Beyond Agreement: Rethinking the Preeminence of Likert-Type Scales in Quantitative Research

Abhilash Ponnam,

Associate Professor, Department of Management,

Central University of Andhra Pradesh – Anantapuramu, India

Abstract

Likert-type scales are currently the most popular type of measurement instrument for latent constructs in quantitative research. While these instruments have been developed to measure attitudes and opinions, they are ubiquitously employed to define variables like behaviors, competencies, organizational performance, innovation, resilience, and leadership. Their popularity is due to ease of measurement and fit with statistical analyses, but they may inadvertently lead to an important methodological flaw — measuring and validating them indiscriminately. This paper posits that the weakness is not with Likert scales per se, but with applying them as one-size-fits-all instruments regardless of construct definition. This causes construct-measurement fit to be sacrificed for ease of measurement, leading to potential weaknesses in construct validity, susceptibility to common method and social desirability bias, and overly simplistic representations of complex, dynamic phenomena that are captured only as simple agreement. This paper proposes a Construct-Measurement Fit Framework to avoid inappropriate measurement in quantitative research and presents several alternative measurement paradigms to illustrate that methodological choice can enhance validity.

Keywords: Likert Scale, Construct Validity, Survey Research, Measurement Theory, Common Method Bias, Quantitative Research

1. Introduction

Perhaps more than any other measurement tool, the Likert scale is central to quantitative research design. Since its inception by Likert (1932) it has become iconic to survey research and provides a seemingly easy way to study latent psychological constructs. Throughout many disciplines such as management, psychology, education, marketing, health, and IS, thousands of researchers use some variety of the five- or seven-point agreement scale to define and operationalize latent variables. Their ubiquity is easily explained. Agreement scales are easy to administer, inexpensive, familiar to respondents, and can be straightforwardly

incorporated into all the modern statistical software, factor analyses, SEM, and PLS SEM modeling. Likert type items thus have become a default measurement device in many quantitative studies.

Arguably, the ubiquity of agreement scales is one of the social sciences' greatest success stories. Since their advent they have provided a systematic method of measurement, allowed comparisons across studies, and stimulated development of valid psychometric scales. Thousands of psychological and organizational constructs are defined and measured by hundreds of validated multi-item scales; allowing researchers to test theoretically complex models with greater efficiency than before. The creation and availability of validated instruments have likely decreased methodological uncertainty, spurred on cumulative research efforts and facilitated replication.

However, one's success can inadvertently lead to over-application. Scales that were originally developed to measures attitudes and opinions are now pervasively applied to theoretically very different constructs such as behaviors (e.g., are you a good team member?), competence (e.g., I am a confident programmer), organizational performance (e.g., I achieve organization goals). Even research on the phenomena of innovation, resilience, leadership relies heavily on employee agreement with behavior or performance descriptions. But if respondents simply report their agreement with an statement, does this truly measure that construct?

Measurement theory tells us construct validity is not simply just statistical reliability, but also whether an instrument captures what it is supposed to (Messick, 1995; Borsboom et al., 2004). Good internal consistency, good factor loadings and a fit for any statistical analysis does not necessarily mean that the researcher is measuring the right construct. A latent construct could be measured by a valid scale, which nonetheless gives very poor and unreliable measurement. Accordingly, valid measurement requires that researchers think beyond the mere statistical estimation and ensure measurement has content validity. The question is not whether we can reliably measure something using an agreement scale; it is whether we are measuring the correct thing.

This paper contends that quantitative research suffers from a measurement monoculture, relying so heavily on agreement scales for conceptually different traits. We contend that it is essential to consider the nature of constructs before choosing and validating instruments. The paper subsequently considers why consensus scales became so prevalent, examines some problems of over-reliance upon them, and offers a Construct-Measurement Fit Framework to address the problem by presenting seven measurement scales that can address the need for a breadth of valid measurements to capture the richness of phenomenon that quantitative researchers seek to study.

2. The Triumphant Dominance of the Likert Scale: A Methodological Triumph Becomes a Default Mechanism

The immense success of the Likert scale can neither be considered a mistake nor unearned. For almost one century, it has provided researchers with a workable, standardized, psychometrically sound way of measuring latent psychological phenomena. As compared to direct observation, interviews, or behavior tasks, Likert-type questionnaires enable researchers to gather a large number of standardized measurements within a relatively short amount of time and at relatively low costs.

Easy to administer, readily understood by participants, and stat-package compatible, the Likert-type questionnaire has become a prerequisite in many empirical fields. Measurement relying on agreement has therefore become ingrained within quantitative measurement practice in those fields where personality, attitudes, or beliefs are the main variables, such as in most of the disciplines in the social sciences (Likert, 1932; DeVellis & Thorpe, 2021). Quantitative methods of data analysis have only increased this dominance. Contemporary statistical procedures, such as confirmatory factor analysis (CFA), covariance based structural equation modeling (CB-SEM), and partial least squares structural equation modeling (PLS-SEM), require multiple observable indicators to estimate latent constructs.

Likert-type items can fulfill this requirement quite efficiently; they produce more or less continuous like variables which are amenable to the computation of both reliability and validity indices. Cronbach's Alpha, Composite Reliability (CR), average variance extracted (AVE), discriminant validity and model fit indices are all indispensable components of empirical studies, and have brought forth methodological criteria in which the Likert scale works quite naturally. As methods of analysis advance, they are matched by instrument measurement procedures; the agreement-based questionnaire becomes a more and more natural way of measuring variables amenable to advanced statistical modeling procedures (Hair et al., 2021).

University practices also explain why such over-reliance on the Likert scale occurs. In most doctoral programs, students are encouraged to adopt already tested instruments instead of developing alternative ways of measuring phenomena. This strategy has obvious benefits: the reliability and validity of existing instruments have been established, they allow comparison across studies, and reduce the uncertainty involved in survey development. Journal reviewers tend to prefer using established measures and therefore are more likely to accept a manuscript containing such measurements.

While these factors explain why Likert-scale measurement has developed, they do not necessarily explain why researchers prefer the Likert scale to other measurement tools. The Likert scale was developed as a means of measuring attitudes, and it does that quite efficiently. Nevertheless, due to the above factors, the Likert scale has become one in a series of appropriate measurements and is now considered the default choice. Researchers don't ask, how should this construct be measured anymore, but instead, which validated Likert scale already exists?

The choice of measurement has turned from being based on the concept to being based on what measure is available. Accordingly, operationalization increasingly looks like adapting existing survey items rather than deciding how the concept could best be measured. Nowhere is this more evident than in contemporary organizational and management research. Constructs as diverse as innovation capability, organizational resilience, employee performance, digital transformation, ethical leadership, knowledge sharing, organizational learning, customer orientation and sustainability practices are all measured using virtually identical 5- or 7-point Likert-type scales with the same questionnaire administered to the same set of respondents in one wave of data collection.

Although inherently different in nature, each of these concepts relies on the exact same method of measurement which is, implicitly, that the researcher can obtain evidence about any concept by using Likert-type questionnaire items completed by the subject under study. This method implies that the response option "agree" for all of these concepts represents the same empirical evidence. For a long time, researchers have argued that there are several forms of evidence that a researcher can use to demonstrate whether a concept truly covers some aspects (Messick, 1995; Borsboom et al., 2004). Beliefs are inherently about what individuals think is true. Attitudes are about preferences. Perceptions are about what individuals perceive of themselves and the world around them. Behaviors are about the actions taken by individuals. Competencies are the demonstrated skills or abilities within specific domains. Organizational characteristics exist in and of themselves (or should reflect the reality of the organizational state), regardless of employee beliefs, attitudes, perceptions and behaviors. Dynamic psychological states may change from situation to situation or across time.

Why would all these intrinsically different constructs be represented using identical response formats? Why does each researcher assume that agreement automatically implies measuring behavior, personality, competence or individual preferences, or some other aspects, or the whole set of aforementioned constructs? By increasing reliance on the Likert scale for all these constructs, this leads to a loss of correspondence between theory and measurement for some aspects of the study. Likert scale measurements, developed to quantify beliefs and attitudes, have become universal measurement tools. We often mistake employees' opinions or perspectives of a situation, event, person, or issue for the object itself. The difference between

what people believe/think/perceive and what really is and, importantly, what employees can do, appears to be erased within a single questionnaire. This difference is, of course, most crucial when researching matters of fact and behavioral performance.

This argument suggests that the extensive use of the Likert scale has potentially resulted in construct-measurement misfit: This is the situation where measurement procedures are driven by, to some extent, convention, and not by the specific demands of the concept(s). Researchers probably fail to detect this condition because reliability and factorial validity refer to the internal consistency of the measures and do not represent an indication of how well the proposed construct captures the measured phenomenon. The research must continue with questioning whether it makes sense to measure anything by using a Likert-type scale and explore the situations where it does not apply.

3. Construct-Measurement Mismatch: How Likert Scales Can Become Methodologically Unsound

One of the primary factors determining how effective an instrument will be at capturing a phenomenon is its correspondence with the phenomenon it represents, irrespective of its psychometric properties. Validity hinges on whether a measure does what it claims to do and captures what it is meant to capture (Messick, 1995; Borsboom et al., 2004). Therefore, the highly reliable Likert scale is not necessarily the appropriate instrument simply because the Likert survey elicits response consistency. The choice of a response scale that cannot accommodate the conceptual nature of a construct leads to construct-measurement misfit. That is, even sophisticated statistical manipulation of a poorly chosen response format cannot ameliorate conceptually flawed measurement.

Indeed, these items ask participants to do little more than offer an opinion about their behavior based on their memory or impression of it, not their actual frequency of engagement in the behavior. As such these types of items are subject to such sources of error as self-preservation, bias, limitations of memory, and a differential understanding of what is encompassed by a frequency term such as "frequently" (Podsakoff et al., 2003). While at first glance agreement might appear correlated with behaviors that will be observed, there may still be many ways one can implement the target behaviors at very different levels or at some significant degree. Another is the assessment of performance: “Employee, organizational, or team performance” has been and continues to be widely measured using agreement-based self-report survey methods, such as Likert scale items (e.g., "I get all the work that needs to be done"; "I maintain productivity"). It is apparent, however, that such measures do little more than tap the respondent’s beliefs

or perceptions of how they perform or achieve. Moreover, employees' beliefs that their work leads to success are often influenced by subjective factors such as ego and a desire to appear capable on questionnaires to employers or supervisors, thus contributing to common method bias (CMB).

CMB is a problematic method variance that can occur when independent, dependent, moderator, and mediator variables are measured using identical survey items (Podsakoff et al., 2003). It should be noted that the procedures/statistical techniques recommended to remedy CMB can only alleviate the consequences of having a common measurement source, not eliminate its origin. Using multiple methods of gathering data, in as varied modes as possible, is thought to be the best approach to reducing common method variance. In addition, there is social desirability bias. Although agreement has been a popular choice to measure various phenomena, researchers should consider whether the phenomena in question lend themselves well to measurement using an agreement format. The likelihood is increased in the organizational setting where workers want to appear successful in the eyes of managers and the company for better career development. Thus, many respond with high levels of agreement that reflect what they think people expect them to do (Podsakoff et al., 2003).

4. The Construct-Measurement Fit (CMF) Framework: Moving Beyond "One-Size-Fits-All"

The goal of quantitative research is to increase our scientific knowledge, and moving beyond the issue of scale reliability to scale fit will go a long way. We are not trying to replace Likert scales with other measurement tools per se; we are trying to encourage the selection of methods based upon theory rather than statistics. The proposal that CMF applies also has implications for doctoral students, journal editors and peer reviewers alike. Researchers should be pushed beyond mere reliability testing of their measures to stating clearly how their measures fit the nature of the constructs they hypothesize exist. Similarly, reviewers should consider construct-measurement fit. If it is easier to use Likert scales to measure a behavior, it is not better; researcher-driven measures that have good construct-measurement fit should be preferred over existing Likert scales when they better capture the phenomena being investigated. Therefore, these are not competing methodologies to each other; quite the contrary, they should all fit within the broad category of methodologies to be used as they fit the nature of the construct. What is being proposed is not that one measure replace another as the status quo, but that an adequate tool be chosen for a purpose rather than an inappropriate tool used based on habit.

One cannot adequately assess preferences with a Likert-type instrument because people do not express preferences via agreement; such tasks should be performed using methods such as ranking exercises, Best-Worst Scaling, pairwise comparisons, forced-choice problems, or conjoint analysis; these require

individuals to make hard trade-offs which can provide a more direct assessment of preferences in such contexts (Louviere et al., 2015). However, some constructs also are temporary state constructs and may not be stable within the individual over time; stress, engagement, fatigue, emotions, or motives are temporary. Assessing such states over time may be best done via longitudinal survey, diary methods, Experience Sampling Method (ESM), or Ecological Momentary Assessment (EMA), which capture changes in the states of individuals over time (Beal, 2015). Likewise, performance measures or outcomes should preferably be measured with objective instruments, or an assessment by another person. Job performance may be assessed using financial indicators, productivity data or records, outcomes, customer or employee retention rates, or operational data. Self-reporting about behavior can be appropriate for exploratory purposes or if no other measure is available; however, the researcher should state up front that the measurement represents perceived, not actual performance (Bommer et al., 1995; Pransky et al., 2006).

A behavioral construct, on the other hand, cannot be assessed on the basis of agreement; actions are performed, there is a behavior, not whether respondents agree that they frequently perform the behavior. Instead of asking if they agree that they exhibit innovative behavior, researchers should consider the frequency, latency or recency of behavior, event counts, objective measures, logs, or existing archives. These methods provide evidence closer to the actual definition of behavior (King et al., 2024). Attitudes, perceptions, beliefs, or subjective evaluative judgments are all natural candidates for Likert-type scales. These constructs are defined within the respondent's cognitive or affective domain; therefore, survey methods are appropriate. Constructs such as job satisfaction, organizational commitment, perceived organizational support, customer satisfaction and perceived usefulness are appropriate uses of agreement scales because the respondent's subjectivity is the construct of interest. (Messick, 1995; DeVellis & Thorpe, 2021)

5. Conclusion

We argue that the pervasive use of agreement-based measurement has led to a construct-measurement mismatch; our choices about measurement are thus driven by convenience and precedent rather than by conceptual correspondence. Concepts such as behaviors, performance, organizations as outcomes, competence, and dynamic psychological states are commonly operationalized using agreement scales that were originally designed for more subjective attitudes. Although the scales may have acceptable psychometric properties, high reliability should never be mistaken for high validity. An instrument that registers a high internal consistency may still fail to accurately reflect the phenomenon that is meant to be tapped by such a scale.

The proposed Construct-Measurement Fit Framework suggests an alternative approach; rather than deciding if an existing Likert scale can be modified to an existing construct, researchers ought to initially consider the fundamental nature of the construct itself, then decide on the appropriate method from which best evidence regarding the construct will be attained. Likert scales are sometimes the appropriate method; other times other means will provide a more conceptually relevant measure. This alternative method may involve measuring performance via a frequency count, using objective measures to define the phenomenon, dynamic assessment via experience sampling, ranking scales, forced choice questionnaires, archival data, or a combination of such measures. The choice of method must thus arise from theory rather than convenience. The importance of such a framework has consequences beyond just developing an instrument. It means that training for doctoral students should increasingly focus on measurement theory and less on merely adapting an existing scale, and journals and reviewer boards should be charged with evaluating not just statistics but the correspondence between construct and method. Journals and reviewers ought to welcome variety when traditional instruments are not conceptually attuned to the phenomenon being investigated.

Future research must systematically study the occurrence of this construct-measurement misfit across studies and whether conceptually aligned measurement paradigms yield evidence with stronger validity and theoretical importance. Researchers may even begin to develop software that enables authors to justify their use of a particular measurement methodology. The core element of an advanced measurement procedure will thus be more theory. Quantitative research is progressing not only through sophisticated statistical analysis, but also through advances in how evidence is obtained. The continuing prevalence of the Likert scale should not prevent innovation in measurement, or prevent thoughtful consideration of measurement alternatives. Instead, the Likert scale should be embraced as but one tool among many available to the researcher. Progress in quantitative research will not be driven by the refusal to use agreement scales, but by using them when appropriate and complementing them with measures that possess a greater conceptually attuned nature when necessary.

References

King, D. L., Billieux, J., & Delfabbro, P. H. (2024). Red box, green box: A self-report behavioral frequency measurement approach for behavioral addictions research. Journal of Behavioral Addictions, 13(1), 21–24. https://doi.org/10.1556/2006.2023.00079

Beal, D. J. (2015). ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. Annual Review of Organizational Psychology and Organizational Behavior, 2, 383–407. https://doi.org/10.1146/annurev-orgpsych-032414-111335

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. Personnel Psychology, 48(3), 587–605. https://doi.org/10.1111/j.1744-6570.1995.tb01772.x

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological Review, 111(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

DeVellis, R. F., & Thorpe, C. T. (2021). Scale development: Theory and applications (5th ed.). SAGE Publications.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2021). A primer on partial least squares structural equation modeling (PLS-SEM) (3rd ed.). SAGE Publications.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140), 1–55.

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). Best-worst scaling: Theory, methods and applications. Cambridge University Press.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, 88(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Pransky, G., Finkelstein, S., Berndt, E., Kyle, M., Mackell, J., & Tortorice, D. (2006). Objective and self-report work performance measures: A comparative analysis. International Journal of Productivity and Performance Management, 55(5), 390–399. https://doi.org/10.1108/17410400610671426