Big Data and the Challenge of Construct Validity

Michael T. Braun
Virginia Tech

Goran Kuljanin
DePaul University

One important issue not highlighted by Guzzo, Fink, King, Tonidandel, and Landis (2015) is that simply establishing construct validity will be significantly more challenging with big data than ever before. One needs to only look as far as the other social sciences analyzing big data (e.g., communications, economics, industrial engineering) to observe the difficulty of making valid claims as to what measured variables substantively "mean." This presents a significant hurdle in the application of big data to organizational research questions because of the critical importance of demonstrating validity in the organizational sciences as highlighted by Guzzo et al.

Construct validity represents the extent to which a test (i.e., operationalization) measures what it purports to (Cronbach & Meehl, 1955). In other words, it evaluates how well a physical representation (i.e., measurement) relates to the psychological or conceptual construct of interest. Typically, researchers obtain construct validity through a process of careful measure construction combined with sufficient statistical justification. This process has been extensively vetted for evaluating the validity of data commonly collected in the organizational sciences (e.g., Bagozzi, Yi, & Phillips, 1991); however, it has not yet been evaluated for the type of data commonly observed in "big" datasets.

Creating Construct Valid Measures

The first steps toward measuring a construct in a valid manner are to understand the construct of interest and to design a measure that is thought to best represent that construct. Traditionally, when researchers are designing a measure or operationalization for a construct, great care is taken to maximize overlap between the measure and the construct while simultaneously minimizing both construct contamination (i.e., irrelevant material included) and construct deficiency (i.e., relevant material excluded; McDonald, 1999). This process is methodologically rigorous and involves a variety of activities such as consulting subject matter experts, pilot testing, and running a plethora

Michael T. Braun, Department of Psychology, Virginia Tech; Goran Kuljanin, Department of Psychology, DePaul University.

Correspondence concerning this article should be addressed to Michael T. Braun, who is now at Department of Psychology, University of South Florida, 4151 PCD, Tampa, FL 33620. E-mail: mtbraun@usf.edu

https://doi.org/10.1017/iop.2015.77 Published online by Cambridge University Press

of statistical analyses (Shadish, Cook, & Campbell, 2001). Alternatively, the volume, variety, and velocity of big data frequently accompany a lack of control over how variables are measured and what variables are included in a dataset. As a result, researchers utilizing big data are forced to establish construct validity using measures they likely did not design and that were never intended to represent the construct of interest.

As alluded to by Guzzo et al., big data in the organizational sciences likely originates from three primary sources with differing degrees of researcher control. At one extreme are naturally occurring behavioral trace data that are readily available from the Internet (e.g., Hernandez, Newman, & Jeon, 2015). Facebook, Twitter, Wikipedia, and others are treasure troves of data related to individuals interacting with one another over time. However, to exploit their potential, researchers must first determine how to convert and combine "likes," friend requests, retweets, edits, and other such Internet behavior into meaningful organizational constructs. The second source of big data is organizations that already track numerous employee behaviors and ratings (e.g., Caputo & Boyce, 2015). E-mails, chat logs, performance reviews, and other data are routinely collected by organizations but rarely utilized to better understand organizational processes. To do so, researchers must first properly synthesize text and code behavioral patterns into commonly studied constructs such as leadership, teamwork, and organizational citizenship behaviors. Finally, the third source of big data represents the other extreme whereby organizational researchers utilize physiological sensors or computer-based laboratory simulations to obtain large quantities of data on a predetermined set of individuals (e.g., Kennedy & McComb, 2014). As a result, researchers control exactly what is measured and at what frequency. Unfortunately, these data likely still require researchers to convert button clicks, mouse movements, location, vocal amplitude, heart rate, and other raw behaviors into something substantively interesting and organizationally relevant. Therefore, overall, each of these big data sources presents differing challenges for understanding and interpreting what was measured and for combining and converting them into valid operationalizations of organizational constructs.

Statistically Establishing Construct Validity

Once a measure of a construct has been created, data are collected with that measure and statistical analyses are used in an attempt to establish construct validity. There are two primary means by which organizational researchers traditionally statistically demonstrate construct validity.[1]

[1] This is not intended to be an exhaustive list as there are additional statistical approaches for establishing construct validity (e.g., Westen & Rosenthal, 2003).

https://doi.org/10.1017/iop.2015.77 Published online by Cambridge University Press

Currently, the dominant approach is to assess the internal structure of a measure using confirmatory factor analysis (CFA). In this approach researchers test an a priori factor structure to determine the overall model fit as well test the hypothesized factor structure against plausible alternatives (Widaman, 1985). If the measure meets the common standards for goodness of fit established by Hu and Bentler (1999) as well as fits the data better than the plausible alternatives according to chi-squared difference tests or likelihood ratio tests, then the measure is thought to be construct valid.[2] Alternatively, historically researchers would utilize a multitrait–multimethod matrix (MTMM) to establish construct validity through a pattern of convergent and discriminant validity evidence (Campbell & Fiske, 1959). Researchers gather data from multiple methods on the same construct as well as on distinct constructs and compare the relative magnitude of the resulting correlations. Typically, to accompany a new scale or measurement, the researcher will choose well-established measures with which to compare. This is of particular relevance to big data because researchers can accompany the behavioral trace data that commonly compose "big" datasets with more traditional self-report measures as suggested by Guzzo et al. Construct validity is established when measures of the same construct measured with different methods (e.g., self-report and behavioral trace) correlate more strongly with one another (i.e., convergent validity) compared with measures of different constructs measured with the same method (e.g., all self-report) and different (e.g., self-report and behavioral trace) methods (i.e., discriminant validity).

Unfortunately, the properties (i.e., volume, variety, velocity) of big data make the use of either of these statistical approaches for establishing construct validity significantly more challenging. CFA typically requires at least three indicators for each construct (Anderson & Rubin, 1956). This minimum is easily achieved when developing a questionnaire with multiple items; however, in big data where behavioral traces are a more common variable type, having at least three behaviors to represent each construct is more difficult to achieve. If less than three behavioral indicators exist for a construct, the CFA model is not identified and cannot be estimated. Even when sufficient indicators exist, behavior is likely to occur at different frequencies, resulting in a large quantity of missing data for some variables that change or are recorded at slower rates. It is well known that missing or incomplete data may result in inaccurate correlational estimates (Roth, 1994), making factor loadings unstable. One proposed advantage of CFA

[2] It is important to note that no single study proves construct validity, but rather, construct validity is established through the accumulation of evidence across multiple studies (Peter, 1981).

https://doi.org/10.1017/iop.2015.77 Published online by Cambridge University Press

over means of establishing construct validity is the ability to utilize goodness-of-fit statistics (indices) to quantify the degree to which construct validity has been achieved. Unfortunately, the chi-square statistic tends to be overpowered in large samples, resulting in apparent poor model fit and biased fit indices that are based on the chi-square value (Zwick & Velicer, 1986). In addition, when a small number of indicators per factor are used in large datasets, fit indices tend to be upwardly biased, resulting in an overestimate of model fit (Marsh, Hau, Balla, & Grayson 1998). Because "big" datasets by definition have large numbers of records (i.e., volume) as well as rather few indicators per construct, the fit statistics estimated in CFA are likely unreliable.

Like utilizing CFA, using an MTMM to establish construct validity also becomes problematic with big data. Even in the ideal case where a researcher is able to control the behavioral trace data collected as well as include self-report survey or other established methods of measuring a construct (as recommended by Guzzo et al.), the MTMM is unlikely to produce the familiar patterns of correlations for convergent and discriminant validity. Behavioral trace data are almost always collected at a significantly higher frequency than typical survey measurement (i.e., velocity). In addition, surveys are typically administered using a Likert response scale, which produces relatively limited variability made even more restricted by common rating biases such as the halo effect (Thorndike, 1920). Alternatively, behavioral measures tend to be on wildly different scales with massive variance. Correlations among variables measured at wildly different frequencies with differing amounts of variability will inevitably be rather low. Therefore, the resulting correlations among measures of the same construct with different methods (i.e., convergent validity) will likely be far lower than is typically observed. As a result, it is significantly more likely that different constructs measured with the same method (i.e., discriminant validity), especially self-report surveys, will demonstrate higher correlations than the same construct measured with different methods. This pattern of results would violate the conditions for establishing construct validity according to Campbell and Fiske (1959). Under other conditions of less researcher control such as using behavioral trace date from the Internet or organizationally collected data, it may not even be possible to utilize an MTMM to demonstrate construct validity due to lack of multiple modes of measurement or lack of an established measure to compare against.

A Path Forward

Given the challenges that big data presents with regard to establishing construct validity using traditional approaches, authors, reviewers, and editors will likely need to adapt their requirements for demonstrating validity if the organizational sciences are going to join the big data revolution. Rather than

https://doi.org/10.1017/iop.2015.77 Published online by Cambridge University Press

a strict reliance on statistical standards, authors will need to justify the validity of a measure utilizing a theoretical approach combined with demonstrating conceptual overlap with other, more traditional, measures of the construct of interest. For instance, an author could use theory to describe the type of behavior and behavioral pattern that would be expected for a construct and highlight how items from an established scale are designed to measure the frequency or presence of the behavior. To further provide support, authors could employ subject matter experts (SMEs) to rate the relevance of behavioral trace variables for measuring a construct of interest and then test the consistency of SME ratings against established standards for test development (e.g., Lawshe, 1975). This process is a proven and accepted practice for validating psychological and education tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), and it could prove invaluable for establishing construct validity when other statistical estimates using big data do not meet current standards. It is important to recognize that behavior is inherently context specific, and thus, the specific instantiation of a behavior will vary across studies and should be identified and justified as is recommended for creating behavioral anchored rating scales (BARS; Schwab, Heneman, & DeCotiis, 1975; Smith & Kendall, 1963). Also, because no single study can establish construct validity (Peter, 1981), authors should be allowed latitude so that a preponderance of evidence across situations can be collected and evaluated to determine whether a behavior is in fact indicative of the construct of interest. Likewise, gathering statistical evidence on the relationships between behavioral trace and self-report measures will also help serve as a baseline for potentially adjusting commonly accepted standards for evaluating the size of a correlation in behavioral science (e.g., Cohen, 1988), making the use of CFA and MTMMs in the future more appropriate.

Another potential avenue forward for utilizing big data in spite of the challenges of statistically establishing construct validity is to focus more on process. McGrath (1964) created the input–process–output model of performance that has dominated organizational thinking and research for over 50 years.[3] Historically, organizational science has been much better at identifying, measuring, and understanding the inputs and outputs rather than the process. This likely stems from the dominance of cross-sectional designs and self-report measurement. Process is inherently a complex dynamic phenomenon that is not easily studied or understood from a static perspective. Although there are excellent theoretical articles outlining

[3] More recently this has been replaced by the input–mediator–output–input (IMOI) model (Ilgen, Hollenbeck, Johnson, & Jundt, 2005).

https://doi.org/10.1017/iop.2015.77 Published online by Cambridge University Press

critical processes for teams (e.g., Marks, Mathieu, & Zaccaro, 2001), leaders (e.g., Morgeson, DeRue, & Karam, 2010), and others, little empirical evidence exists to test them. Big data presents an opportunity to begin to test these theories and unpack the black box of process by examining the behaviors that individuals, teams, and organizations engage in over time.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Newman (Ed.), *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability: Vol. 5. Contributions to econometrics, industrial research, and psychometry* (pp. 111–150). Berkeley, CA: University of California Press.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Construct validity in organizational research. *Administrative Science Quarterly*, *36*, 421–458.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Caputo, P., & Boyce, A. (2015, April). *Data science in human capital research and analytics*. Symposium presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Guzzo, R. A., Fink, A. A., King, E., Tonidandel, S., & Landis, R. S. (2015). Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*(4), 491–508.

Hernandez, I., Newman, D., & Jeon, G. (2015). Twitter analysis: Methods for data management and validation of a word count dictionary to measure city-level job satisfaction. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology*. New York, NY: Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology*, *56*, 517–543.

Kennedy, D. M., & McComb, S. A. (2014). When teams shift among processes: Insights from simulation and optimization. *Journal of Applied Psychology*, *99*, 784–815.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*, 356–376.

Marsh, H. W., Hau, K., Balla, J. R., & Grayson, D. (1998). Is more ever too much: The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181–220.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

https://doi.org/10.1017/iop.2015.77 Published online by Cambridge University Press

McGrath, J. E. (1964). *Social psychology: A brief introduction*. New York, NY: Holt, Rinehart and Winston.
Morgeson, F. P., DeRue, S., & Karam, E. P. (2010). Leadership in teams: A functional approach to understanding leadership structures and processes. *Journal of Management*, *36*, 5–39.
Peter, J. P. (1981). Construct validity: A review of basic issues and marketing practices. *Journal of Marketing Research*, *18*, 133–145.
Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*, 537–560.
Schwab, D., Heneman, H., & DeCotiis, T. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, *28*, 549–562.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
Smith, P. C., & Kendall, L. M. (1963). Re translation of expectations: An approach to the construction of unambiguous anchors to rating scales. *Journal of Applied Psychology*, *47*, 149–155.
Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 25–29.
Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*, 608–618.
Widaman, K. F. (1985). Hierarchically nested covariance models for multitrait–multimethod data. *Applied Psychological Measurement*, *9*, 1–26.
Zwick, W. R., & Velicer, W. F. (1986). A comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *22*, 432–442.

Big Data, Little Individual: Considering the Human Side of Big Data

Michael N. Karim
Fors Marsh Group, LLC, Arlington, Virginia

Jon C. Willford and Tara S. Behrend
The George Washington University

Guzzo, Fink, King, Tonidandel, and Landis (2015) provide a clear overview of the implications of conducting research using big data. One element we believe was overlooked, however, was an individual-level perspective on big data; that is, what impact does this sort of data collection have on the individuals being studied? As psychologists, the ethics and impact of big data

Michael N. Karim, Fors Marsh Group, LLC, Arlington, Virginia; Jon C. Willford and Tara S. Behrend, Department of Organizational Sciences and Communication, The George Washington University.

Correspondence concerning this article should be addressed to Michael N. Karim, Fors Marsh Group, LLC, 1010 North Glebe Road, Number 510, Arlington, VA 22201. E-mail: mkarim@forsmarshgroup.com

https://doi.org/10.1017/iop.2015.77 Published online by Cambridge University Press