

Mitigating the Impact of Invalid Responding to Improve Data Quality in Psychopathology & Personality Disorder Research

Danielle Burchett, Ph.D.

Associate Professor of Clinical Psychology

Director, Clinical and Forensic Evaluation (CAFE) Lab

California State University, Monterey Bay

**March 23, 2023 | APA Center for Learning & Career Development
Statistical Methods & Data Collection Series**





Learning Objective:

Attendees will review the negative impacts that invalid response styles have on self-report questionnaire data quality and learn strategies to prevent or remove low-quality data before conducting analyses.

Disclosures:

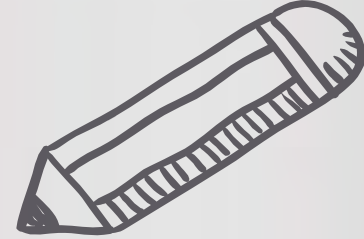


Danielle Burchett has received research grant funds from the University of Minnesota Press–Test Division to study the Minnesota Multiphasic Personality Inventory (MMPI) measures. The use of the MMPI measures (and other commercial and non-commercial instruments) will be discussed as part of the workshop.

Questions to Discuss Today

- ◆ What are the Types of Invalid Responding?
 - ◆ *When Does it Happen?*
- ◆ What's Its Impact on Research Conclusions?
- ◆ What Strategies Can Reduce the Risk of Erroneous Conclusions?
- ◆ Where to Go Next?

1.



Types of Invalid Responding

Protocol Validity

What is Protocol Validity?

- ◆ Does an individual examinee's test protocol (results) validly reflect what's true for them?
- ◆ **Clinicians:** Implications for assessment conclusions
- ◆ **Researchers:** Implications for the validity of sample data

1	☒ T	☐ F
2	☐ T	☒ F
3	☐ T	☒ F
4	☒ T	☐ F

Threats to Protocol Validity

Non-Content-Based Invalid Responding

- ◆ Unscorable
 - ◆ Skipped Items
 - ◆ Multiple Answers
- ◆ Variable/Random
- ◆ Fixed Acquiescent
- ◆ Fixed Counteracquiescent



- ◆ Reading & Language Comprehension
- ◆ Confusion
- ◆ Thought Disorganization
- ◆ Noisy Environment
- ◆ Careless Approach
- ◆ Recording Errors

Ben-Porath (2013)

Threats to Protocol Validity

Content-Based Invalid Responding

Overreporting

- ◆ Symptom (or Undesired Trait)
Exaggeration or Fabrication
 - ◆ Personality, Psychopathology,
Cognitive, & Somatic
Domains
- ◆ *Unintentional Overreporting*
- ◆ *Feigning; Malingering*

Underreporting

- ◆ Denial of Symptoms or
Exaggeration of Virtues (or
Desired Traits)
- ◆ *Self-Deception; Social
Desirability*
- ◆ *Other-Deception;
Impression Management;
Minimization; Denial;
Defensiveness*

Ben-Porath (2013); Burchett & Bagby (2014); Rogers (2018)

2.

Relevant Contexts



Data Collection Contexts

- ◆ **Collected for Research Purposes**

- ◆ College Students, Local Community Members, Online Participants

- ◆ **Collected for Real-World Decisions**

- ◆ Outpatient, Inpatient, Criminal and Civil Forensic, Medical, Pre-Employment



Data Collection Contexts

- ◆ Collected for Research Purposes

- ◆ Low-Stakes Context

- ◆ Potential for Rushing or Disengagement
Leading to **Non-Content-Based Invalid
Responding**





Data Collection Contexts

◆ Collected for Real-World Decisions

◆ High-Stakes Context



◆ Potential Motivation to **Underreport**

◆ e.g., Pre-Employment; Custody



◆ Potential Motivation to **Overreport**

◆ e.g., Forensic, Worker's Compensation

3.



Impact on Research Conclusions

Negative Impacts on Conclusions

- ◆ Variable & Fixed Responding

- ◆ ↑ Mean Substantive Scale Scores

- ◆ *Especially with rare-item scales*

- ◆ Changed Reliability & Validity Coefficients, Between-Groups Variance, & Classification Accuracy Metrics

Burchett et al. (2016); Chandler et al. (2020); Cheung et al. (2017); Handel et al. (2010) ¹⁴

Negative Impacts on Conclusions

◆ Underreporting

- ◆ ↓ Neuroticism; ↑ Extraversion, Agreeableness, & Conscientiousness
- ◆ ↓ Psychopathology
- ◆ Degraded Concurrent Validity (with Criteria Uninfluenced by Underreporting)

Anderson et al. (2013); Dhillon et al. (2017); Sellbom & Bagby (2008); Whitman et al. (2021) ¹⁵

Negative Impacts on Conclusions

◆ Overreporting

- ◆ ↑ Neuroticism; ↓ Extraversion, Agreeableness, & Conscientiousness
- ◆ ↑ Psychopathology
- ◆ Degraded Concurrent Validity (with Criteria Uninfluenced by Overreporting)

Burchett & Ben-Porath (2010); Dhillon et al. (2017); Sellbom & Bagby (2008); Whitman et al. (2021) ¹⁶

Degraded Concurrent Validity

- ◆ Burchett & Ben-Porath (2010) Example
 - ◆ Criteria Unaffected by Overreporting
 - ◆ Self-Report Measures Administered *Before* Overreporting Instructions
 - ◆ Self-Report Measures While Asked to Overreport Psychopathology or Somatic Symptoms

Impact on Means & SDs	Standard Instructions ($n = 124$)		Psychopathology Instructions ($n = 94$)		Somatic Instructions ($n = 94$)	
	M	SD	M	SD	M	SD
MMPI-2-RF Scale						
Emotional / Internalizing Dysfunction (EID)	53.05	11.90	80.96	12.04	62.54	12.94
Thought Dysfunction (THD)	55.42	11.34	95.43	13.09	65.19	17.18
Behavioral / Externalizing Dysfunction (BXD)	53.41	11.12	80.16	16.66	51.33	12.50

Impact on EID Correlations (r)	Standard Instructions (<i>n</i> = 124)	Psychopathology Instructions (<i>n</i> = 94)	Somatic Instructions (<i>n</i> = 94)
EASI-II Emotionality	.50	.22	.18
Rosenberg Self- Esteem Scale	-.78	.04	-.25
SCL-90-R Interpersonal Sensitivity	.70	.02	.19
SCL-90-R Depression	.78	-.02	.26
SCL-90-R Positive Symptom Total	.57	-0.05	.21

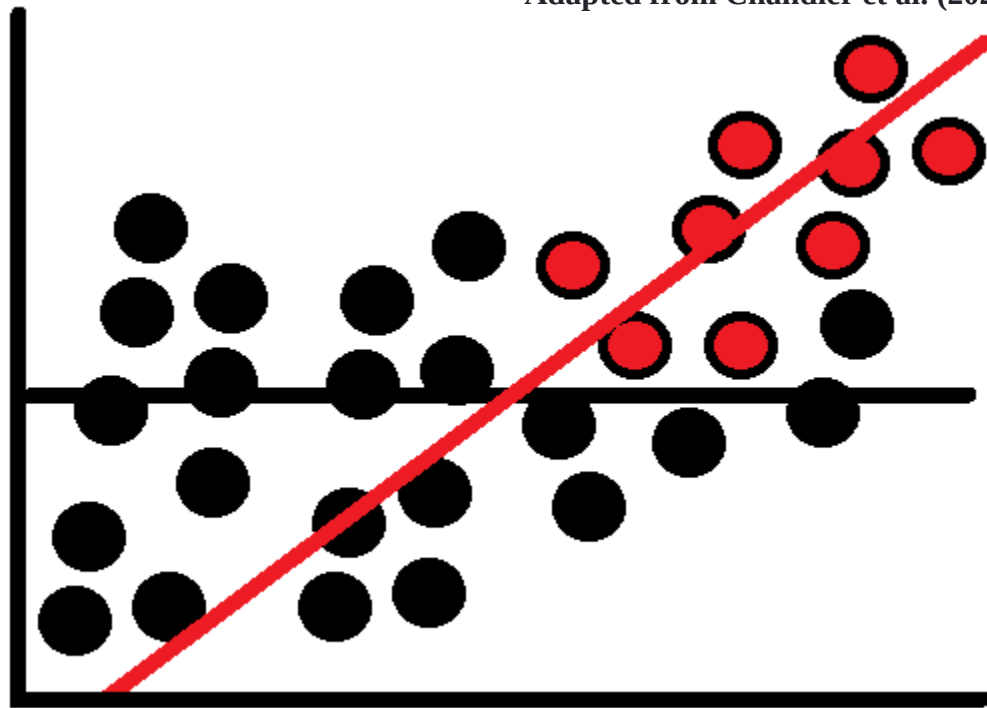
Adapted from Burchett & Ben-Porath (2010)

Erroneously Strong Associations

- ◆ Results Can Be Deceptively Strong
 - ◆ Spurious Self-Report Measure Associations & Between-Group Differences
 - ◆ Impact is Particularly Problematic for Skewed Variables & Rare Subgroups
 - ◆ (*Like Psychopathology & PD Scales*)

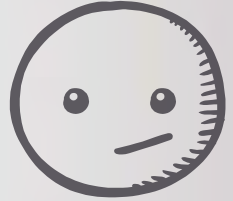
Burchett, Sellbom, & Bagby (2023); Chandler et al. (2020) ²⁰

Adapted from Chandler et al. (2020)



4.

Strategies to Reduce Risks of Erroneous Conclusions



Existing Strategies

- ◆ **Prevent Invalid Data**
 - ◆ Can We Stop It Before It Happens?
- ◆ **Identify and Remove Invalid Data**
 - ◆ Reduce Error Variance
 - ◆ Obtain Less Biased Associations Between Constructs Under Study

Burchett et al. (2023); Chandler et al. (2020); Goldammer et al. (2020) ²³

Preventing Invalid Responding

- ◆ Promising with Limitations:
 - ◆ Low Face-Valid Instruments
 - ◆ Forced-Choice Personality Measures
 - ◆ Scales with True/False Item Balance

Preventing Invalid Responding

- ◆ Communicate Value of Participation
- ◆ Inform About Length and Nature of Study
- ◆ Build in Breaks for Longer Studies
- ◆ Incentives
- ◆ Avoid Study Goal Descriptions that Cue Social Desirability
- ◆ Integrating Interview, Informant, & Behavioral Data

Aguinis et al. (2021); Burchett & Ben-Porath (2019); Rogers (2018) ²⁵

Identifying & Removing

- ◆ **Sample More than Desired Number**
- ◆ **Integrate Validity Scales**
 - ◆ Several long commercial measures have well-studied validity indicators / cut scores
 - ◆ *Explore research grants & discounts*
 - ◆ Promising shorter / free options, too

Burchett et al. (2023); Chandler et al. (2020); Goldammer et al. (2020) ²⁶

Omnibus Measure	Inconsistent Responding	Overreporting	Underreporting
MMPI-3* : Minnesota Multiphasic Personality Inventory-3 [335 items]	CRIN, VRIN, TRIN	F, Fp, Fs, FBS, RBS	L, K
PAI* : Personality Assessment Inventory [344 items]	ICN, INF	NIM, MAL, RDF	PIM, DEF, CDF
NEO PI-R* : NEO-Personality Inventory-Revised [240 items]	INC**	NPM**	PPM**
PID-5 : Personality Inventory for DSM-5 [220 items]	PID-5-INC**	PID-5-ORD**	PRD**, SD-TD**
SNAP-2 : Schedule for Nonadaptive and Adaptive Personality-2 [390 items]	VRIN, TRIN, BD	D	DRIN, RV
PPI-R* : Psychopathic Personality Inventory-Revised [154 items]	INC	DR	VR
TriPM : Triarchic Psychopathy Measure [58 items]	TAPIR		
EPA : Elemental Psychopathy Measure [178 items]	Infrequency		Virtue
BIDR : Balanced Inventory of Desirable Responding [40 items]			SDE, IM

Adapted from Burchett et al. (2023) – See Paper for Full Scale Names

*Commercially Sold; **Not Officially Adopted by Test Authors

Best Practices with Online Platforms

◆ There Are Known Data Quality Concerns

- ◆ Financial Incentive to Move Quickly

 - ◆ Financial > Fun, Interest, Self-Knowledge

- ◆ Automated Bots

- ◆ Careless Responding Higher than in College Student Samples

 - ◆ **9-46%** Attention Check Failure Rates)



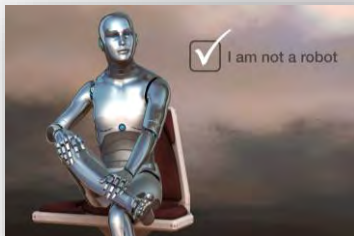
Agley et al., (2022); Barends & de Vries (2019); Burnette et al. (2022); Burchett & Ben-Porath (2019); 28
Chmielewski & Kucker (2020); Goldammer et al. (2020); Kim & Hodgins (2017); Paas et al. (2018)

Best Practices with Online Platforms

- ◆ Hire Highly Qualified Workers
 - ◆ \$\$\$; Possibly Reduced Effectiveness
 - ◆ Doesn't Detect All Poor-Quality Data
- ◆ Block Suspicious IP Addresses & Workers with Poor Data Quality from Coming Back
- ◆ Disallow Same IP Address More Than Once

Best Practices with Online Platforms

- ◆ Reduce Fabricated Screening Answers
 - ◆ Lower-Pay Pre-Screening Study → Invite to Higher-Pay Extensive Study
- ◆ Reduce Bots: Require CAPTCHA Verification



Chandler et al. (2020); Chandler & Paolacci (2017); Loepp & Kelly (2020) 30

Best Practices with Online Platforms

- ◆ Ask Open-Ended Questions on Engagement & Whether the Data are Useable
- ◆ Inform Them – Payment Not Withheld for Honesty
- ◆ Or, Warn of Screening; Offer Second Chances
- ◆ Not Naïve Samples; Online Communities Discuss Study Details & Screening Methods

Chandler et al. (2020); Cheung et al (2017);

Kim & Hodgins (2017); Meade & Craig (2012); Schmidt (2015) ³¹

Attention Check Items

- ◆ “Do This” or “Refrain from Answering This”
 - ◆ Obvious Items May Annoy
- ◆ Bogus Items - So Extreme, Most Answer the Same Way
- ◆ Open-Ended Qualitative Items / Writing Tasks
 - ◆ Coding Scheme; Inter-Rater Agreement
- ◆ Use 2-3 Attention Check Items to Avoid False Positives

Aguinis et al. (2021); Chandler et al. (2020); Curran & Hauser (2019);

Edwards (2019); Meade & Craig (2012) ³²

Response Latencies

- ◆ **The Length of Time to Complete a Measure**
- ◆ Can Help Identify Careless Responding
 - ◆ Evidence for Cut Score of ~1 to 5 Seconds/Item
- ◆ Evidence is Mixed on Detecting Feigning
 - ◆ Variability & Coaching
- ◆ More Research is Needed to Standardize Methods

Chandler et al. (2020); Dilchert & Ones (2012);

Goldammer et al. (2020); Huang et al. (2012); Meade & Craig (2012)

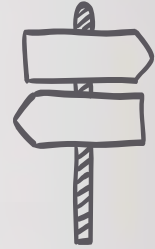
Inconsistency Item Pairs

- ◆ **A.K.A. Synonyms & Antonyms**
 - ◆ (Similar to VRIN or TRIN Validity Scales)
 - ◆ Matched Item Pairs
 - ◆ Direct Replication
 - ◆ Rational Selection
 - ◆ Empirical Selection
 - ◆ More Research is Needed to Standardize Methods

Statistical Approaches

- ◆ **Personal Reliability** – Low item-pair within-person correlations (e.g., even-odd items)
- ◆ **Mahalanobis distance** – Inconsistency with the sample's normative pattern (multivariate outliers)
- ◆ **Longstring** – String answered in same way; May help for fixed responding

5.



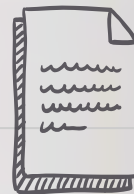
Next Steps for the Use & Creation of Strategies

Are Researchers Screening?



- ◆ Sharpe et al. (2022)
 - ◆ Reviewed 239 studies in personality disorder and clinical science journals (but not assessment journals)
 - ◆ Only 19% of studies tested for the validity of self-report data
 - ◆ Online (70%) vs. Other (14%)

Are Researchers Screening?



- ◆ **Sharpe et al. (2022) - Most Common:**
 - ◆ Formal Validity Scales
 - ◆ Attention Checks
 - ◆ Response Timing
 - ◆ Informal Inspections of Data
- ◆ **Let's More Frequently Screen!**

What Can Personality & Psychopathology Researchers Do?

- ◆ **Rely on Well-Studied Validity Scales**
 - ◆ Commercial Instruments
 - ◆ Free Measures
- ◆ **Follow Best Practices for Online Platforms**
 - ◆ Reduce Bots & Low-Quality Workers
 - ◆ Address Motivational Issues
 - ◆ Attention Checks, Response Times, & Developing Inconsistency Indices

“

...we recommend personality researchers using online platforms **use as many of these recommended (and particularly, validated) practices as possible, in conjunction with well-validated inconsistency, underreporting, and overreporting measures.** The chosen methods and cut scores should be established **a priori** (Curran, 2016).”

-Burchett, Sellbom, & Bagby (2023)

What Can Response Bias Researchers Do?

◆ Validate & Standardize Tools & Cut Scores!

- ◆ Develop scales/cut scores to identify inconsistent responding, underreporting, and overreporting on free commonly-used measures or batteries
- ◆ Study utility of prevention methods
- ◆ Study utility of multi-method approaches
- ◆ Attend to classification accuracy to set cut scores

Recap



Attendees will review the **negative impacts that invalid response styles have on self-report questionnaire data quality** and learn **strategies to prevent or remove low-quality data** before conducting analyses.

1. Types of Invalid Responding
2. Relevant Contexts
3. Impact on Research Conclusions
4. Strategies to Reduce Risks of Erroneous Conclusions
5. Next Steps for the Use & Creation of Strategies

Recommended Readings



Preventing and Identifying Poor Self-Report Data

Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2022). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior Research Methods*, 54(2), 885-897.

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, 47(4), 823-837.

Burchett, D., Sellbom, M., & Bagby, R. M. (2023). Assessment of response bias in personality disorder research. *Personality Disorders: Theory, Research, and Treatment*, 14(1), 93-104.

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5), 500-508.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, 32, 347-361.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4-19.

Recommended Readings



Preventing and Identifying Poor Self-Report Data, Continued

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.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27, 99-114.

Sharpe, B. M., Lynam, D. R., & Miller, J. D. (2022). Check and report: The state of data validity detection in personality disorder science. *Personality Disorders: Theory, Research, and Treatment*.

Zickar, M. J., & Keith, M. G. (2023). Innovations in sampling: Improving the appropriateness and quality of samples in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, 10.

Considerations for New Detection Strategies

Burchett, D., & Ben-Porath, Y. S. (2019). Methodological considerations for developing and evaluating response bias indicators. *Psychological Assessment*, 31(12), 1497-1511.

Reference List



Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2022). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior Research Methods*, 54(2), 885-897.

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, 47(4), 823-837.

Anderson, J. L., Sellbom, M., Wygant, D. B., & Edens, J. F. (2013). Examining the necessity for and utility of the Psychopathic Personality Inventory-Revised (PPI-R) Validity Scales. *Law and Human Behavior*, 37, 312-320.

Barends, A. J., & de Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences*, 143, 84-89.

Ben-Porath, Y. S. (2013). Self-report inventories: Assessing personality and psychopathology. In J. Graham & J. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 622-644). Wiley.

Burchett, D., & Bagby, R. M. (2014). Multimethod assessment of response distortion: Integrating data from interviews, collateral records, and standardized assessment tools. In C. Hopwood & R. Bornstein (Eds.), *Multimethod clinical assessment* (pp. 345-378). New York, NY: Guilford Press.

Burchett, D. L., & Ben-Porath, Y. S. (2010). The impact of overreporting on MMPI-2-RF substantive scale score validity. *Assessment*, 17(4), 497-516.

Reference List



Burchett, D., & Ben-Porath, Y. S. (2019). Methodological considerations for developing and evaluating response bias indicators. *Psychological Assessment*, 31(12), 1497-1511.

Burchett, D., Dragon, W. R., Smith Holbert, A. M., Tarescavage, A. M., Mattson, C. A., Handel, R. W., & Ben-Porath, Y. S. (2016). "False feigners": Examining the impact of non-content-based invalid responding on the Minnesota Multiphasic Personality Inventory-2 Restructured Form content-based invalid responding indicators. *Psychological Assessment*, 28(5), 458-470.

Burchett, D., Sellbom, M., & Bagby, R. M. (2023). Assessment of response bias in personality disorder research. *Personality Disorders: Theory, Research, and Treatment*, 14(1), 93-104.

Burnette, C. B., Luzier, J. L., Bennett, B. L., Weisenmuller, C. M., Kerr, P., Martin, S., Keener, J., & Calderwood, L. (2021). Concerns and recommendations for using Amazon MTurk for eating disorder research. *International Journal of Eating Disorders*, 55(2), 263-272.

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5), 500-508.

Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, 129(1), 49-55.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, 32(4), 347-361.

Reference List



Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study result Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4-19.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4-19.

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality*, 82, 103849.

Dhillon, S., Bagby, R. M., Kushner, S. C., & Burchett, D. (2017). The impact of underreporting and overreporting on the validity of the Personality Inventory for DSM-5 (PID-5): A simulation analog design investigation. *Psychological Assessment*, 29(4), 473.

Dilchert, S., & Ones, D. S. (2012). Application of preventive strategies. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 177-200). Oxford University Press.

Edwards, J. R. (2019). Response invalidity in empirical research: Causes, detection, and remedies. *Journal of Operations Management*, 65(1), 62-76.

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.

Reference List



Handel, R. W., Ben-Porath, Y. S., Tellegen, A., & Archer, R. P. (2010). Psychometric functioning of the MMPI-2-RF VRIN-r and TRIN-r scales with varying degrees of randomness, acquiescence, and counter-acquiescence. *Psychological Assessment*, 22, 87-95.

Huang, J. L., Curran, P. G., Keeney, J., Popski, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business Psychology*, 27, 99-114.

Kim, H., & Hodgins, D. C. (2017). Reliability and validity of data obtained from alcohol, cannabis, and gambling populations on Amazon's Mechanical Turk. *Psychology of Addictive Behaviors*, 31(1), 85-94.

Loepp, E., & Kelly, J. T. (2020). Distinction without a difference? An assessment of MTurk worker types. *Research and Politics*, 7(1), 2053168019901185.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437-455.

Paas, L. J., Dolnicar, S., & Karlsson, L. (2018). Instructional manipulation checks: A longitudinal analysis with implications for MTurk. *Instructional Journal of Research in Marketing*, 35(2), 258-269.

Rogers, R. (2018). Detection strategies for malingering and defensiveness. In R. Rogers & S. D. Bender (Eds.) *Clinical assessment of malingering and deception* (pp. 18-41).

Reference List



Schmidt, G. B. (2015). Fifty days an MTurk worker: The social and motivational context for Amazon Mechanical Turk workers. *Industrial and Organizational Psychology*, 8(2), 165-171.

Sellbom, M., & Bagby, R. M. (2008). The validity and utility of the positive presentation management and negative presentation management scales for the Revised NEO Personality Inventory. *Assessment*, 15(2), 165-176.

Sharpe, B. M., Lynam, D. R., & Miller, J. D. (2022). Check and report: The state of data validity detection in personality disorder science. *Personality Disorders: Theory, Research, and Treatment*.

Whitman, M. R., Tylicki, J. L., & Ben-Porath, Y. S. (2021). Utility of the MMPI-3 Validity Scales for detecting overreporting and underreporting and their effects on substantive scale validity: A simulation study. *Psychological Assessment*, 33(5), 411-426.

Zickar, M. J., & Keith, M. G. (2023). Innovations in sampling: Improving the appropriateness and quality of samples in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, 10 315-337.

THANKS!

Any questions?

Please email me at
dburchett@csumb.edu

Special thanks to those who released these
resources for free:

Presentation template by [SlidesCarnival](#)

Photographs by [Unsplash](#), [Pixabay](#), & [Pexels](#)

