Presented at the 2022 Conference on Test Security

October 26-28 Princeton, NJ

# A Comparison of an Approximation Score Similarity Index to Other Similarity Indices

Amanda A. Wolkowitz
Senior Psychometrician

Russell W. Smith
Chief Innovation Officer

Alpine Testing Solutions

# Purpose

- Compare relative performance of an approximate score similarity index to other indices
  - Without Item Response Theory (IRT)/specialized software
  - Less computationally intense
  - Can be run in real time or near real time

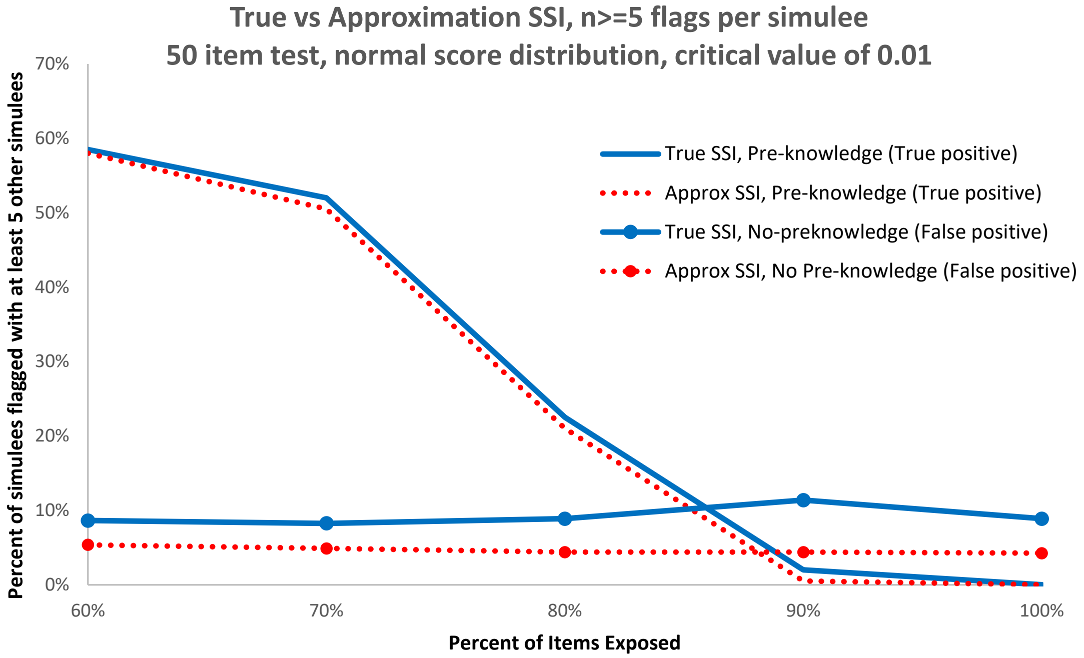
- The purpose is *not* to determine whether one method performs better than another, but rather if the relative outcome of the methods are comparable

# Purpose

- Extension of an NCME presentation by Smith (2021, 2022) and CoTS presentation at 11am in the Palmer Room

  Approximation answer and response similarity analyses: A practical approach

  - 60% or more of the content exposed, 20% of examinees with pre-knowledge



# Method

## Comparison of 4 methods

- aSSI
    - Approximation of the GBT method using the Bock's model

- Wollack's omega
    - Using the Bock model
    - Using the Rasch model

- Generalized binomial model (GBT) (van der Linden and Sotaridona (2006))
    - Using the Bock model
    - Using the Rasch model

- B3 (residual correlations for persons) (Foley 2019, Smith 2019)

## Datasets

- 1 simulation
- 3 real datasets

# aSSI

- aSSI = z-score (for persons 1 and 2)

$$z_{12} = \frac{(M_{12} - E^*_{12})}{\sqrt{npq}} \text{ , where}$$

***M*** is count of observed matches

***n*** is the number of items

***p*** is $E^*_{12}/n$ and ***q*** is $(1 - p)$

$E^*_{12}$ is the adjusted expected value:

$$E^*_{12} = n \cdot [s_1 s_2 + (1 - s_1)(1 - s_2)] + n \cdot b(1 - |s_1 - s_2|)(1 - |s_1 + s_2 - 1|)$$

where, $s_i$ is proportion correct score for person *i*, ***b*** is an adjustment to the magnitude of the correction set at 12.5%

# Wollack's (1996) Omega

- Calculation involves the use of a z-score (i.e., assumes a normal distribution)
- IRT-based model

$$\boldsymbol{\omega = \frac{O - E}{\sqrt{var(E)}}}$$ , where

*E* = expected agreement between two examinees

- sum of the probabilities that the copier's response (0,1) equals the observed source's response given the ability of the copier and the item's IRT parameters

*O* = observed agreement between two examinees

# GBT

- Calculation involves the use of a z-score (i.e., assumes a normal distribution)
- IRT-based model

$$\boldsymbol{GBT = \frac{O - E}{\sqrt{npq}}}$$

, where

*E* = sum of joint probabilities of matching scores (0,1) between two examinees given the ability of each examinee and the item's IRT parameters

*O* = observed agreement between two examinees

# B3

- Correlation of residuals for two people – indicates they are locally dependent
- Use interquartile range (IQR) to detect outliers
  - In a N(0,1):
    - z-score for Q1 = -0.6745
    - z-score for Q3 = 0.6745
    - → IQR = 1.349
  - Q1 – ***m***(IQR) = ***z***, where ***z*** = outlier z-score value, *m* = IQR multiplier

Example: Assume mean = 0, SD = 1 of the correlations of the person residuals

| One-tail p | ***z*** | ***m*** |
|---|---|---|
| 0.025 | 1.96 | 0.95 |
| 0.005 | 2.58 | 1.41 |
| 0.001 | -3.09 | 1.79 |
| 0.0001 | -3.72 | 2.26 |
| 0.00001 | -4.26 | 2.66 |

# Study

- **Simulation**
  - 100 item test, 1000 examinees, normal distribution, 60% of exposed content, 1% of examinees with pre-knowledge
  - Multiple critical values

- **Real data**
  - Exam A – 66 items, 416 examinees, known security issues
  - Exam B – 60 items, 1992 examinees
  - Exam C – 66 items, 1109 examinees
  - Focus on correlation of methods, not accuracy of detection

- **Data notes**
  - Focused on SSI, so only detection based on 0/1 responses
  - Methods applying Bock’s model were collapsed to 0/1 responses
  - Methods applying Bock’s model were also applied using the Rasch model

# Simulation: 1% of people with pre-knowledge on first 60% of items

N items = 100
N examinees = 1,000
N items exposed = 60

N examinees with pre-knowledge = 10
N pairs of examinees with pre-knowledge = c(10,2) = 45
Total N pairs of examinees = c(1000,2) = 499,500

For each critical value,

**True positive =**
N pairs of correctly flagged pairs of examinees/45

**False positive =**
N pairs of incorrectly flagged pairs of examinees/(499,500-45)

## True Positives

| 1-tail probability | z | Wollack's ω - Rasch | Wollack's ω - Bock | GBT - Bock | GBT - Rasch | B3 | aSSI |
|---|---|---|---|---|---|---|---|
| 0.025 | 1.96 | 91.1% | 91.1% | 46.7% | 46.7% | 57.8% | **37.8%** |
| 0.005 | 2.58 | 80.0% | 80.0% | 22.2% | 26.7% | 26.7% | **17.8%** |
| 0.001 | 3.09 | 53.3% | 55.6% | 6.7% | 6.7% | 4.4% | **4.4%** |
| 0.0001 | 3.72 | 31.1% | 31.1% | 0.0% | 2.2% | 0.0% | **0.0%** |
| 0.00001 | 4.26 | 8.9% | 6.7% | 0.0% | 0.0% | 0.0% | **0.0%** |

## False Positives

| 1-tail probability | z | Wollack's ω - Rasch | Wollack's ω - Bock | GBT - Bock | GBT - Rasch | B3 | aSSI |
|---|---|---|---|---|---|---|---|
| 0.025 | 1.96 | 2.3% | 2.4% | 0.8% | 0.8% | 0.1% | **0.5%** |
| 0.005 | 2.58 | 0.4% | 0.4% | 0.1% | 0.1% | 0.0% | **0.0%** |
| 0.001 | 3.09 | 0.1% | 0.1% | 0.0% | 0.0% | 0.0% | **0.0%** |
| 0.0001 | 3.72 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | **0.0%** |
| 0.00001 | 4.26 | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | **0.0%** |

# Simulation – Correlations of SSI Statistics

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | | 0.999 | 0.950 | 0.951 | 0.938 | 0.928 |
| B | | | 0.952 | 0.95 | 0.937 | 0.926 |
| C | | | | 0.998 | 0.869 | 0.978 |
| D | | | | | 0.870 | 0.981 |
| E | | | | | | 0.853 |

Which method is represented by the letters shown in the correlation matrix above?

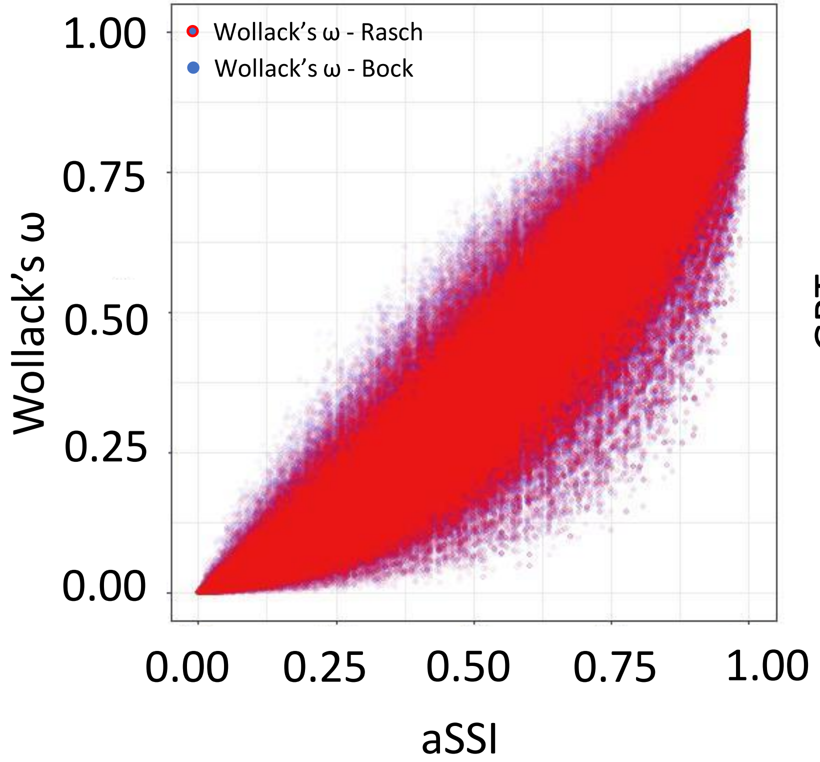
- aSSI
- B3
- GBT - Bock
- GBT – Rasch
- Wollack's ω - Bock
- Wollack's ω - Rasch

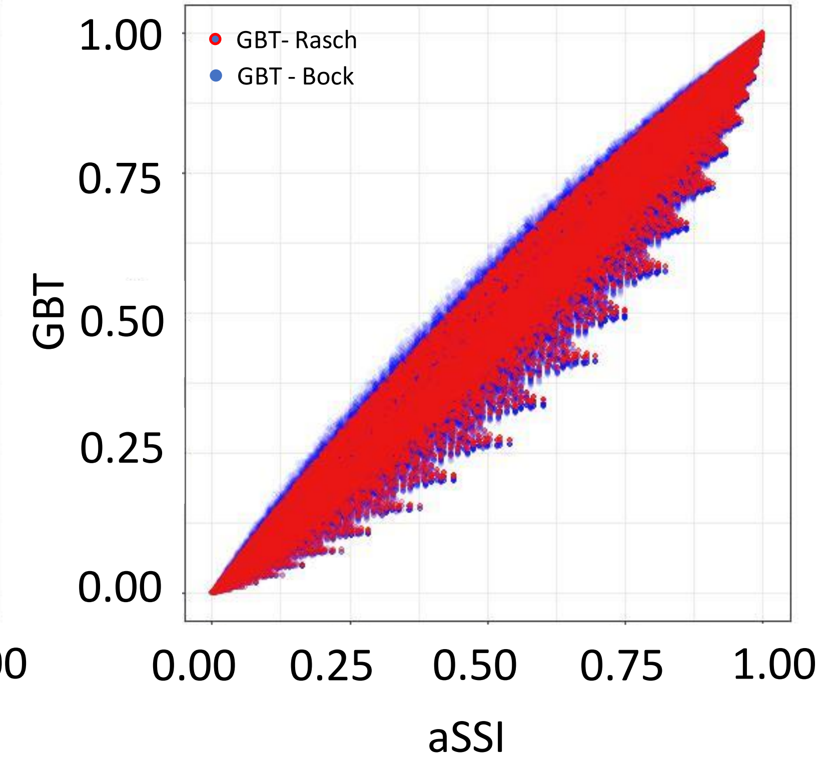
# Simulation – Correlations of SSI Statistics

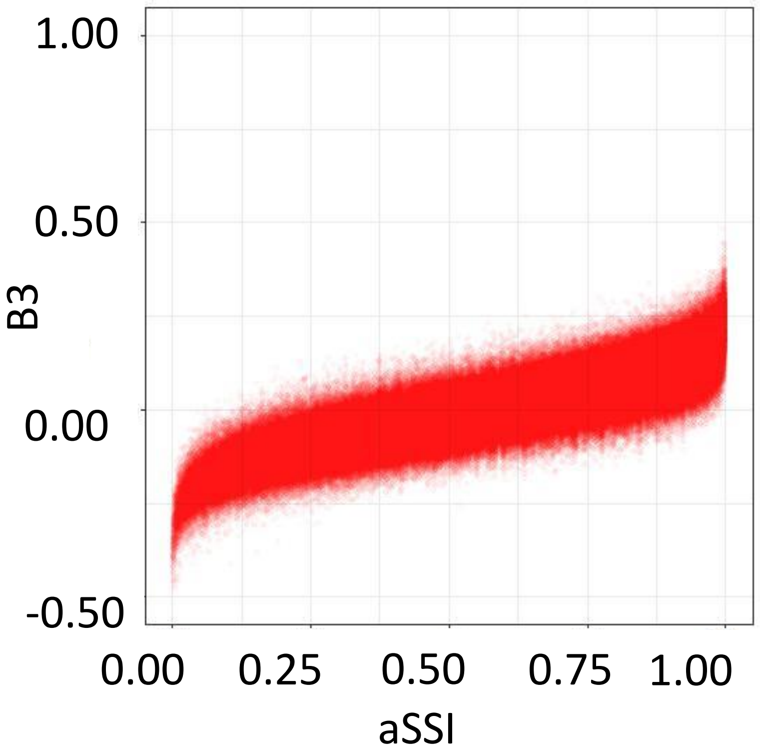
| | Wollack's ω - Rasch | Wollack's ω - Bock | GBT - Bock | GBT - Rasch | B3 | aSSI |
|---|---|---|---|---|---|---|
| Wollack's ω - Rasch | | 0.999 | 0.950 | 0.951 | 0.938 | 0.928 |
| Wollack's ω - Bock | | | 0.952 | 0.950 | 0.937 | 0.926 |
| GBT - Bock | | | | 0.998 | 0.869 | 0.978 |
| GBT - Rasch | | | | | 0.870 | 0.981 |
| B3 | | | | | | 0.853 |

# Simulations – Correlations of SSI Statistics

| | Wollack's ω - Rasch | Wollack's ω - Bock | GBT - Bock | GBT - Rasch | B3 | aSSI |
|---|---|---|---|---|---|---|
| Wollack's ω - Rasch | | 0.999 | 0.950 | 0.951 | 0.938 | 0.928 |
| Wollack's ω - Bock | | | 0.952 | 0.950 | 0.937 | 0.926 |
| GBT - Bock | | | | 0.998 | 0.869 | 0.978 |
| GBT - Rasch | | | | | 0.870 | 0.981 |
| B3 | | | | | | 0.853 |

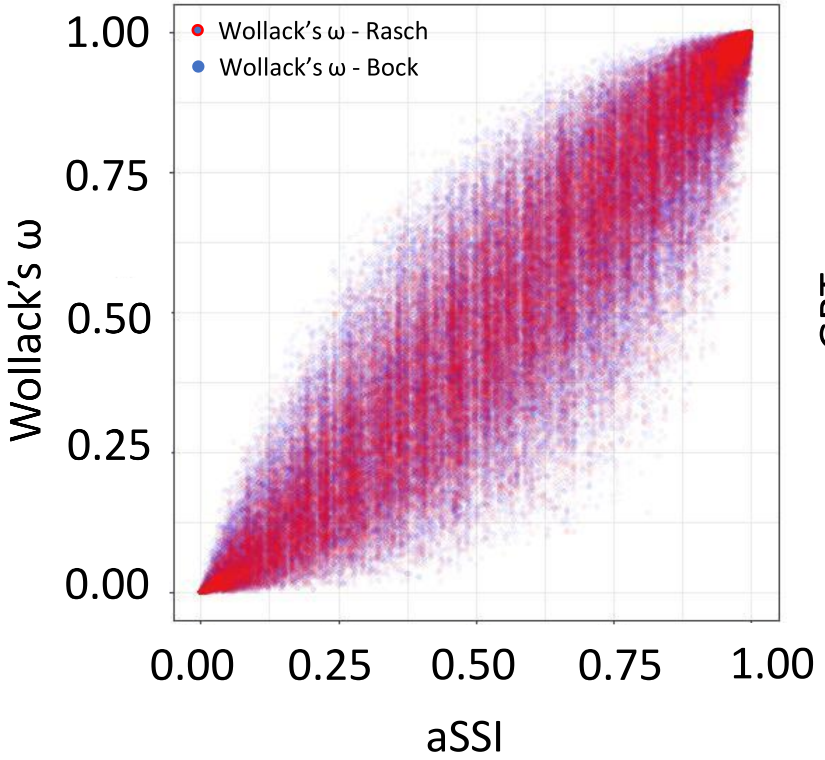


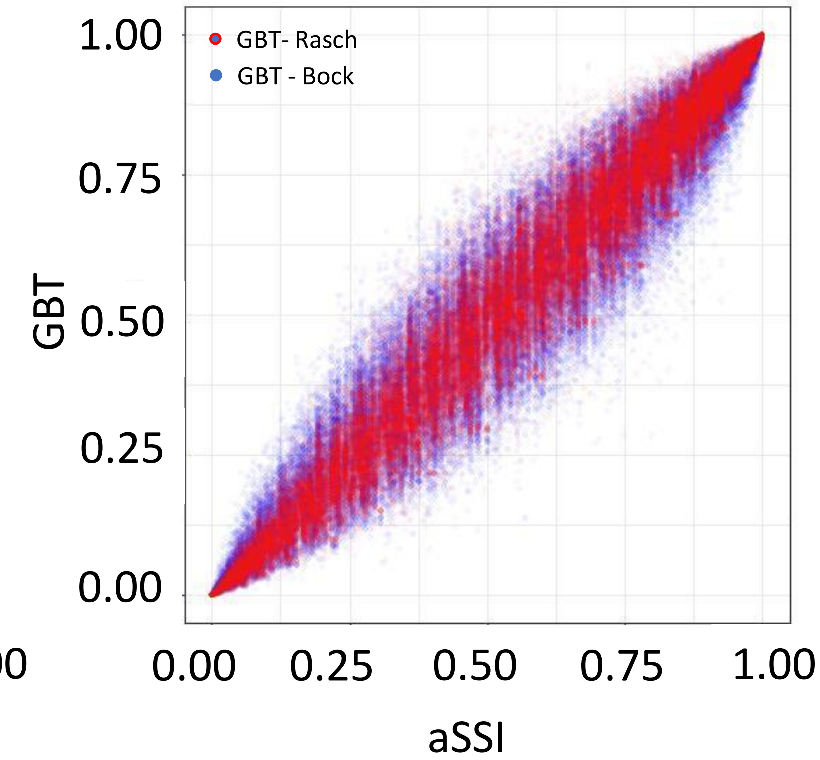



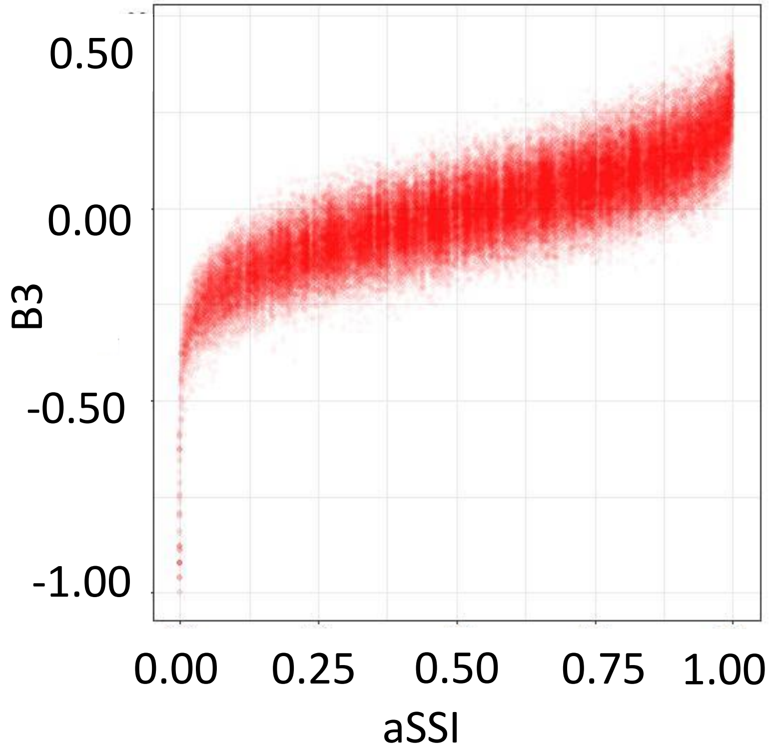



# Exam A – Correlations of SSI Statistics

| | Wollack's ω - Rasch | Wollack's ω - Bock | GBT - Bock | GBT - Rasch | B3 | aSSI |
|---|---|---|---|---|---|---|
| Wollack's ω - Rasch | | 0.961 | 0.907 | 0.922 | 0.886 | 0.909 |
| Wollack's ω - Bock | | | 0.942 | 0.891 | 0.891 | 0.894 |
| GBT - Bock | | | | 0.951 | 0.850 | 0.951 |
| GBT - Rasch | | | | | 0.832 | 0.973 |
| B3 | | | | | | 0.839 |

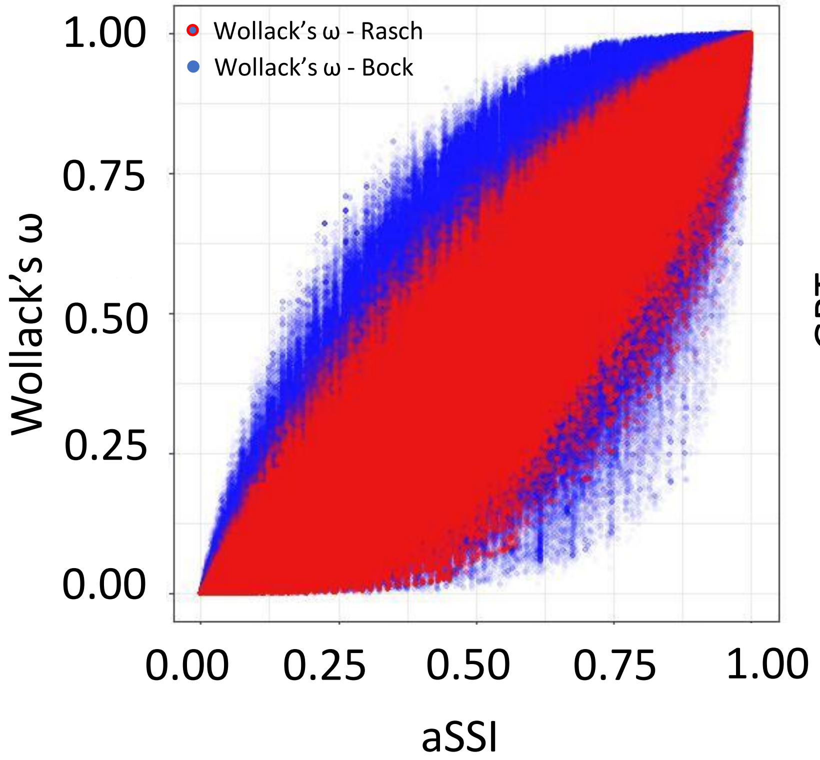


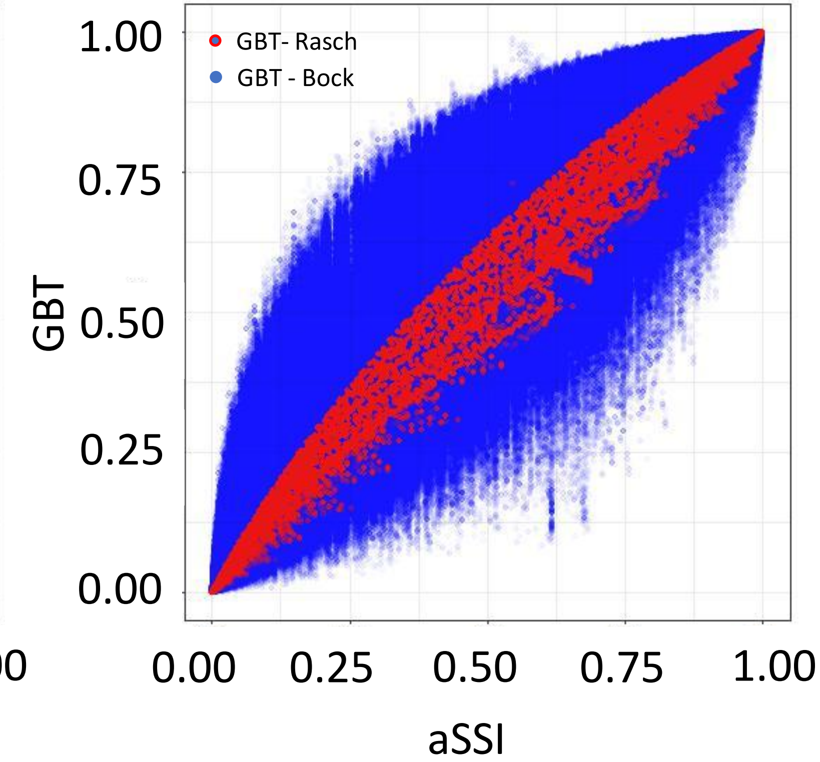



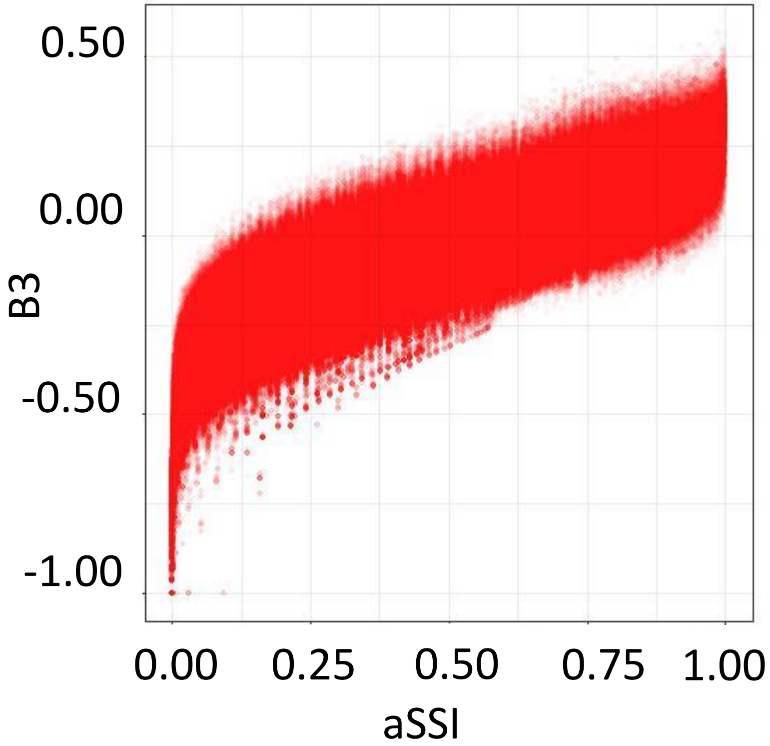



# Exam B – Correlations of SSI Statistics

| | Wollack's ω - Rasch | Wollack's ω - Bock | GBT - Bock | GBT - Rasch | B3 | aSSI |
|---|---|---|---|---|---|---|
| Wollack's ω - Rasch | | 0.788 | 0.740 | 0.924 | 0.888 | 0.930 |
| Wollack's ω - Bock | | | 0.962 | 0.854 | 0.701 | 0.844 |
| GBT - Bock | | | | 0.868 | 0.656 | 0.855 |
| GBT - Rasch | | | | | 0.851 | 0.986 |
| B3 | | | | | | 0.835 |

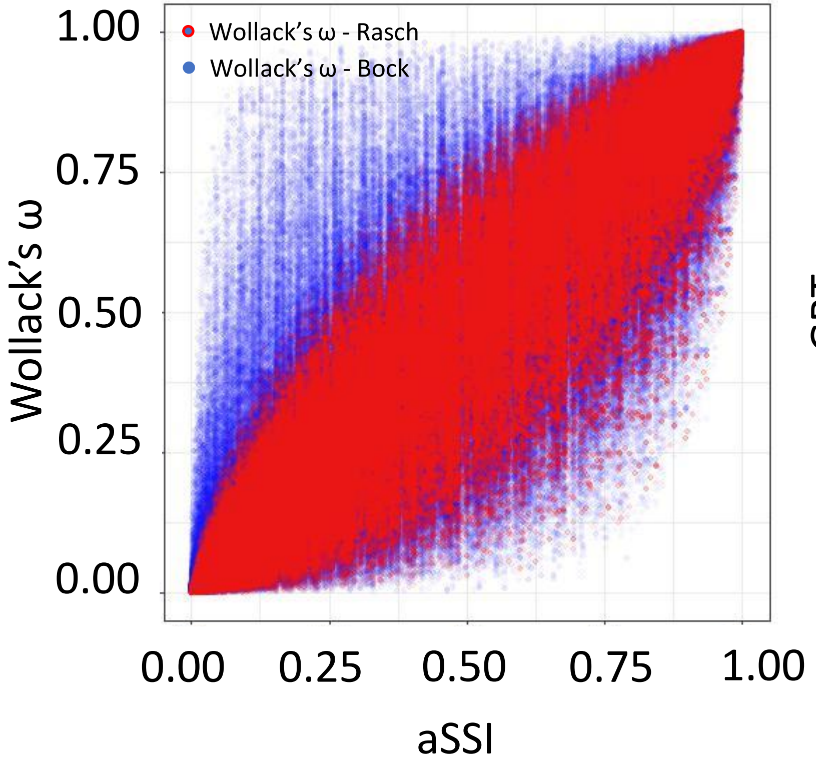


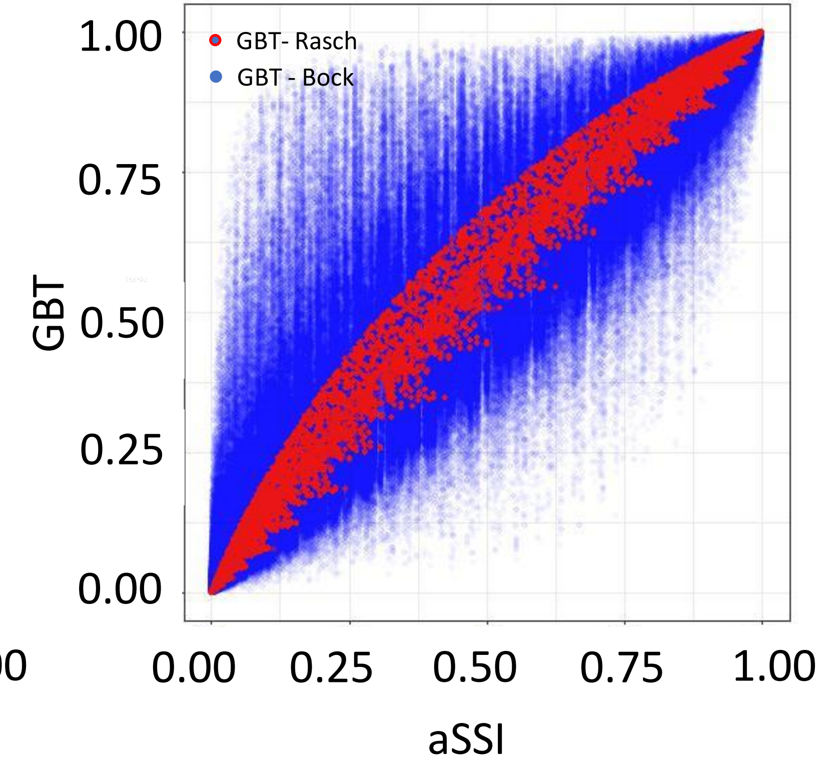



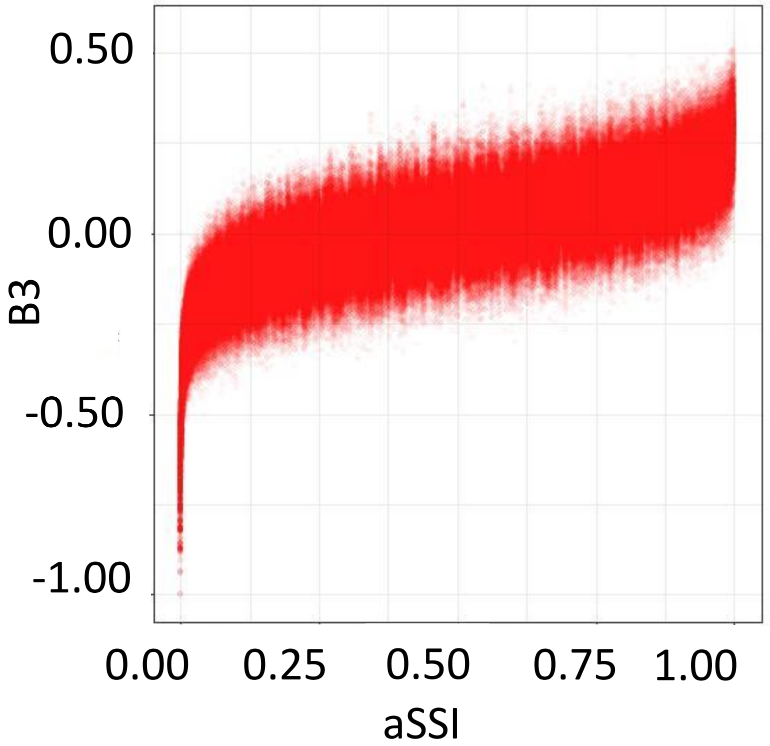



# Exam C – Correlations of SSI Statistics

| | Wollack's ω - Rasch | Wollack's ω - Bock | GBT - Bock | GBT - Rasch | B3 | aSSI |
|---|---|---|---|---|---|---|
| Wollack's ω - Rasch | | 0.929 | 0.892 | 0.927 | 0.891 | 0.908 |
| Wollack's ω - Bock | | | 0.948 | 0.845 | 0.858 | 0.813 |
| GBT - Bock | | | | 0.904 | 0.807 | 0.860 |
| GBT - Rasch | | | | | 0.810 | 0.982 |
| B3 | | | | | | 0.765 |



Contributions: Approximation SSI

- Does aSSI do a good job and detecting pre-knowledge compared to other methods?

  **Yes. It is also much simpler to compute and does not require IRT.**

- Does aSSI work well with real and simulated data?

  **Yes. The strength of the correlation of the methods is strong under both situations.**

- With which true SSI method does aSSI most strongly correlate?

| | Min. *r* | Max. *r* |
|---|---|---|
| GBT - Rasch | 0.973 | 0.986 |
| Wollack's ω - Rasch | 0.908 | 0.930 |
| GBT - Bock | 0.855 | 0.951 |
| Wollack's ω - Bock | 0.813 | 0.894 |
| B3 | 0.765 | 0.839 |

# Contributions: Approximation SSI

- aSSI is good for "real-time"
- Solves a real-world problem
- No IRT, no calibration, computationally less intense
- Provides pairwise estimated probabilities
- All that's required is scores, test length, and a weight
- It does well under the right conditions (see Smith's presentation at 11am)
- It correlates well with other known SSI statistics, but critical value may need to be adjusted for similar effectiveness
- B3 is less well known, but does a reasonable job, too!

# References

Foley, B. P. (2019). Collusion Detection Using an Extension of Yen's Q3 Statistic. Presented at the 8th Annual Conference on Test Security. Miami, FL.

Smith, R. W. (2019). Comparing B3 to Answer Similarity Index for Detecting Collusion. Presented at the 8th Annual Conference on Test Security. Miami, FL.

Smith, R. W. (2021, October 6-7). *A Practical Approximation of Response Similarity*. Conference on Test Security, online.

Smith, R. W. (2022, April). *Approximation answer and response similarity analyses: A practical approach* [Paper presentation]. Annual meeting of the National Council on Measurement in Education (NCME), San Diego, CA.

van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*(3), 283-304.

Wollack, J.A.(1996). Detection of answer copying using item response theory. *Dissertation Abstracts International, 57/05*, 2015.

# Thank you!

amanda.wolkowitz@alpinetesting.com
russell.smith@alpinetesting.com