Detecting Carelessness in High Stakes Cognitive Ability Assessments – Practitioner's Guide

Julia L. Walsh | Kyle J. Mann
Infoscitex Corporation
Dayton, OH

Sophie Romay | John D. Trent
Air Force Personnel Center's Strategic Research and Assessment Branch,
St. Antonio, TX

Thomas R. Carretta
Air Force Research Laboratory
Wright-Patterson AFB, OH

Disclaimer

- The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense or of the United States Air Force.

- The content or appearance of hyperlinks does not reflect an official DoD, Air Force, Air Force Research Laboratory position or endorsement of the external websites, or the information, products, or services contained therein.

Who is the intended audience?

- ✓ HR professionals, I/O psychologists, researchers, and psychometricians who design, administer, and analyze large scale assessments to make generalizable inferences from these assessments about the constructs being measured and their predictive power.

- Not for applicants and for HR professionals who work with third party assessment vendors and have no access to the raw data.

Introduction: Example 1

- Marissa is a workforce analyst at an HR department of a mid-size firm.
- She is tasked with administering an organization-wide employee engagement survey every quarter.
- She designs the survey to include 75 self-report questions that will take 20-30 minutes for employees to complete.
- She sends the survey link to all employees.

While waiting for the data to come in, Marissa is wondering, “Should I expect carelessness in my data?”

Introduction: Carelessness in Low Stakes Surveys

- **To answer this question, Marissa reviewed her notes:**

- **What are assessment stakes?**
  - Low stakes = no direct impact on the survey-taker's valued outcomes (e.g., promotion, raise);
  - High stakes = direct impact on the survey-taker's valued outcomes (e.g., promotion, raise).

- **What is carelessness?**
  - Carelessness, or Insufficient Effort Responding (IER), represents a lack of motivation when responding to surveys (DeSimone et al., 2018).

Introduction: Carelessness in Low Stakes Surveys (cont'd)

- **Why should we care about carelessness?**
  - IER is pervasive in I/O and HR research (Moore et al., in progress)
    - Up to 22.7% of low stakes survey responses are careless
  - Research suggests that carelessness influences psychometrics of the assessment thereby resulting in potentially erroneous conclusions, but
  - Research also suggests that some prevalence of IER may be tolerable (Huang et al., 2012; Huang & DeSimone, 2017)

- **Is there research to guide practitioners?**
  - Much research has focused on IER detection and removal strategies in low stakes settings.

Equipped with this knowledge, Marissa knows the employee engagement data will likely have carelessness, but she is prepared to tackle it!

Introduction: Example 2

- John is a cognitive ability assessment specialist at a large-scale telecommunications organization.
- He is tasked with administering a personnel selection assessment to applicants for call center positions.
- He designs the survey to include 75 verbal and quantitative questions that will take 20-30 minutes for applicants to complete.
- He follows all the best practices in the design, validation, and administration of the battery.

While waiting for the data to come in, John is wondering, "Should I expect carelessness in my data?"

Introduction: Carelessness in High Stakes Assessments

- **Would job applicants engage in IER? Why?**
  - Logically, job applicants should be motivated to pay attention because their career outcomes depend on how well they do on the test.

  - However, as practice shows, some job applicants may engage in IER because:
    a) they may not care for the outcomes of the test (e.g., USAFA);
    b) they may perceive the stakes of various parts of the test as low or high depending on whether each test makes up an operational composite;
    c) they may use their first attempt as an opportunity to practice and then retake the test; or
    d) they are incumbents who partake in concurrent validation data collection involving potential new cognitive predictor.*
    e) temporary factors such as lack of focus, anxiety, etc.

Introduction: Carelessness in High Stakes Assessments (cont'd)

- **Assuming IER exists in high stakes assessments:**
  - It is important to separate careless (unmotivated) cases from careful (motivated) cases who may have lower ability and therefore may engage in rapid-guessing behaviors (Schnipke & Scrams, 1997).
    - This is especially true for speeded tests where the test-taking behavior for low ability and IER may look the same.
  - The removal of IER may reduce data noise and thereby purify population estimates. However, the removal of low ability cases may obfuscate population estimates thereby leading to erroneous conclusions.
- **Is there research to guide practitioners?**
  - Not much. So, what should John do?

Introduction: Purpose of Current Study

- **To help practitioners like John, this study was designed:**
  1. to examine the prevalence of IER in a large-scale cognitive ability assessments;
  2. to recommend the IER detection methods that may be helpful in data cleaning;
  3. to examine the impact the IER detection methods might have on psychometric properties of these type of assessments; and
  4. to examine the convergence among the IER detection methods.

Method: Data & Perceived Stakes

- **Data & Perceived Stakes**
  - Data came from a large-scale archival dataset containing the Air Force Officer Qualifying Test (AFOQT)* administered 2015-2020;
  - AFOQT consists of 9 scored and 1 unscored cognitive ability tests.

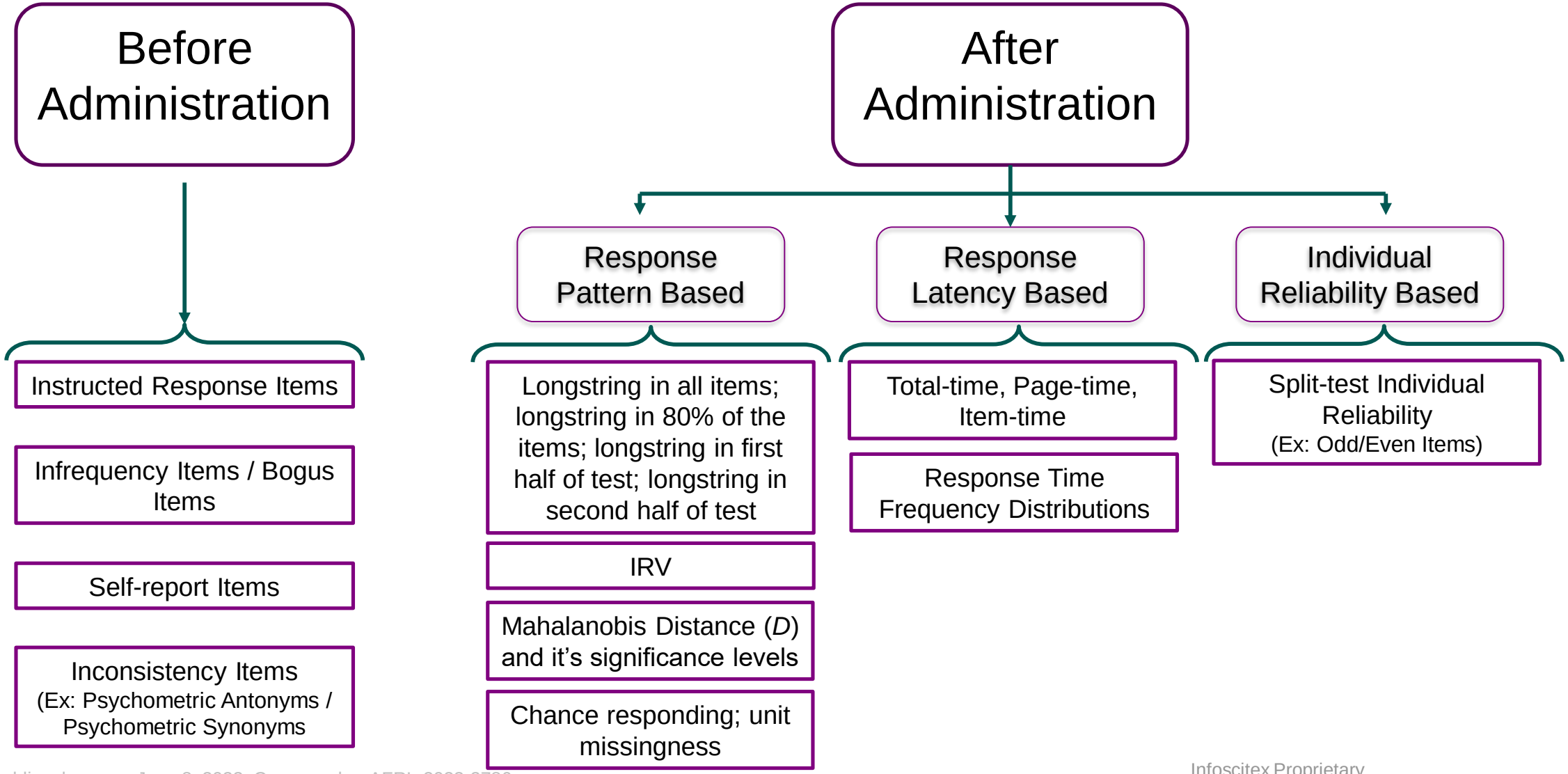
Part	Subtests (in order of administration)	Abbrev.	Included in one of six operational composites?	Hypothesized Stakes
Commissioning	Verbal Analogies	VA	Yes	High
	Arithmetic Reasoning	AR	Yes	High
	Word Knowledge	WK	Yes	High
	Math Knowledge	MK	Yes	High
	Reading Comprehension	RC	Yes	High
Classification	Physical Science	PS	No	Low
	Table Reading	TR	Yes	Variable
	Instrument Comprehension	IC	Yes	Variable
	Block Counting	BC	Yes	Variable
	Aviation Information	AI	Yes	Variable

Method: Sample Size and Demographics

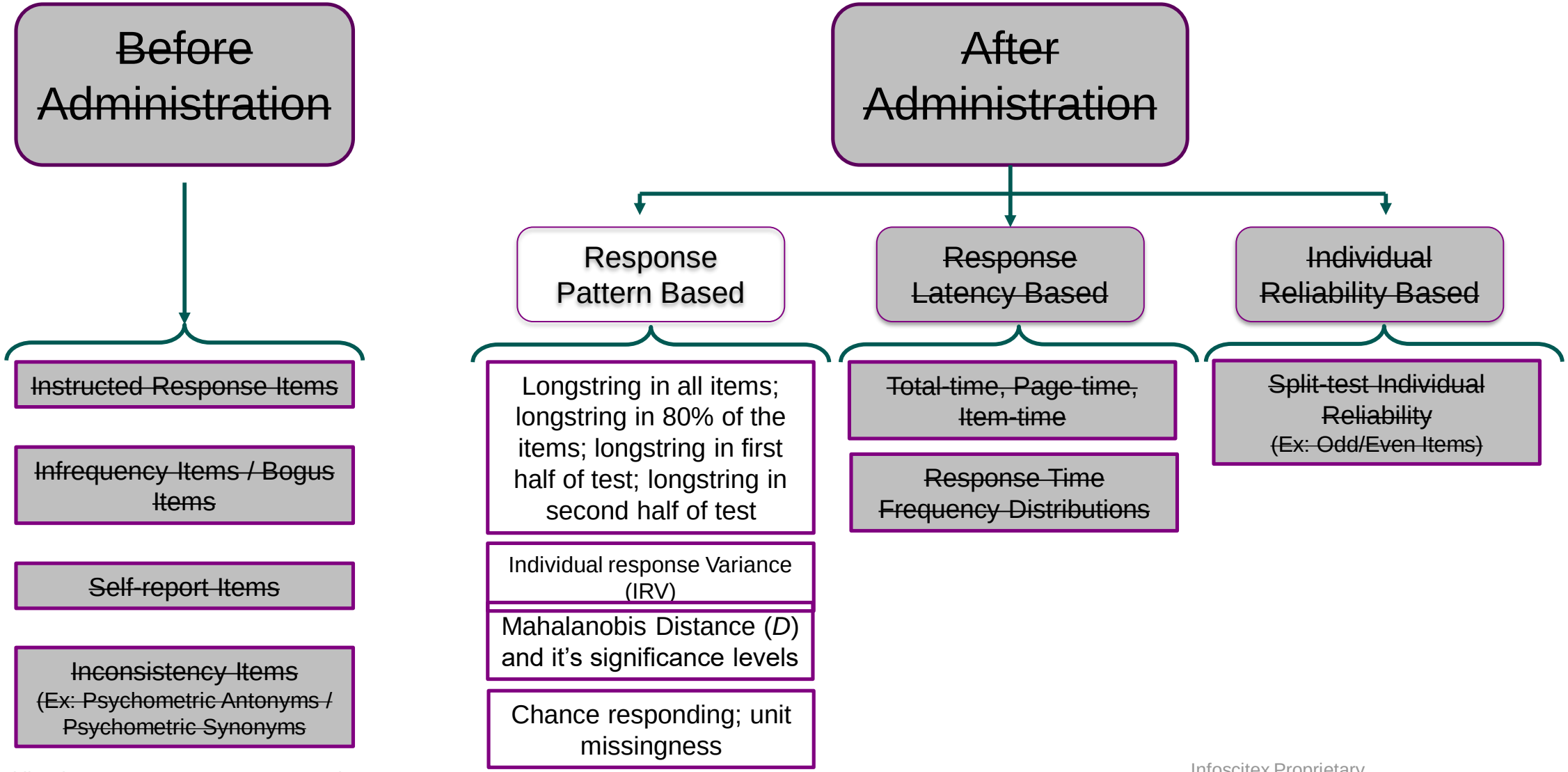
- **Sample Size**
  - 36,099 applicants who are applying for commissioning or who are interested in pursuing a rated career field with the United States Air Force (manned/unmanned aircraft pilot, combat systems officer, air battle manager)

- **Demographics**
  - The majority of the sample were White (67%) non-Hispanic (92%) Males (59%)

Method: IER Detection Methods



Method: IER Detection Methods



Results: Prevalence

- The IER detection methods seem to have split into – true carelessness and mixed (true carelessness and low ability).

Part	True Carelessness	Mixed
Commissioning (high stakes)	0% to 5.26%	.08% to 11.85%
Classification (low-to-variable stakes)	0% to 15.74%	.36% to 22.74%

- Together these results suggest that:
  1. there is IER in high-stakes cognitive ability tests;
  2. it varies based on perceived stakes of each subtest;
  3. IER detection methods need to be chosen wisely to avoid removing low ability test-takers

Results: The Choice of IER

- **RECOMMEND REMOVING** – True Carelessness
  - Unit carelessness
  - Longstring in the first half of the test
  - IRV (3 SD above or below mean IRV)
  - Mahalanobis Distance 3 SD above mean Mahalanobis Distance
  - Chance responding

- **DO NOT RECOMMEND REMOVING** – Mixed (true carelessness and low ability)
  - Longstring in all items
  - Longstring in the second half of the test
  - Longstring in the first 80% of the test
  - Mahalanobis Distance Significance

Results: Psychometric Impact

Part	Decrease in *N* ranged from	Difficulty	Internal Consistency	Alpha	Cohen's *d*
Commissioning	.53% to 5.26%	easier	lower	lower	Same or lower
Classification	2.75% to 14.15%	easier	lower	lower	Same or lower

Results: IER Detection Method Convergence

Part	Convergence between Dissimilar IER Detection Methods	Convergence between Similar IER Detection Methods
Commissioning	0% to .03%	.01% to 1.52%
Classification	0% to .04%	0% to .33%

Discussion & Conclusion

- There is a gap in the literature regarding the detection and removal of carelessness in high stakes cognitive ability tests. This study is one of the first to fill the gap.

- **Major takeaways of the current research:**
  1. there *is* IER in high-stakes cognitive ability tests and it varies based on perceived stakes of each subtest;
  2. to avoid the risk of removing low ability cases, more conservative IER detection methods are recommended;
  3. the removal of IER has a nontrivial impact on psychometrics;
  4. different IER detection methods capture different types of carelessness and may need to be used simultaneously.

Equipped with this knowledge, John knows the data will likely have carelessness, but he is prepared to tackle it!

References

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, *67*(2), 309-338.

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), 101384.

Kline, T. J. (2005). Classical test theory: Assumptions, equations, limitations, and item analyses. *Psychological testing: A practical approach to design and evaluation* (pp. 91-105). Thousand Oaks, CA: Sage.

Miller, R. E. (1968). *Development of officer selection and classification tests-1968*. AFHRL-TR-68-104. Lackland AFB, TX; Personnel Research Division.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74-104.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213-232.

Tabachnick, B., & Fidell, L. (2019). *Using multivariate statistics* (7th ed.). Boston: Allyn and Bacon.

Questions?