quadrants

		Screener Predict T2	
		No	Yes
Screener	No	E	F
Predict T1	Yes	G	H



Classification Consistency: $(E+H) / (E+F+G+H)$

Conditional distribution

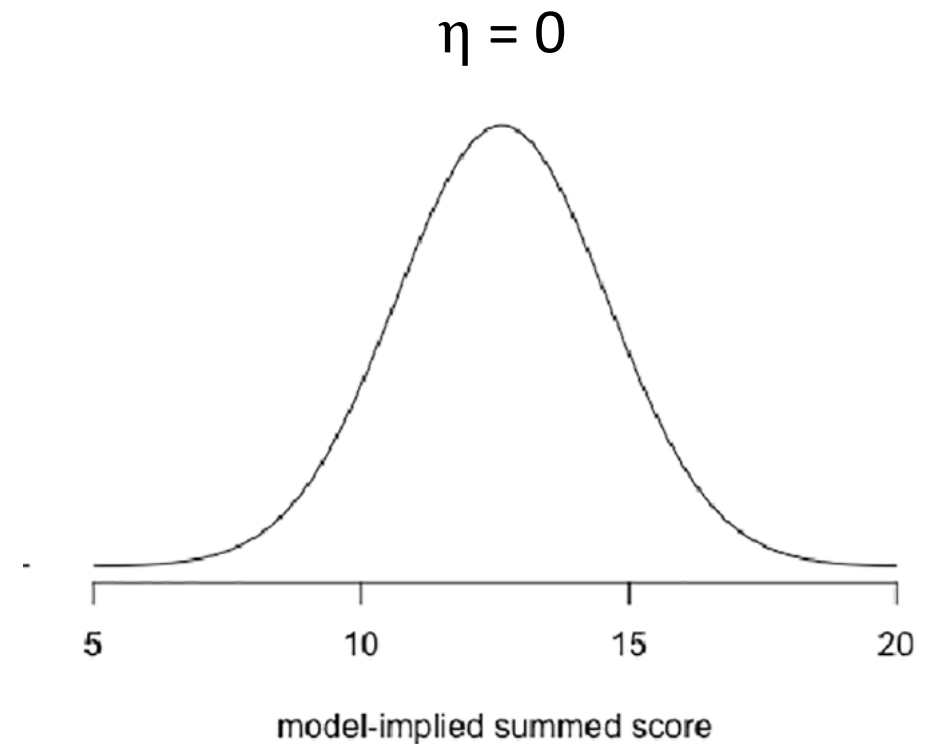


Gonzalez et al., 2023 *Assessment*

Conditional Estimates

At each level of η , there is a distribution of X^*

- $\text{Var}(X^* | \eta) = \text{Var}(\varepsilon)_j^* = \psi_j^*$
- Standard error of measurement



Gonzalez et al., 2023 *Assessment*

Conditional Estimates

At each level of η , there is a distribution of X^*

- $\text{Var}(X^* | \eta) = \text{Var}(\varepsilon)_j^* = \psi_j^*$
- Standard error of measurement

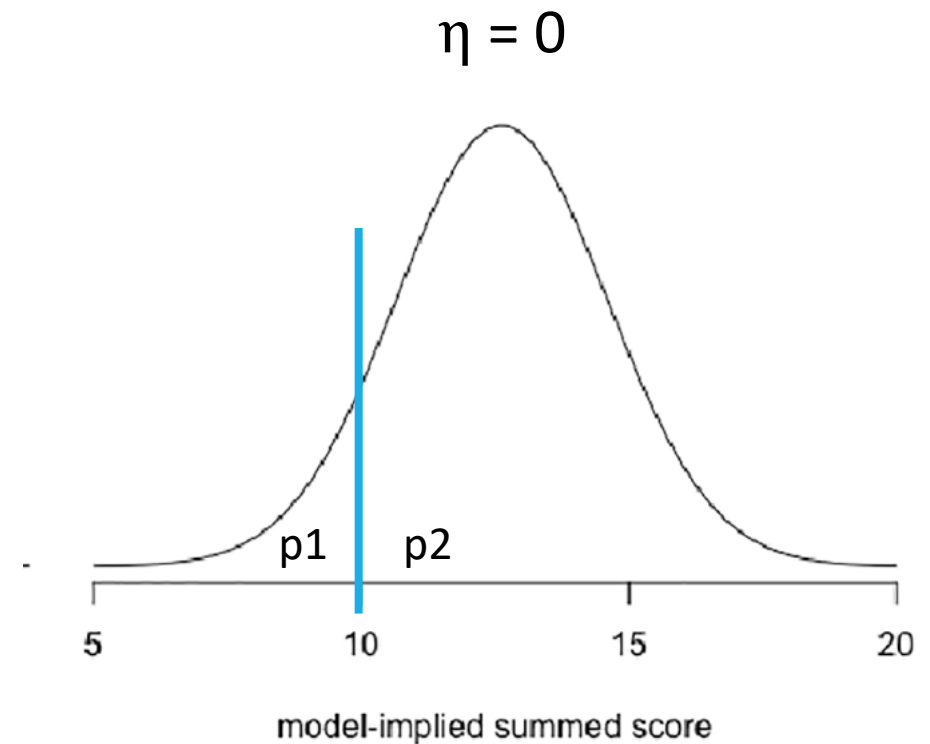
Impose cutpoint on distribution

Conditional Accuracy

- $\max(p1, p2)$

Conditional Consistency

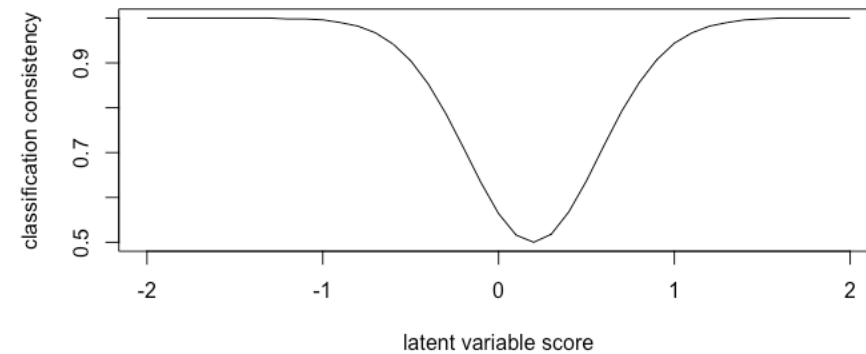
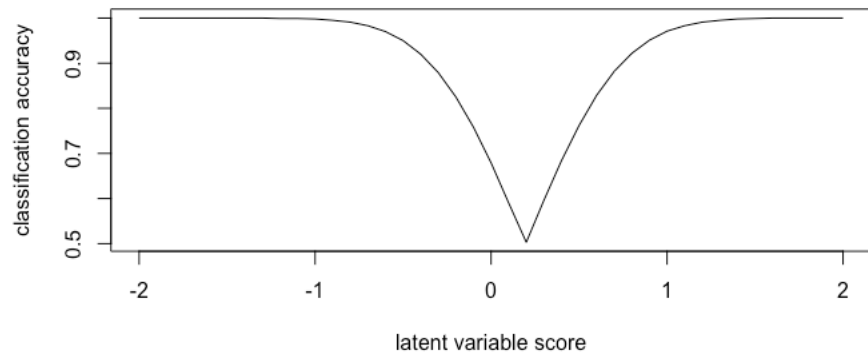
- $p1^2 + p2^2$



Gonzalez et al., 2023 *Assessment*

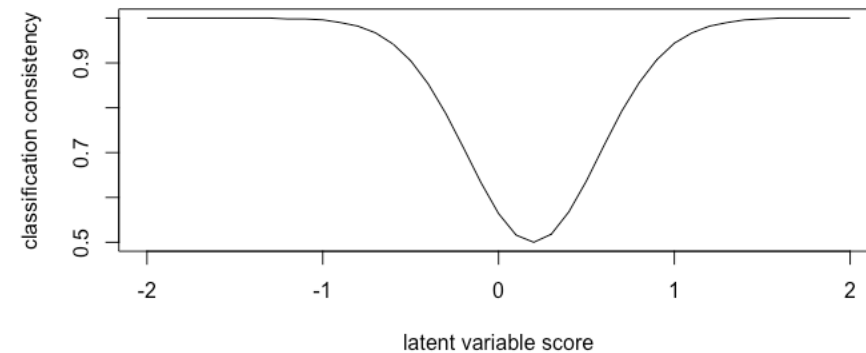
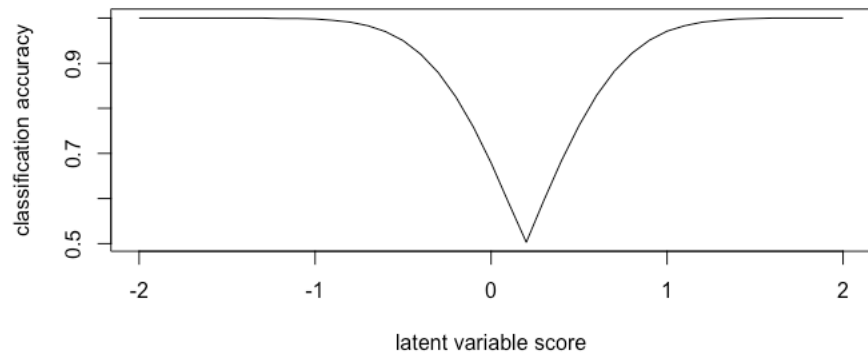
Added value: Plots

Classification accuracy and consistency of X^* as a function of η



Added value: Plots

Classification accuracy and consistency of X^* as a function of η

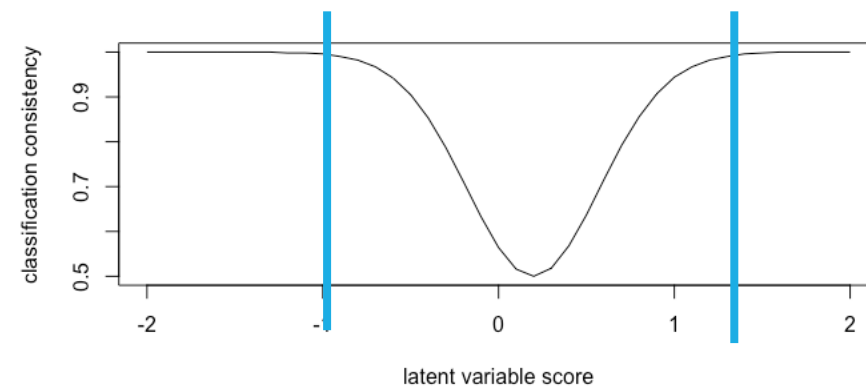
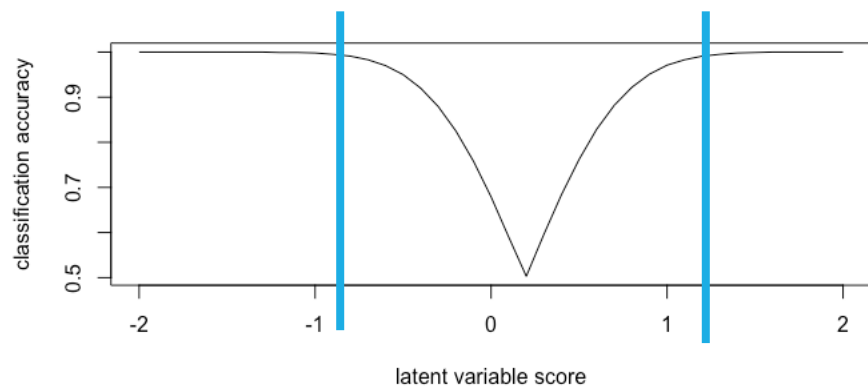


Marginal estimates: (un)weighted average across η

- same as quadrants

Added value: Plots

Classification accuracy and consistency of X^* as a function of η

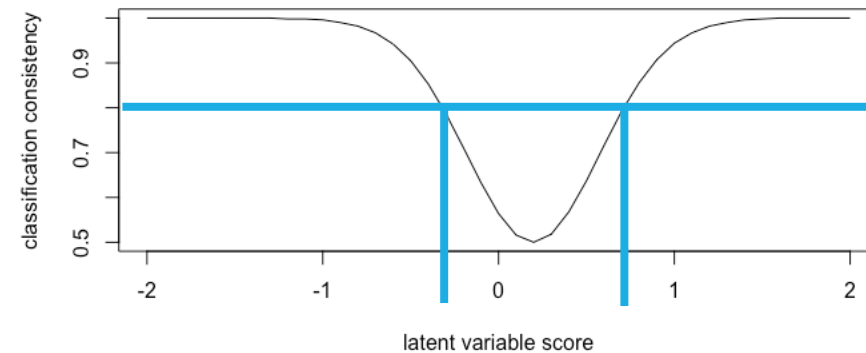
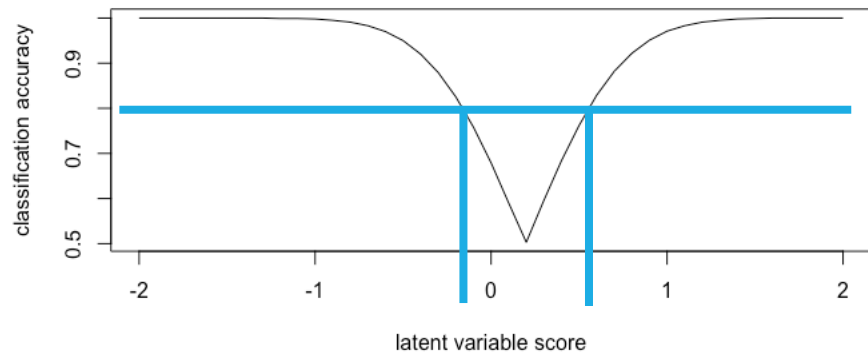


There might be regions where accuracy and consistency are not affected by measurement error...

Gonzalez et al., 2023 *Assessment*

Added value: Plots

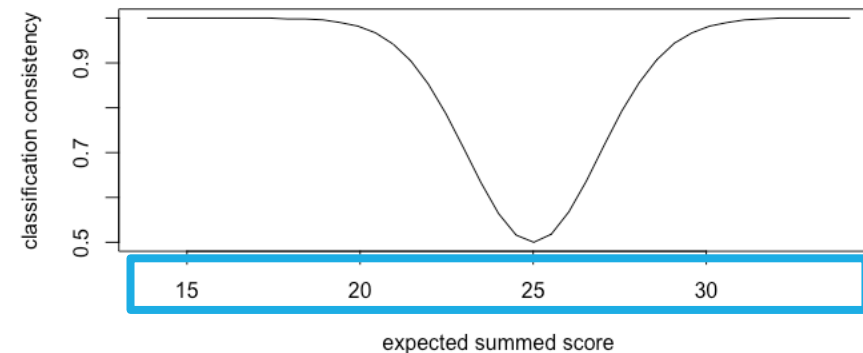
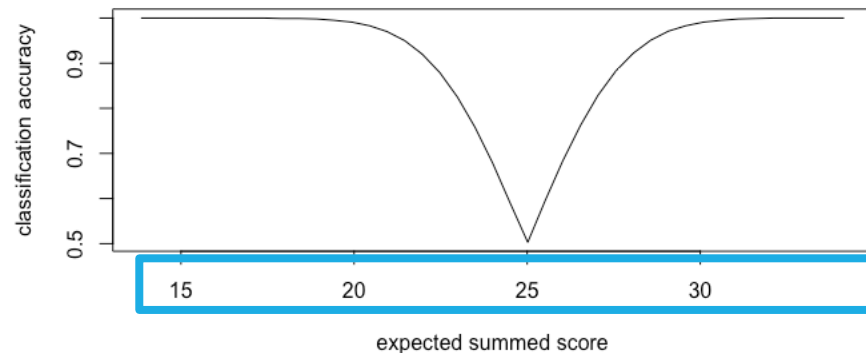
Classification accuracy and consistency of X^* as a function of η



... or where there is a minimum accuracy/consistency achieved

Added value: Plots

Classification accuracy and consistency of X^* as a function of η



... and we can transform to expected summed score for better interpretation

Example

Mindfulness Attention Awareness Scale [MAAS]

- 15-items, 6 response categories
- Possible scores = 15-90 [recoded]
- How about finding people with low mindfulness (cutpoint = -1SD, or 53)?

Eisenberg et al. (2018) data

- N=522, 50.2% female, 78.1% White
- $\chi^2(90) = 494.214, p < .001, CFI = .924, RMSEA = .079, SRMR = .042$

The Mindful Attention Awareness Scale (MAAS)

The trait MAAS is a 15-item scale designed to assess a core characteristic of mindfulness, namely, a receptive state of mind in which attention, informed by a sensitive awareness of what is occurring in the present, simply observes what is taking place.

Brown, K.W. & Ryan, R.M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology, 84*, 822-848.

Carlson, L.E. & Brown, K.W. (2005). Validation of the Mindful Attention Awareness Scale in a cancer population. *Journal of Psychosomatic Research, 58*, 29-33.

Instructions: Below is a collection of statements about your everyday experience. Using the 1-6 scale below, please indicate how frequently or infrequently you currently have each experience. Please answer according to what really reflects your experience rather than what you think your experience should be. Please treat each item separately from every other item.

	1	2	3	4	5	6
	almost always	very frequently	somewhat frequently	somewhat infrequently	very infrequently	almost never
___ 1.	I could be experiencing some emotion and not be conscious of it until some time later.					
___ 2.	I break or spill things because of carelessness, not paying attention, or thinking of something else.					
___ 3.	I find it difficult to stay focused on what's happening in the present.					
___ 4.	I tend to walk quickly to get where I'm going without paying attention to what I experience along the way.					

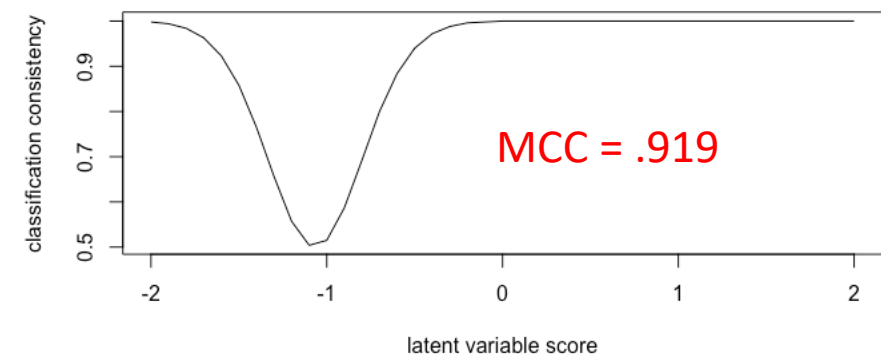
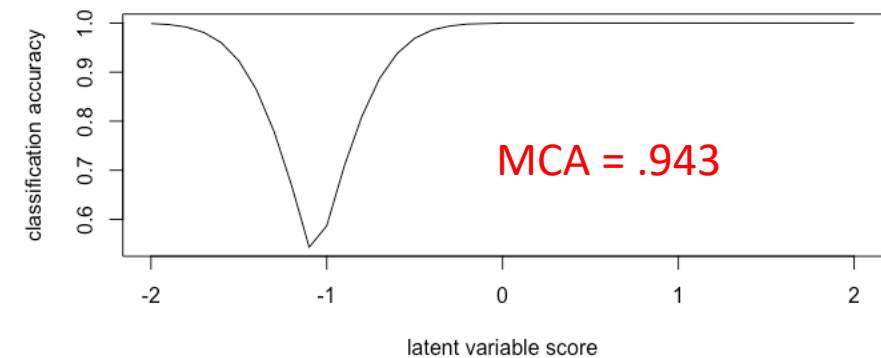
Example

Mindfulness Attention Awareness Scale [MAAS]

- 15-items, 6 response categories
- Possible scores = 15-90 [recoded]
- How about finding people with low mindfulness (cutpoint = -1SD, or 53)?

Eisenberg et al. (2018) data

- N=522, 50.2% female, 78.1% White
- $\chi^2(90) = 494.214$, $p < .001$, CFI = .924, RMSEA = .079, SRMR = .042



Uncertainty

What about confidence intervals?

Bootstrap confidence intervals

- Resample data with replacement many times
- Fit factor model
- Estimate CA and CC over resamples
- Outperform Bayesian intervals

Summary Intervals for Model-Based Classification Accuracy and Consistency Indices

Oscar Gonzalez¹ 

Abstract

When scores are used to make decisions about respondents, it is of interest to estimate classification accuracy (CA), the probability of making a correct decision, and classification consistency (CC), the probability of making the same decision across two parallel administrations of the measure. Model-based estimates of CA and CC computed from the linear factor model have been recently proposed, but parameter uncertainty of the CA and CC indices has not been investigated. This article demonstrates how to estimate percentile bootstrap confidence intervals and Bayesian credible intervals for CA and CC indices, which have the added benefit of incorporating the sampling variability of the parameters of the linear factor model to summary intervals. Results from a small simulation study suggest that percentile bootstrap confidence intervals have appropriate confidence interval coverage, although displaying a small negative bias. However, Bayesian credible intervals with diffused priors have poor interval coverage, but their coverage improves once empirical, weakly informative priors are used. The procedures are illustrated by estimating CA and CC indices from a measure used to identify individuals low on mindfulness for a hypothetical intervention, and R code is provided to facilitate the implementation of the procedures.



Uncertainty

MAAS example

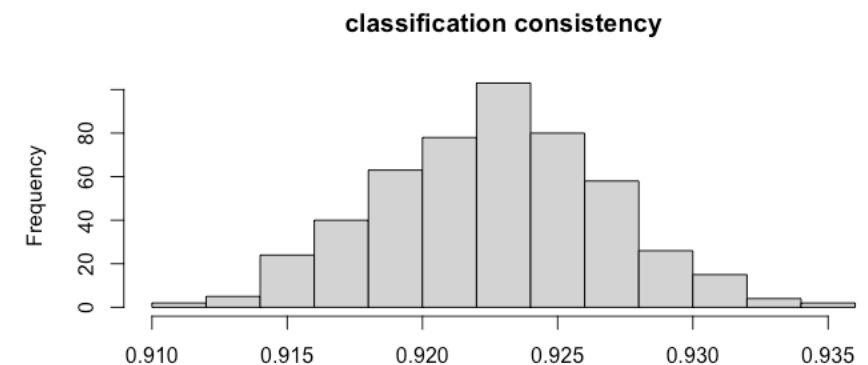
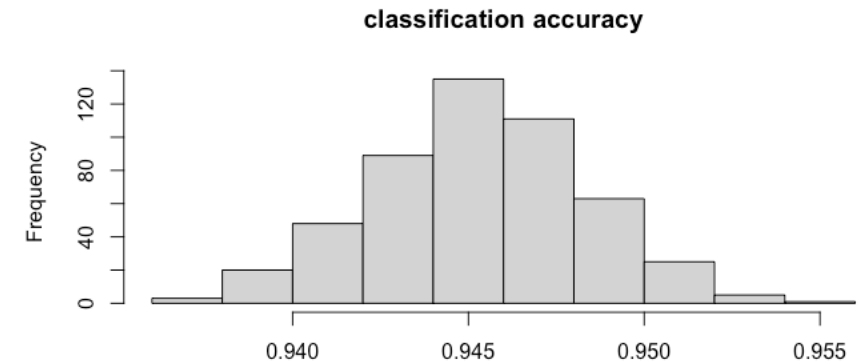
- Histograms with marginal estimates across 500 bootstrapped samples

MAAS Classification accuracy

- 95%CI [.939, .951]

MAAS Classification consistency

- 95%CI [.914, .931]

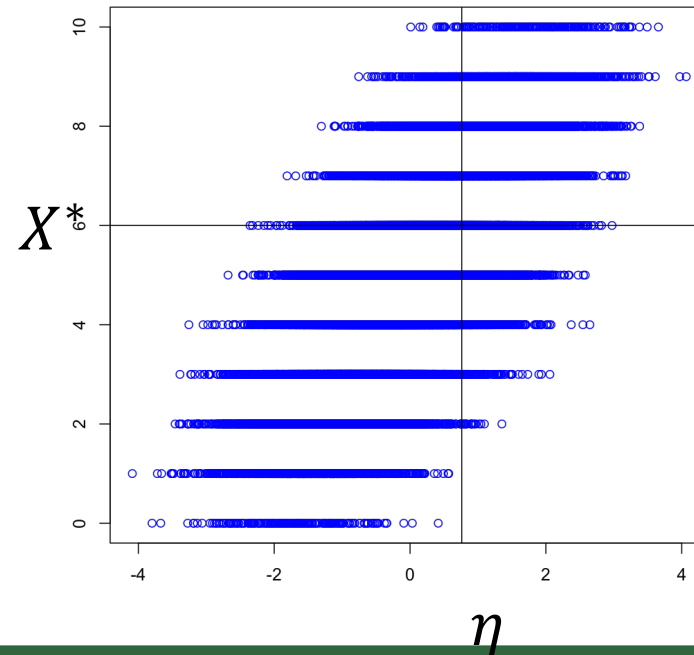
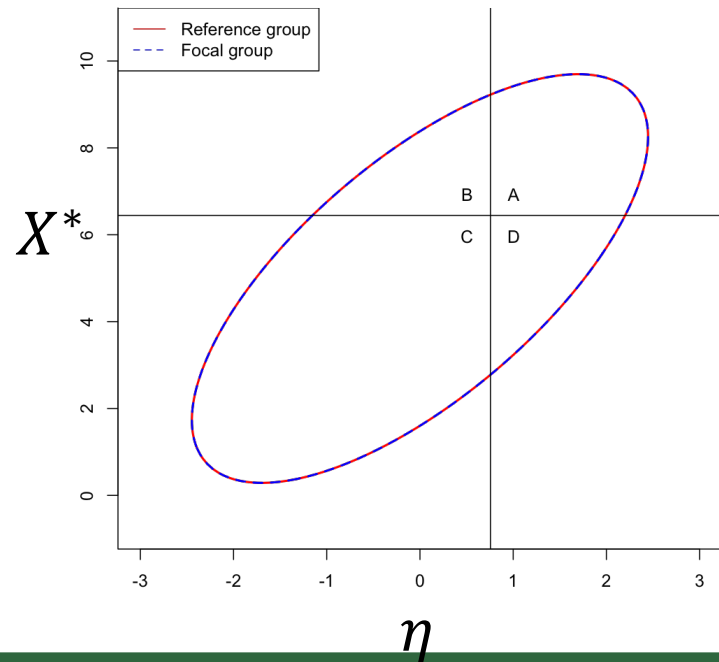


Gonzalez, 2023 *EPM*

Side Note: Discrete items

In short assessments with discrete items, X^* is not continuous ...

- Approximate relations via simulations
- We have worked out these extensions

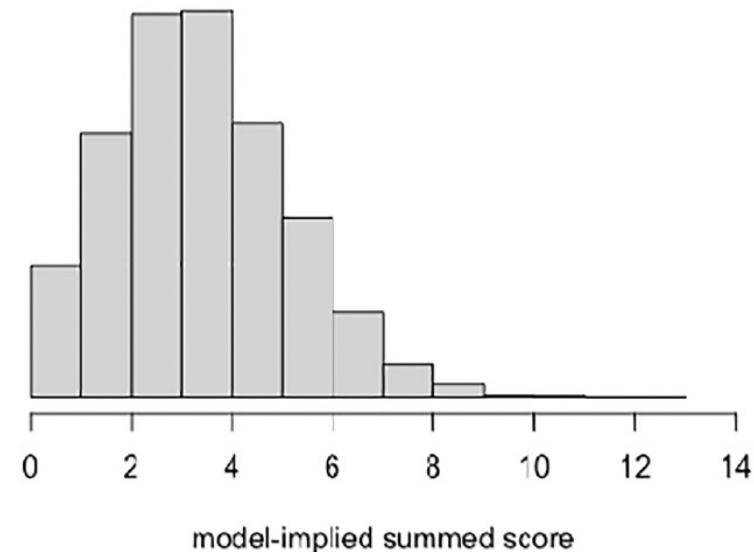
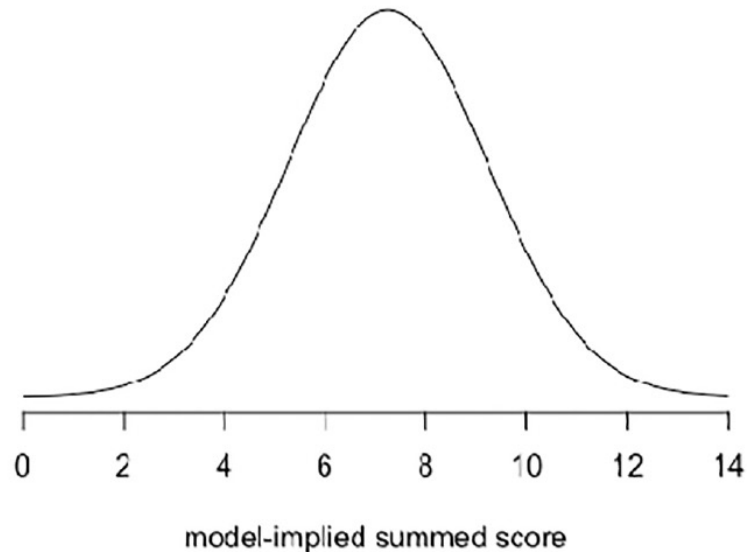


Gonzalez & Pelham, 2021 *Assessment*
Gonzalez et al., 2021 *PsychAssmt*

Side Note: Discrete items

In short assessments with discrete items, X^* is not continuous ...

- Approximate relations via simulations
- We have worked out these extensions



Gonzalez & Pelham, 2021 *Assessment*
Gonzalez et al., 2021 *PsychAssmt*

Impact of item bias on accuracy and consistency

OSCAR GONZALEZ, PHD

TEST THEORY – PSYCHOMETRICS – MEASUREMENT

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

Invariance Testing

Comparing multiple groups g to the same cut score assumes measurement invariance:

In practice, we map the relation between η and X_j in each group g ...

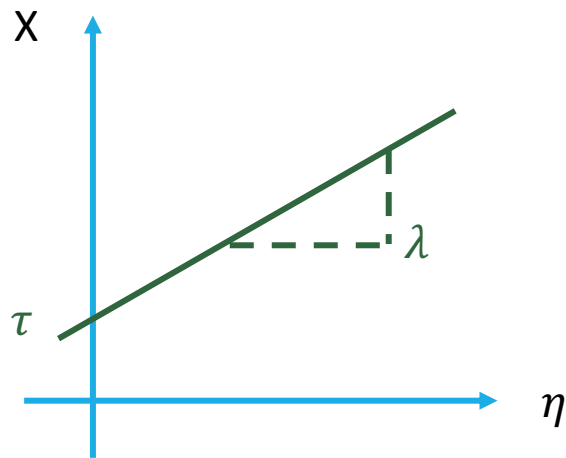
$$X_j = \tau_{jg} + \lambda_{jg}\eta + \varepsilon_j, \text{ with } \text{Var}(\varepsilon_j)_g$$

... and then we test if ...

$$\lambda_{jg} = \lambda_j, \quad \tau_{jg} = \tau_j, \quad \text{Var}(\varepsilon_j)_g = \text{Var}(\varepsilon_j)$$

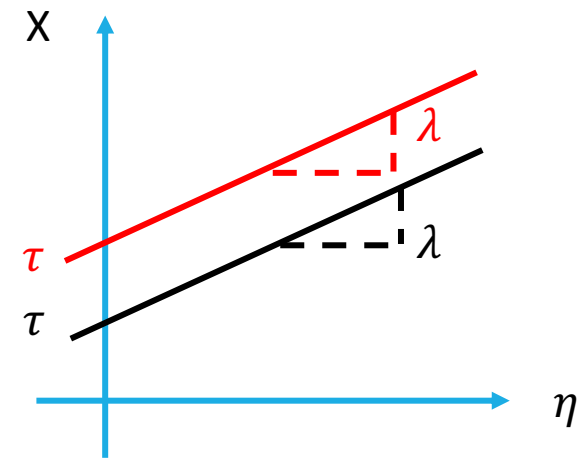
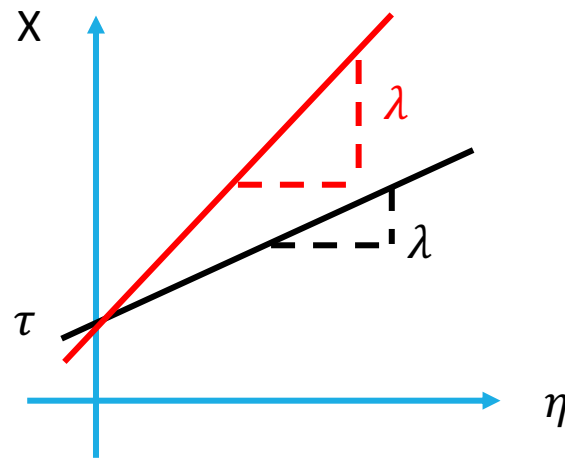
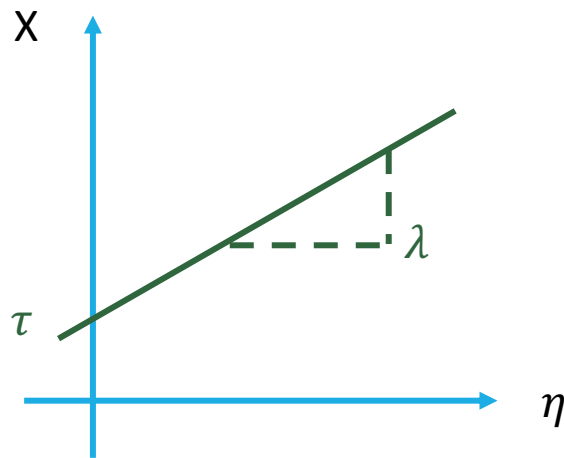
Graphically

The factor model describes the linear relation between the latent variable and the item



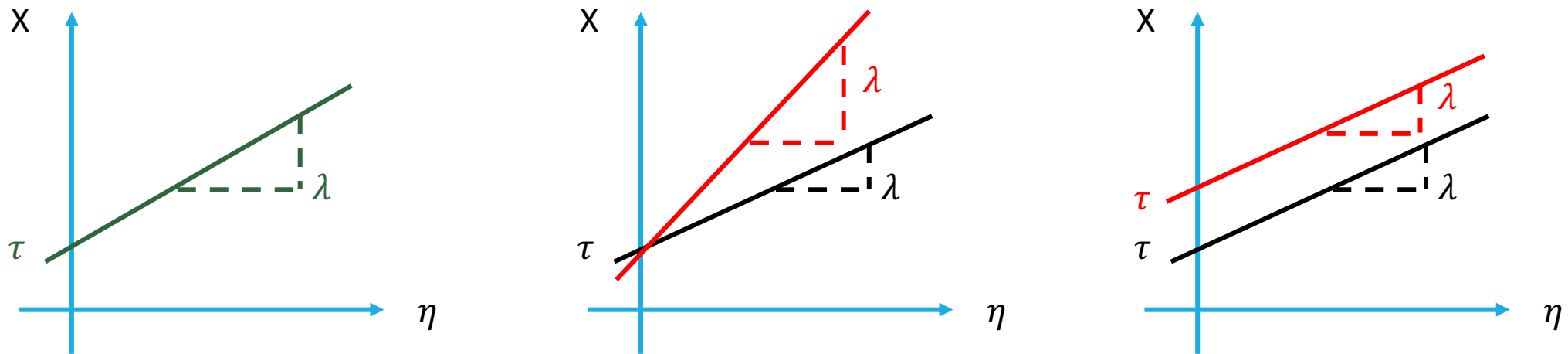
Graphically

The factor model describes the linear relation between the latent variable and the item



Graphically

The factor model describes the linear relation between the latent variable and the item



At the same level of the latent variable, we might have different observed scores... and this may systematically place individuals above/below the cutpoint!

When Does Differential Item Functioning Matter for Screening? A Method for Empirical Evaluation

Assessment

1–11

© The Author(s) 2020

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/1073191120913618

journals.sagepub.com/home/asm



Oscar Gonzalez¹  and William E. Pelham III²

Abstract

When items in a screening measure exhibit differential item functioning (DIF) across groups (e.g., males vs. females), DIF might affect which individuals are “caught” in the screening. This phenomenon is common, but DIF detection procedures do not typically provide guidance on whether the presence of DIF will meaningfully affect screening accuracy. Millsap and Kwok proposed a method to quantify the impact of DIF on screening accuracy, but their approach had limitations that prevent its use in scenarios where items are discrete. We extend the Millsap and Kwok procedure to accommodate discrete items and provide *R* functions to apply the procedure to the user’s own data. We illustrate our approach using published screening information and evaluate the proposed methodology with a small simulation study. Overall, we encourage researchers to use empirical methods to evaluate the extent to which the presence of DIF in a screening measure materially affects screening performance.

Estimating Classification Consistency of Screening Measures and Quantifying the Impact of Measurement Bias

Oscar Gonzalez¹, A. R. Georgeson¹, William E. Pelham III², and Rachel T. Fouladi³

¹ University of North Carolina at Chapel Hill

² University of California, San Diego

³ Simon Fraser University

Screening measures are used in psychology and medicine to identify respondents who are high or low on a construct. Based on the screening, the evaluator assigns respondents into classes corresponding to different next steps: make a diagnosis versus reject a diagnosis; provide services versus withhold services; and conduct further assessment versus conclude the assessment process. When measures are used to classify individuals, it is important that the decisions be consistent and equitable across groups. Ideally, if respondents completed the screening measure repeatedly in quick succession, they would be consistently assigned into the same class each time. In addition, the consistency of the classification should be unrelated to the respondents' background characteristics, such as sex, race, or ethnicity (i.e., the measure is free of measurement bias). Reporting estimates of classification consistency is a common practice in educational testing, but there has been limited application of these estimates to screening in psychology and medicine. In

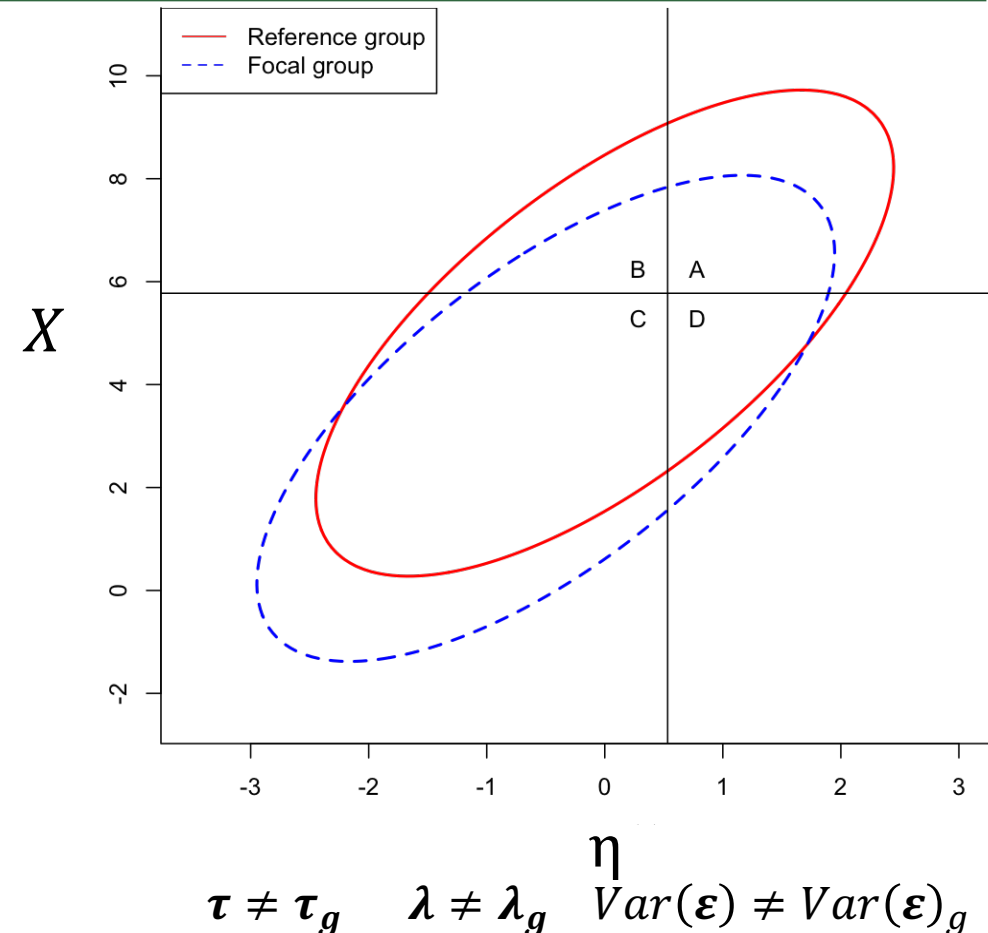
Noninvariance and Classification

I've shown that we can determine classification accuracy and consistency from model parameters

- If model parameters are different across groups... then it might affect the estimates of accuracy and consistency

Our procedure entails looking at Δ accuracy and Δ consistency per group vs pooled

Gonzalez & Pelham, 2021 *Assessment*
Gonzalez et al., 2021 *PsychAssmt*



Noninvariance and Classification

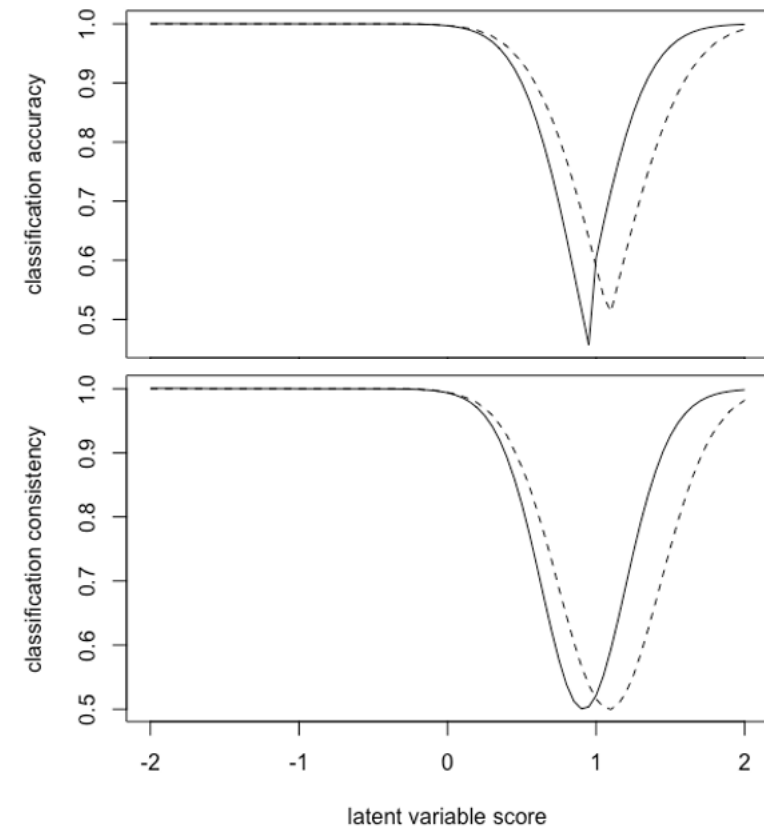
I've shown that we can determine classification accuracy and consistency from model parameters

- If model parameters are different across groups... then it might affect the estimates of accuracy and consistency

Our procedure entails looking at Δ accuracy and Δ consistency per group vs pooled

Gonzalez & Pelham, 2021 *Assessment*

Gonzalez et al., 2021 *PsychAssmt*



$$\tau \neq \tau_g \quad \lambda \neq \lambda_g \quad \text{Var}(\epsilon) \neq \text{Var}(\epsilon)_g$$

Example

CES-D data

- Screener for depression, self-reported symptoms in past week
- 0 to 3, screen if score > 16

Wallis (2013)

- N = 845, 35.1% male

Invariance analysis

- Detection method: linear regression
- 8 items were noninvariant wrt sex

Blue:
High intercept
& slope for males

Pink:
High intercept
& slope for females

1. I was bothered by things that usually don't bother me.
2. I did not feel like eating; my appetite was poor.
3. I felt that I could not shake off the blues even with help from my family or friends.
4. I felt I was just as good as other people.
5. I had trouble keeping my mind on what I was doing.
6. I felt depressed.
7. I felt that everything I did was an effort.
8. I felt hopeful about the future.
9. I thought my life had been a failure.
10. I felt fearful.
11. My sleep was restless.
12. I was happy.
13. I talked less than usual.
14. I felt lonely.
15. People were unfriendly.
16. I enjoyed life.
17. I had crying spells.
18. I felt sad.
19. I felt that people dislike me.
20. I could not get "going."

Example

Model Fit

- CFI = .851, RMSEA = .083, SRMR = .064

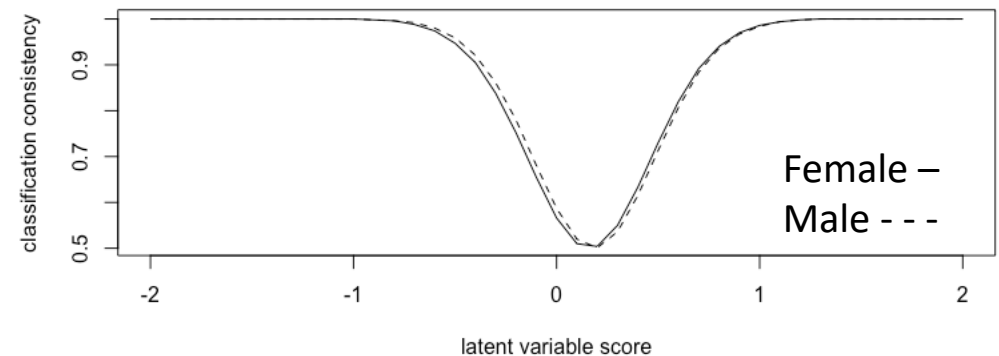
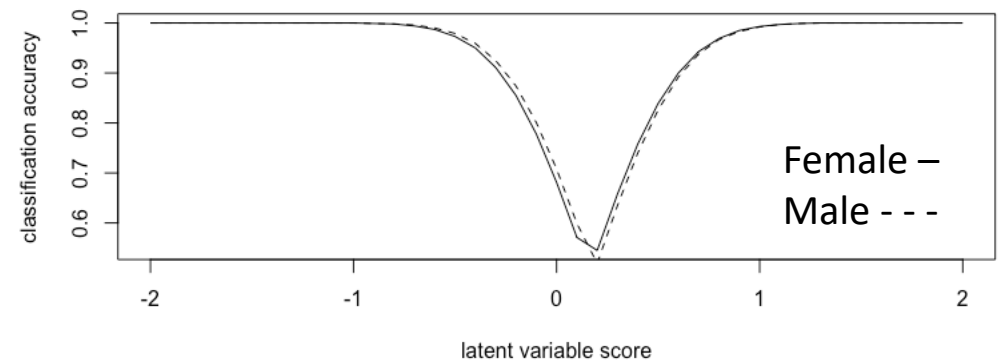
Marginal classification accuracy:

- Females: .892
- Males: .894

Marginal classification consistency:

- Females: .849
- Males: .852

Relatively close, but the conditional estimates vary a bit



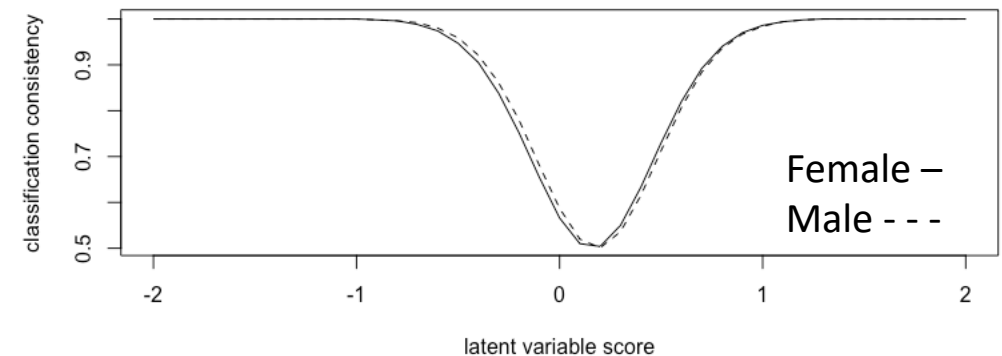
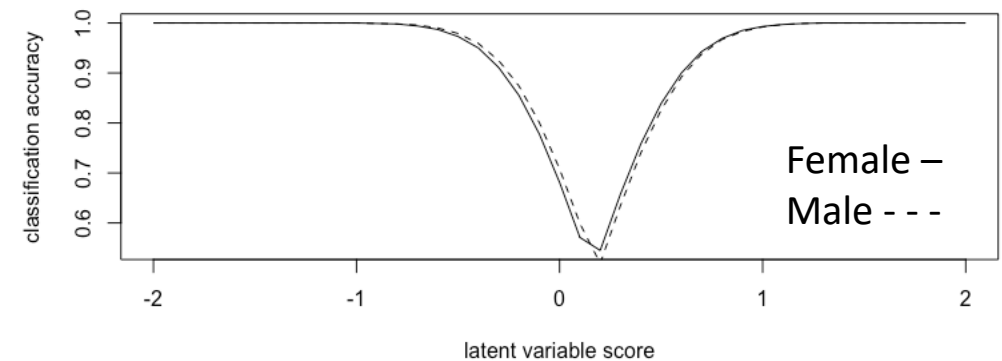
Conclusions

Quantify how item bias impacts the screening procedure

- Beyond just identifying bad items

Empowers clinicians to decide if the measure is still fit for purpose

- If we are making decisions based on the measure, how are groups impacted?
- Make process more **equitable** and **transparent**



Extensions and Future Directions

OSCAR GONZALEZ, PHD

TEST THEORY – PSYCHOMETRICS – MEASUREMENT

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL



Extensions

Applications

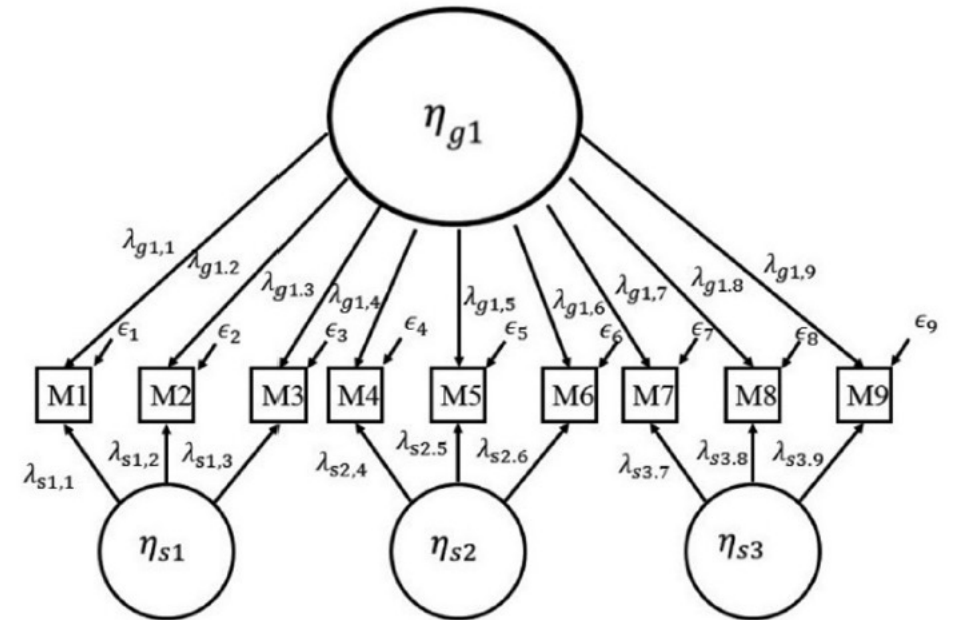
- Consulting in an R21 on substance use screening

Current methods rely on meeting factor model assumptions...

- What if we don't?

Handling other scenarios

- Multidimensionality (bifactor, corr residuals)
- DIF with continuous variables
- Other scores



Machine Learning for Screening

Predict diagnosis from items

- *Train* model to learn patterns
- *Evaluate* performance in a test data

Comparable accuracy to psychometrics when items are highly related to diagnosis

- Might outperform psychometrics when items interact and have nonlinear relations



© 2020 American Psychological Association
ISSN: 1082-989X

Psychological Methods

<http://dx.doi.org/10.1037/met0000317>

Psychometric and Machine Learning Approaches for Diagnostic Assessment and Tests of Individual Classification

Oscar Gonzalez
University of North Carolina at Chapel Hill

Abstract

Assessments are commonly used to make a decision about an individual, such as grade placement, treatment assignment, job selection, or to inform a diagnosis. A psychometric approach to classify respondents based on the assessment would aggregate items into a score, and then each respondent's score is compared to a cut score. In contrast, a machine learning approach to classify respondents would build a model to predict the probability of belonging to a specific class from assessment items, and then respondents are classified based on their predicted probability of belonging to that class. It remains unclear whether psychometric and machine learning methods have comparable classification accuracy or if 1 method is preferable in all or some situations. In the context of diagnostic assessment, this study used Monte Carlo simulation methods to compare the classification accuracy of psychometric and machine learning methods as a function of the diagnosis-test correlation, prevalence, sample size, and the structure of the diagnostic assessment. Results suggest that machine learning models using logistic regression or random forest could have comparable classification accuracy to the psychometric methods using estimated item response theory scores. Therefore, machine learning models could provide a viable alternative for classification when psychometric methods are not feasible. Methods are illustrated with an empirical example predicting an oppositional defiant disorder diagnosis from a behavior disorders scale in children of age seven. Strengths and limitations for each of the methods are examined, and the overlap between the field of machine learning and psychometrics is discussed.

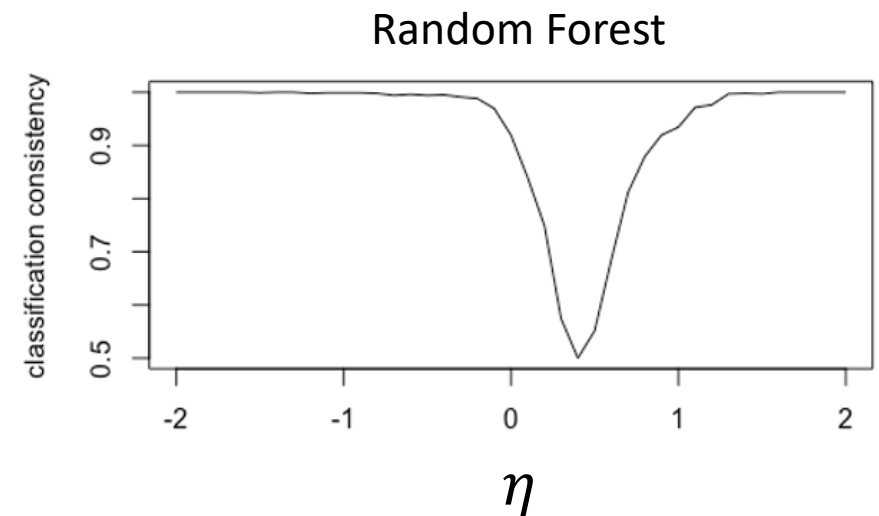
Measurement and Machine Learning

Intersection of machine learning and psychometrics for assessment

- Prediction models with item responses as predictors
- Develop measures that are reliable and that predict well

Recently –classification consistency for machine learning models

- Accepted at *PsychAssessment*



Measurement and Machine Learning

Intersection of machine learning and psychometrics for assessment

- Prediction models with item responses as predictors
- Develop measures that are reliable and that predict well

Recently –classification consistency for machine learning models

- Accepted at *PsychAssessment*



© 2024 American Psychological Association
ISSN: 1040-3590

Psychological Assessment

<https://doi.org/10.1037/pas0001313>

Estimating Classification Consistency of Machine Learning Models With Screening Measures

Oscar Gonzalez¹, A. R. Georgeson², and William E. Pelham III³

¹ Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill

² Department of Psychology, Arizona State University

³ Department of Psychiatry, University of California San Diego

This article illustrates novel quantitative methods to estimate classification consistency in machine learning models used for screening measures. Screening measures are used in psychology and medicine to classify individuals into diagnostic classifications. In addition to achieving high accuracy, it is ideal for the screening process to have high classification consistency, which means that respondents would be classified into the same group every time if the assessment was repeated. Although machine learning models are increasingly being used to predict a screening classification based on individual item responses, methods to describe the classification consistency of machine learning models have not yet been developed. This article addresses this gap by describing methods to estimate classification inconsistency in machine learning models arising from two different sources: sampling error during model fitting and measurement error in the item responses. These methods use data resampling techniques such as the bootstrap and Monte Carlo sampling. These methods are illustrated using three empirical examples predicting a health condition/diagnosis from item responses. R code is provided to facilitate the implementation of the methods. This article highlights the importance of considering classification consistency alongside accuracy when studying screening measures and provides the tools and guidance necessary for applied researchers to obtain classification consistency indices in their machine learning research on diagnostic assessments.

Overall

My research heavily emphasizes...

- How we select individuals
- For whom interventions work

Better methods can help us achieve health equity

- Make accurate and consistent decisions across groups
- Determine if measures are too biased to be used

