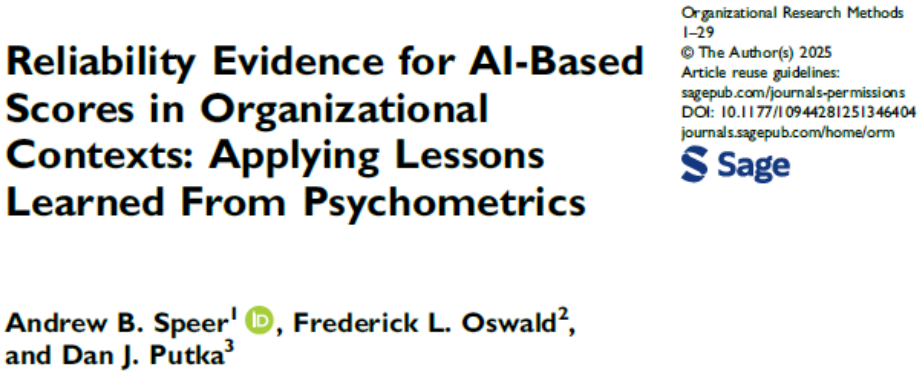
Psychometrics of AI-Scores

Andrew Speer
Indiana University
February 2026

Today's Agenda

- Discuss how psychometric properties can be established for AI scoring
- Based on two recent publications and a series of current projects



Speer, Oswald, & Putka (in press). **Establishing Reliability and Validity Evidence for AI-Based Scores: Application to the Organizational Sciences**. In *AI for I-O psychologists: Research and applications* (Eds: Thompson, Yankov, & Hernandez)

The Context

Rapidly Changing AI/ML Landscape (in IO/OB/HR)

1st Major ML Applications IO/OB/HR

What most people thought
- What is ML?
- ML is overfitted slop

Methods
- Varied ML algos for tabular data
- NLP relied upon simple bag of words

Major Use
- HR prediction and psychological scoring

Maturing Field and Advanced NLP

What most people thought
- ML is sexy!
- Lots of potential, particularly NLP
- Only handful of IO/OB/HR experts

Methods
- Varied ML algos for tabular data
- Transformer neural networks for NLP

Major Use
- Prediction and psychological scoring (more broadly)

Adoption of Artificial Intelligence

What most people think
- AI is a disruptive technology
- Massive automation potential
- Much more accessible for people

Methods
- Primarily zero-shot LLMs
- Traditional ML still useful

Major Uses
- Everyday applications in life
- Diverse research applications
- Prediction and psychological scoring

~10 years ago ⟷ ~ Now

KELLEY SCHOOL OF BUSINESS

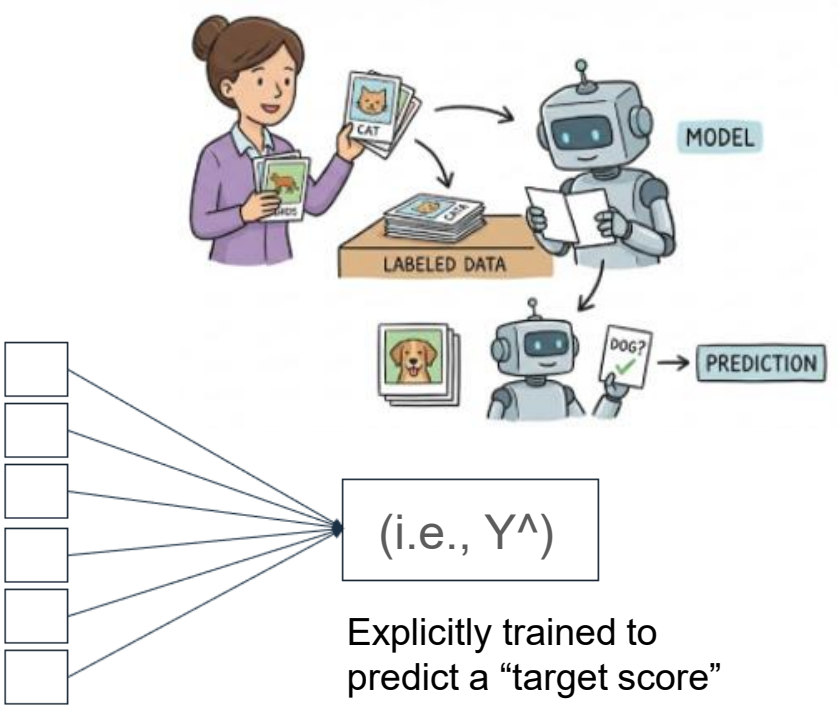
Prediction and Psychological Scoring: What does this mean?

- **Our Focus** is when AI/ML is used to produce scores that represent some construct (e.g., social effectiveness) or predict some behavior/event (e.g., risk of leaving job/organization), via either supervised ML or AI systems (e.g., LLMs)

- Examples
  - Scoring personality from chat bot interactions or text (e.g., Fan et al., 2023; Speer et al, in press)
  - Scoring interview performance (e.g., Hickman et al., 2022; Liff et al., 2024)
  - Scoring assessment center centers to reflect dimensions (e.g., Yankov & Speer, 2023)
  - Predicting employee turnover from HRIS predictors (e.g., Sajjadiani et al., 2019)
  - Predicting employee job performance from assessment data (e.g., Affourtit et al., 2022)
  - And so on…

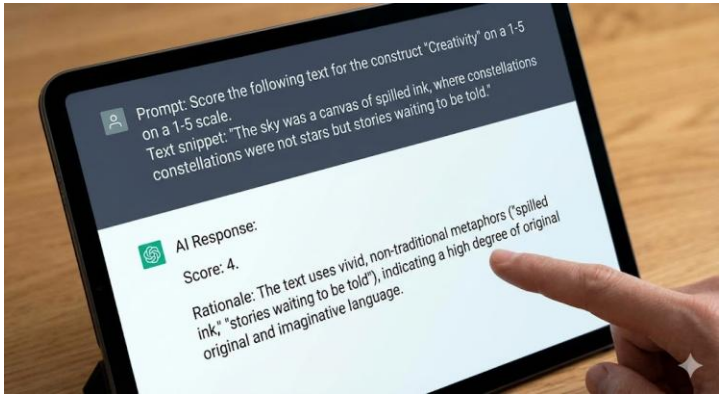
Prediction and Psychological Scoring: What does this mean?

Supervised Modeling



Explicitly trained to predict a “target score”

Zero-shot/In-Context Scoring



Need for Thoughtful Consideration of AI/ML Psychometrics

- For traditional assessments (e.g., surveys), there is literature and professional standards for establishing psychometric evidence in support of score use (reliability, validity)…

Challenges Applied to AI/ML

- 1. Messiness: AI/ML scoring is based on complex and unstructured data with “black box” algorithms
- 2. Inconsistency when interpreting the same data (what is reliability versus validity)
- 3. AI-specific sources of scoring inconsistency (e.g., instructions/prompts, hyper-parameters used, when the algorithm is run)….

Psychometric Overview

From Broad Considerations to More Nuanced

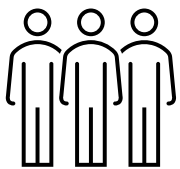
What Do We Mean by "Reliability"

- Per classical test theory (CTT), **reliability** is the proportion of total variance in scores that is attributed to true score variance (Bollen, 1989; Nunnally & Berstein, 1994)

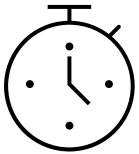
$$r_{XX} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_X - \sigma^2_e}{\sigma^2_X}$$

- Reliability is the consistency in scores across replications of a measurement procedure (Putka & Sackett, 2010):

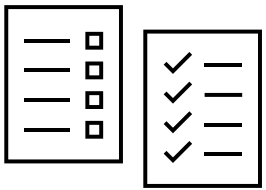
Across Raters



Across Measurement Occasions



Across Content (items/forms)

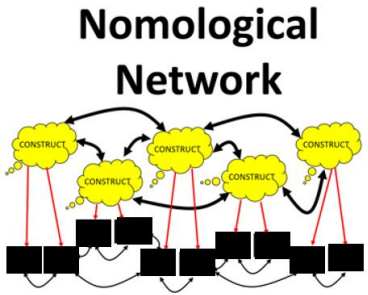


What Do We Mean by “Validity”

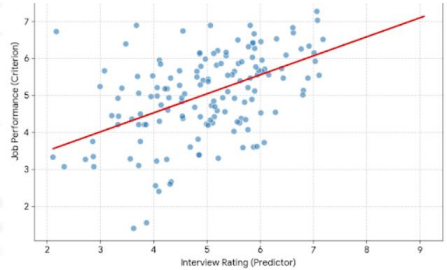
- Validity: degree to which evidence and theory support the interpretations of scores
- Per unitarian framework (Binning & Barrett, 1989), validity is supported via multiple sources of information

Content Information

Convergent & Discriminant Correlations

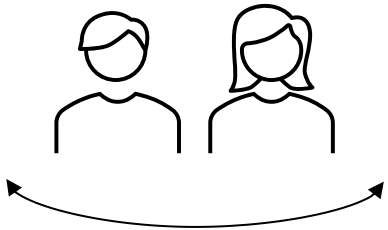


Prediction of Criterion-Related Variables

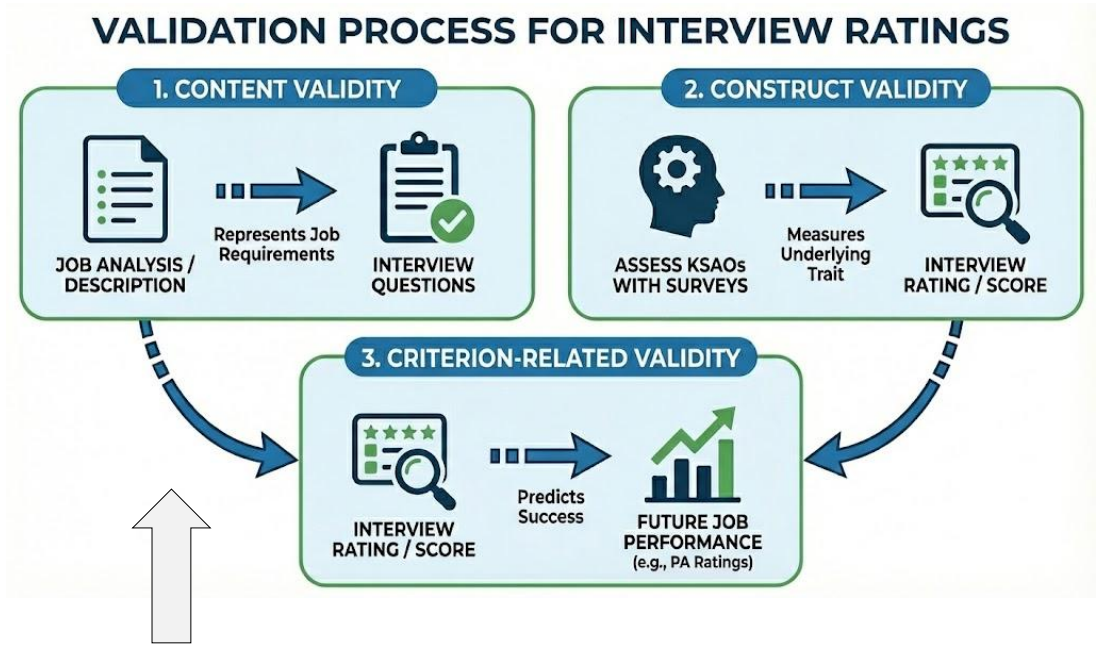


Reliability and Validity in Traditional Context

Two Interviewers Assess Job Applicants

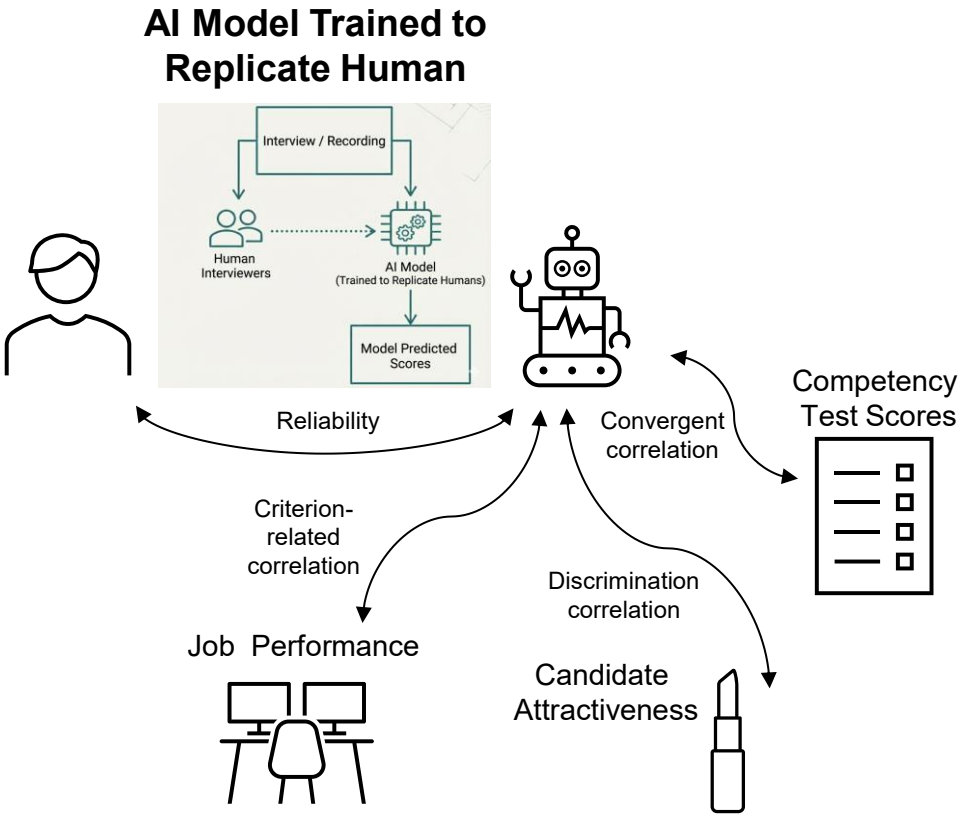


Their ratings are strongly correlated = high inter-rater reliability evidence

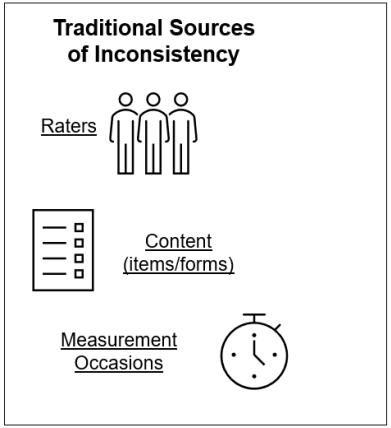


However, what exactly does that reliable variance represent?

Reliability and Validity in AI/ML Context



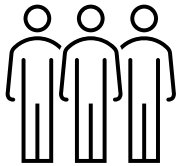
Traditional Sources of Inconsistency with AI/ML

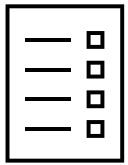


Sources of Inconsistency

Traditional Sources of Inconsistency

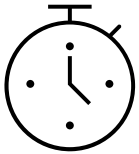
Raters





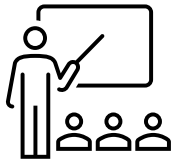
Content (items/forms)

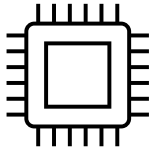
Measurement Occasions



AI-Specific Sources of Inconsistency

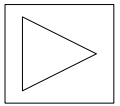
Instructions / Prompts

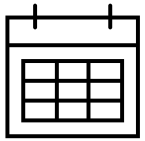




Algorithms

Hyper-parameters





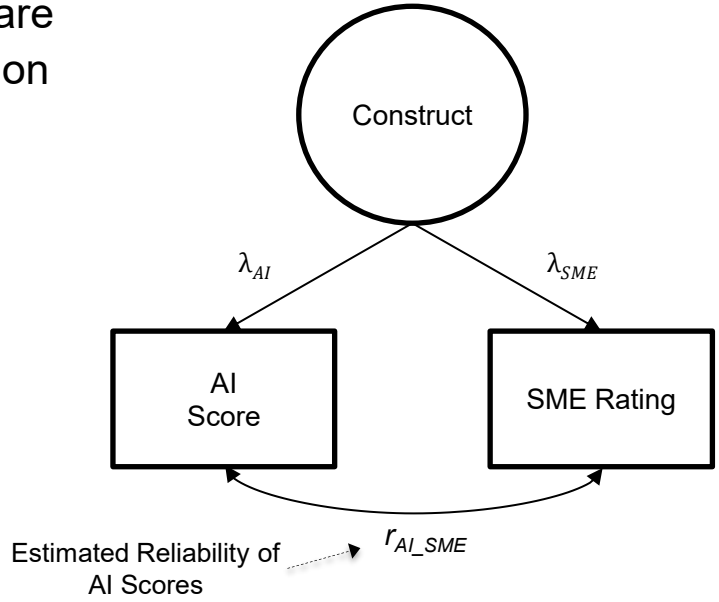
Run Occasion

Inter-Rater Reliability for AI Scores

- AI/ML is used to "replicate" human raters (i.e., SMEs)
- In a common design, AI-scores and SME ratings are correlated… If measures are parallel, the correlation between the two estimates reliability (for either)
  - Same construct (congeneric)
  - Same true score and error variance

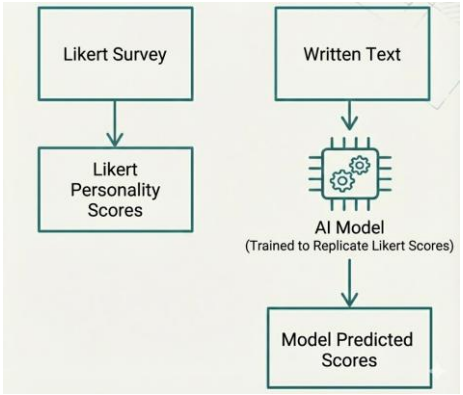
$\lambda_{AI}^2 = r_{xx}$
$\lambda_{AI} * \lambda_{SME} = r_{AI\_SME}$
When $\lambda_{AI} = \lambda_{SME}$, $r_{AI\_SME} = r_{xx}$
When $\lambda_{AI} \neq \lambda_{SME}$, $r_{AI\_SME} \neq r_{xx}$

How reasonable is it to assume parallel measures (same constructs, $\lambda_{AI} = \lambda_{SME}$), and what are the implications if measures are not parallel?



Parallel Assumption: Same Construct, True Score, and Error Variance

AI Model Trained to Replicate Likert Personality



- Proposition: Target scores and AI scores are more likely to measure the same construct when they are both formed based on the same input data

Source of Supervised ML	External Target *r*	SME Target *r*
Speer et al. (in press)	.51	.83
Speer et al. (2024)	.65	.87

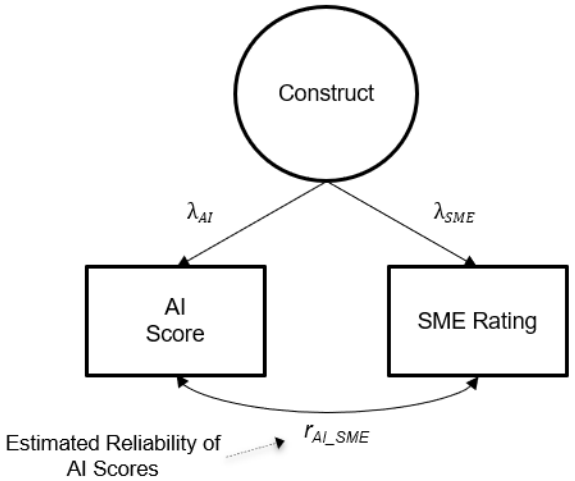
NEO-PI-R Facets (Example Items)

1. Achievement (I work hard)
2. Self-Discipline (I get to work at once)
3. Self-Efficacy (I excel in what I do)
4. Orderliness (I like to tidy up)
5. Dutifulness (I follow the rules)
6. Cautious (I don't make rash decisions)

Text Prompt: Tell me about a time you had to be very organized.

Parallel Assumption: Same Construct, True Score, and Error Variance

- If AI/ML scores have more/less error than target SME ratings, the correlation between measures will not equal reliability



$\lambda_{AI}^2 = r_{xx}$
$\lambda_{AI} * \lambda_{SME} = r_{AI\_SME}$
When $\lambda_{AI} = \lambda_{SME}$, $r_{AI\_SME} = r_{xx}$
When $\lambda_{AI} \neq \lambda_{SME}$, $r_{AI\_SME} \neq r_{xx}$

Simulation: 1000 repeated runs where reliability of AI & SME ratings were varied

True AI r_{xx}	True SME r_{xx}	r_{AI_SME}
.4	.4	.40 =
.6	.4	.49 ↓
.8	.4	.57 ↓
.4	.6	.49 ↑
.6	.6	.60 =
.8	.6	.69 ↓
.4	.8	.57 ↑
.6	.8	.69 ↑
.8	.8	.80 =

Parallel Assumption: Same Construct, True Score, and Error Variance

When might AI scores and SME ratings be more likely to be parallel

- Supervised learning where AI/ML model matches complexity of target scores (e.g., large transformer models) and AI and target scores based on same predictor inputs
  - Thus, most supervised AI/ML projects to replicate SME ratings can use correlational designs to estimate inter-rater reliability for AI scores

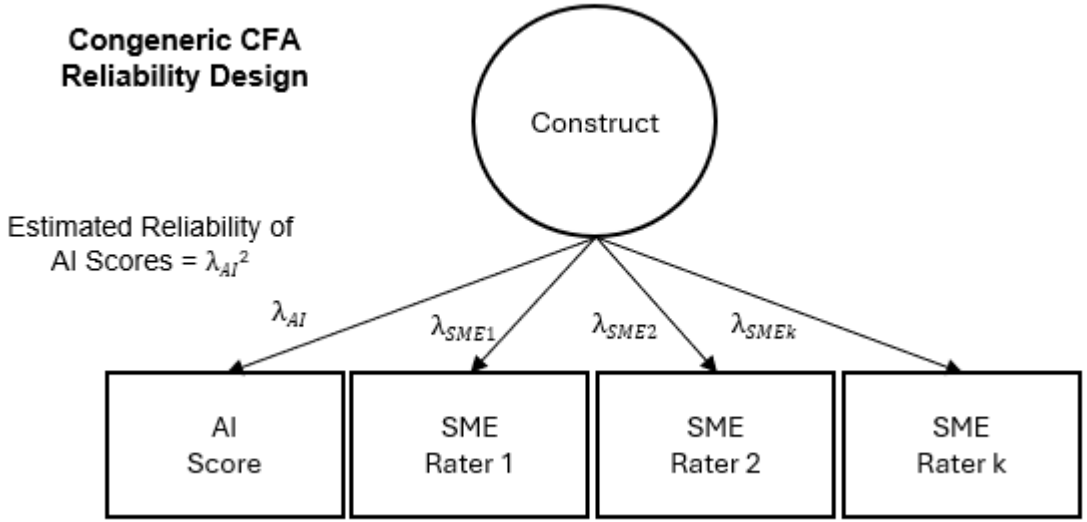
When might AI scores and SME ratings be less likely to be parallel

- Supervised learning where AI/model and target scores are not based on the same input data
- Supervised learning where AI/ML model cannot capture the complexity reflected in target scores (e.g., bag of words)
- Non-supervised scoring via LLMs…
  - Unlike supervised ML, we have no idea how reliable LLMs might be relative to a single human SME or a composite SME score…

Establishing Inter-Rater Reliability for LLMs

So How Should We Establish Inter-Rater Reliability for LLM Scores?

- Confirmatory Factor Analysis (allows for congeneric estimation) with >= 2 SMEs

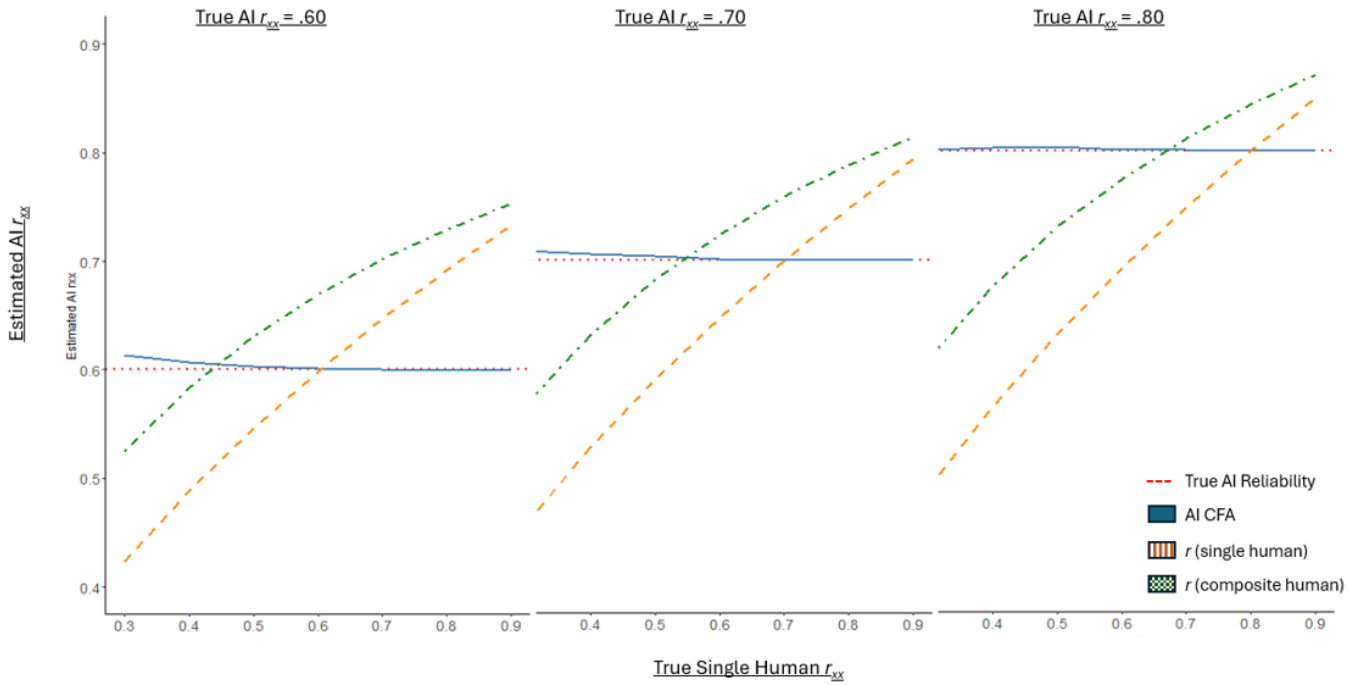


Speer (under review)

Establishing Inter-Rater Reliability for LLMs

So How Should We Establish Inter-Rater Reliability for LLM Scores?

- Large scale simulation of estimating AI reliability using CFA or correlational designs



Speer (under review)

Establishing Inter-Rater Reliability for LLMs

- CFA design superior to correlational approaches across (nearly) all conditions
- Yet, resources are needed to achieve adequately accurate r_{xx} estimates

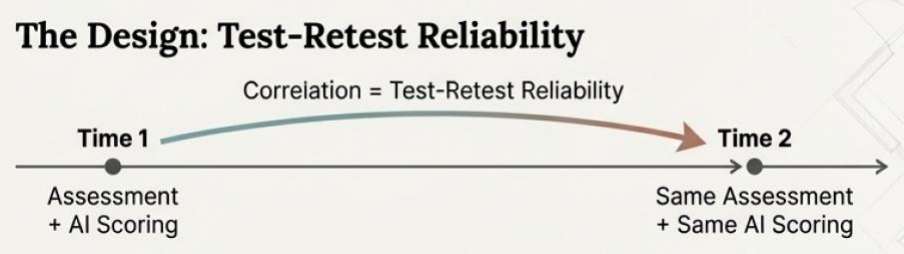
True AI $r_{xx} > .50$					True AI $r_{xx} <= .50$				
True Single SME r_{xx}	N	Abs r_{xx} Diff: |actual-observed|			True Single SME r_{xx}	N	Abs r_{xx} Diff: |actual-observed|		
		r 1 rater	CFA 2 raters	CFA 4 raters			r 1 rater	CFA 2 raters	CFA 4 raters
<= 0.50	50	0.23	0.15	0.10	<= 0.50	50	0.11	0.17	0.12
	100	0.22	0.11	0.07		100	0.08	0.11	0.08
	200	0.21	0.08	0.05		200	0.07	0.08	0.06
	400	0.21	0.06	0.03		400	0.06	0.05	0.04
	600	0.21	0.05	0.03		600	0.05	0.04	0.03
	800	0.21	0.04	0.02		800	0.05	0.04	0.03
	1000	0.21	0.04	0.02		1000	0.05	0.03	0.03
> .50	50	0.09	0.08	0.06	> .50	50	0.14	0.10	0.09
	100	0.07	0.05	0.04		100	0.14	0.07	0.06
	200	0.07	0.04	0.03		200	0.14	0.05	0.05
	400	0.07	0.03	0.02		400	0.14	0.03	0.03
	600	0.07	0.02	0.02		600	0.14	0.03	0.03
	800	0.07	0.02	0.01		800	0.14	0.02	0.02
	1000	0.07	0.02	0.01		1000	0.14	0.02	0.02

Showing aggregated conditions

Speer (under review)

Test-Retest Reliability: Inconsistency Over Time

- Transient error: fluctuations due to temporary factors (e.g., mood, environment)
- Interested in generalizing scores beyond specific measurement occasion

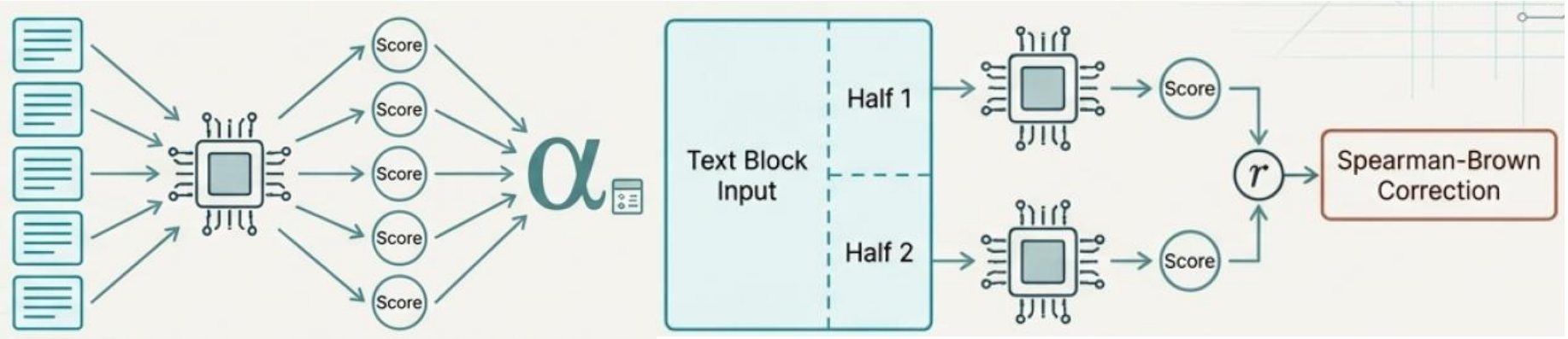


Examples: interview responses (Hickman et al., 2022; Liff et al., 2024), employee attitudes (Speer et al., 2023), chatbots (Fan et al., 2023), assessment center performance (Yankov & Speer, 2023)

- Concerns to be aware of:
  - Construct should theoretically remain stable over time
  - Requires parallel test assumption
    - Re-testing may change behavior, especially for AI-scored assessments (e.g., writing an essay twice)

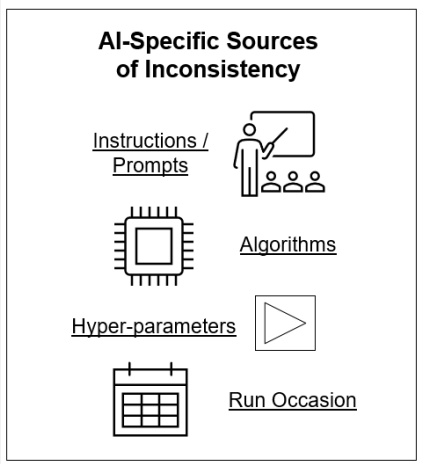
Composite Reliability: Inconsistency over Content

- Form-specific errors occur due to construct-unrelated inconsistencies between different sets of items or predictor inputs (e.g., sets of items or assessment stimuli).



- Concerns to be aware of:
  - Unlike traditional measures, may be large number of diverse and unstructured inputs. Thus, traditional assumptions (e.g., congeneric) may be hard to meet
  - May be less value in generalizing over content (e.g., thousands of predictors to score perf), and more likely when predictor data are in distinct subsets (e.g., interview questions, Hickman et al., 2024)

New Sources of Inconsistency with AI/ML



Imagine the Following Scenario…

- You use an LLM to score construct X based on applicant resumes
  - How different would the ratings be when using other prompts to measure construct (instruction/ prompt specific-error)?
  - How consistent would ratings be if run using the same prompt but another LLM (algorithm-specific error)?
  - How consistent would ratings be if the same prompt and LLM were used but with different hyper-parameters (hyper-parameter-specific error)
  - How consistent would ratings be if the same prompt, LLM, and hyper-parameters were used but on a different occasion (algorithm run occasion-specific error)

AI-Specific Inconsistency: Lots of Potential Influences

- Algorithm:
  - Family (e.g., GPT, Claude, Gemini)
  - Size (i.e., number parameters)
  - Type: reasoning vs. non-reasoning
- Hyper-Parameters & Run Occasion
  - Fixed weight model?
  - Deterministic or no (temperature, top-k)
  - Quantization / hardware used

- Prompts:
  - Zeroshot / few shot
  - Style (e.g., chain of thought)
  - Persona
  - Definition
  - Rating scale (e.g., 1-5, 1-100)
  - Structure (e.g., JSON)
  - Minor variations (spacing, line breaks)

Note. These examples pertain to LLMs. However, they can easily be extended to supervised ML

AI-Specific Factors: Impact on LLM Scoring Inconsistency

Jiang, Hickman, Michaelides, Jackson, & Speer (in progress)

Fully-crossed manipulation of LLM ratings of interview transcripts….
Holding other facets constant…

Type of Inconsistency	Facet	*M* Correlation within Facet
Algorithm	LLM Model	.69
Algorithm	Model Size	.74
Hyper / Run Occasion	GPU Hardware	.91
Hyper / Run Occasion	Deterministic	1.0
Hyper / Run Occasion	Mid-High Temp	.92
Prompt	Minor/Wording	.96
Prompt	Scale	.87
Prompt	Structure	.87

Meaningful variation in LLM scores, particularly by algorithm

LLM results are quite variable and unlikely to generalize unless settings are held firm

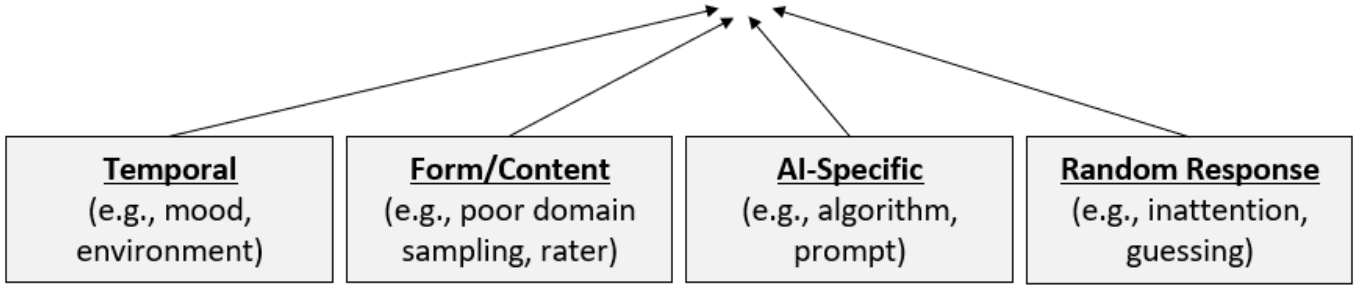
What impact does AI-specific inconsistency have on other psychometric properties?

Does AI-Specific Inconsistency Affect Traditional Psychometrics

Speer, Hong, & Brenner (in progress)

- Based on psychometric theory, AI-specific inconsistency is a form of error that should negatively impact scores

$$X = T + e$$



Although designs do not always measure these, all forms of error should nonetheless affect scores

$$r_{xx} = \frac{{\sigma_T}^2}{{\sigma_T}^2 + {\sigma_e}^2}$$

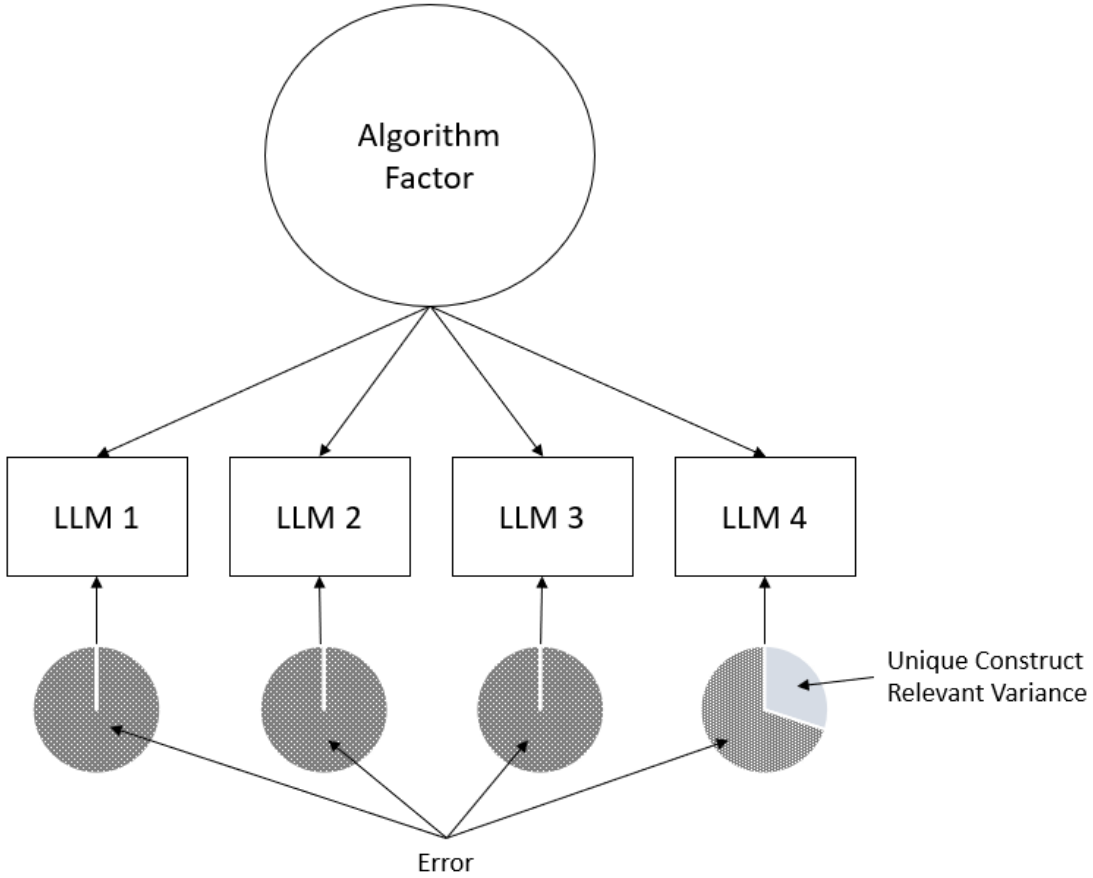
$$e_{temporal} + e_{form} + e_{AI} + e_{random}$$

Any reductions in error, of any type, should improve reliability

Does AI-Specific Inconsistency Affect Traditional Psychometrics

Speer et al. (in progress)

- However, some forms of inconsistency are systematic and meaningful (e.g., Hoffman & Woehr, 2009; Mõttus et al., 2017; Revelle, 2024)
- Might it be possible that certain AI observations contain unique but meaningful variance…?



Effects of AI-Specific Inconsistency: Speer, Hong, & Brenner (in progress)

- Major Goals: (a) examine AI-specific facet effects on score consistency, (b) investigate the effect of algorithmic inconsistency on traditional forms of reliability/validity, and (c) examine whether unique LLM variance contains systematic construct-related information
- Methods:
  - Sample 1: measuring conscientiousness of 262 Prolific employees who completed a single open-ended personality assessment question
  - Sample 2: measuring job satisfaction of 397 employees at large US company who completed engagement survey open-ended responses
  - Manipulation: fully-crossed LLM model (four levels) and four separate prompt factors (style, persona, scoring scale, and definition used, each with four levels) = 1024 sets of LLM scores
  - For each sample, collected human SME ratings (for inter-rater reliability), convergent measures (self-report Likert), and criterion outcomes (employee turnover, self-rated job performance)

Effects of AI-Specific Inconsistency: Speer, Hong, & Brenner (in progress)

Range for single observations / Comp = Composite

Facet	Sample	Single r_{xx}	Comp r_{xx}	Single SME r	Comp SME r	Single Conv r	Comp Conv r	Single r_{xy}	Comp r_{xy}
LLM	1	.67 to .89	.94	.75 to .87	.90	.58 to .66	.69	.25 to .32	.31
	2	.69 to .82	.93	.70 to .76	.81	.55 to .64	.67	.10 to .12	.12
Prompt Style	1	.81 to .93	.97	.79 to .86	.87	.60 to .66	.66	.27 to .31	.31
	2	.91 to .94	.98	.72 to .74	.76	.59 to .61	.62	.11 to .11	.11
Persona	1	.93 to .97	.99	.82 to .84	.84	.62 to .64	.64	.28 to .29	.29
	2	.96 to .97	.99	.74 to .74	.74	.60 to .60	.61	.11 to .11	.11
Scoring Scale	1	.77 to .91	.96	.78 to .86	.88	.58 to .65	.67	.26 to .30	.30
	2	.79 to .92	.96	.71 to .76	.78	.58 to .62	.64	.10 to .12	.12
Definition	1	.90 to .96	.98	.81 to .84	.85	.62 to .64	.65	.28 to .30	.29
	2	.92 to .96	.99	.73 to .74	.75	.59 to .60	.61	.10 to .11	.12

Ensemble composites had better psychometric properties

AI r_{xx} correlates with psychometric properties (more in a bit)

Effects of AI-Specific Inconsistency: Speer, Hong, & Brenner (in progress)

Profile Correlations

	1.	2.	3.	4.
1. Algorithmic r_{xx}		0.34	0.34	0.37
2. SME r_{xx}	0.69		0.96	0.75
3. Convergent r	0.67	0.97		0.79
4. r_{xy}	0.67	0.92	0.91	

Bottom diagonal (Sample 1)
Top diagonal (Sample 2)

AI-specific inconsistency negatively impacts psychometric properties

Used to calculate….

Factor	Level	Algo rxx	SME rxx	Conv r	rxy
Persona	Level0	0.94	0.84	0.64	0.29
Persona	Level1	0.96	0.83	0.63	0.28
Persona	Level2	0.93	0.82	0.62	0.28
Persona	Level3	0.97	0.83	0.64	0.28
Persona	comp	0.99	0.84	0.64	0.29
Model	Level0	0.67	0.75	0.58	0.25
Model	Level1	0.88	0.87	0.64	0.29
Model	Level2	0.89	0.87	0.66	0.32
Model	Level3	0.81	0.84	0.64	0.29
Model	comp	0.94	0.90	0.69	0.31
Prompt Style	Level0	0.92	0.86	0.66	0.31
Prompt Style	Level1	0.93	0.84	0.64	0.28
Prompt Style	Level2	0.93	0.84	0.65	0.29
Prompt Style	Level3	0.81	0.79	0.60	0.27
Prompt Style	comp	0.97	0.87	0.66	0.30
Def	Level0	0.90	0.81	0.62	0.28
Def	Level1	0.93	0.84	0.64	0.29
Def	Level2	0.96	0.84	0.63	0.29
Def	Level3	0.92	0.84	0.64	0.30
Def	comp	0.98	0.85	0.65	0.29
Scale	Level0	0.77	0.78	0.58	0.26
Scale	Level1	0.86	0.84	0.65	0.28
Scale	Level2	0.91	0.85	0.65	0.30
Scale	Level3	0.90	0.86	0.65	0.29
Scale	comp	0.96	0.87	0.67	0.3

Effects of AI-Specific Inconsistency: Speer, Hong, & Brenner (in progress)

- Were residuals (unique variance) systematically related to other variables?
  - Correlated residuals within each facet with validity measures
- Few residuals were significantly related to convergent/criterion measures
  - Going to discuss a few of these significant, unique effects…

Effects of AI-Specific Inconsistency: Speer, Hong, & Brenner (in progress)

LLM Model

Observation	Sample	Targeted Construct	Convergent *r*	Criterion *r*
Llama Residual	1	Conscientiousness	.04 (.00 to .08)	ns

Versus: Qwen, Mistral, Gemma

Terms significantly correlated with residual

Effects of AI-Specific Inconsistency: Speer, Hong, & Brenner (in progress)

Definition

Observation	Sample	Targeted Construct	Convergent *r*	Criterion *r*
High/Low FOR	1	Conscientiousness	.05 (.03 to .06)	ns
High/Low FOR	2	Job Satisfaction	.02 (.00 to .03)	ns

Versus: None, Def, High FOR

Terms significantly correlated with residual

Conscientiousness Scores

Job Satisfaction Scores

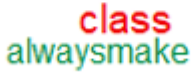
Effects of AI-Specific Inconsistency: Speer, Hong, & Brenner (in progress)

Prompt Style

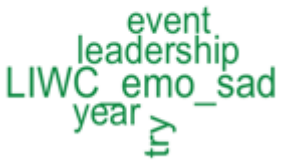
Observation	Sample	Targeted Construct	Convergent *r*	Criterion *r*
None (zero shot)	1	Conscientiousness	.05 (.03 to .07)	.06 (.03 to .08)
None (zero shot)	2	Job Satisfaction	.03 (.01 to .04)	ns

Versus: Chain of thought, comparative, self-ask

Terms significantly correlated with residual



Conscientiousness Scores



Job Satisfaction Scores

In Conclusion

In Conclusion

- Traditional psychometrics still apply to AI-based scoring, but there are unique design features that should be considered when estimating reliability and validity
- If using LLM scores, inter-rater reliability should be estimated using CFA designs rather than correlational designs
- New forms of AI inconsistency are important to consider
  - Inconsistency reflects error
  - Development of ensemble models, particularly across different LLMs, is recommended

Thank you!