

Volume 5

Issue 3 *Advanced Technologies in Assessment:
A Science–Practice Concern*

Article 6

2019

Crowdsourcing Job Satisfaction Data: Examining the Construct Validity of Glassdoor.com Ratings

Richard N. Landers

University of Minnesota, rlanders@umn.edu


Robert C. Brusso

Capital One, robertconrad.brusso@capitalone.com

Elena M. Auer

University of Minnesota, auer0027@umn.edu

Follow this and additional works at: <https://scholarworks.bgsu.edu/pad>

 Part of the [Human Resources Management Commons](#), [Industrial and Organizational Psychology Commons](#), and the [Other Psychology Commons](#)

Recommended Citation

Landers, Richard N.; Brusso, Robert C.; and Auer, Elena M. (2019) "Crowdsourcing Job Satisfaction Data: Examining the Construct Validity of Glassdoor.com Ratings," *Personnel Assessment and Decisions*: Vol. 5 : Iss. 3 , Article 6.

DOI: 10.25035/pad.2019.03.006

Available at: <https://scholarworks.bgsu.edu/pad/vol5/iss3/6>

This Research Article is brought to you for free and open access by the Journals at ScholarWorks@BGSU. It has been accepted for inclusion in Personnel Assessment and Decisions by an authorized editor of ScholarWorks@BGSU.

CROWDSOURCING JOB SATISFACTION DATA: EXAMINING THE CONSTRUCT VALIDITY OF GLASSDOOR.COM RATINGS

Richard N. Landers¹, Robert C. Brusso², and Elena M. Auer¹

1. University of Minnesota

2. Capital One

ABSTRACT

KEYWORDS

glassdoor, validation, job satisfaction, construct validity, crowdsourcing, web scraping, API, ratings

Researchers, practitioners, and job seekers now routinely use crowdsourced data about organizations for both decision-making and research purposes. Despite the popularity of such websites, empirical evidence regarding their validity is generally absent. In this study, we tackled this problem by combining two curated datasets: (a) the results of the 2017 Federal Employee Viewpoint Survey (FEVS), which contains facet-level job satisfaction ratings from 407,789 US federal employees, and which we aggregated to the agency level, and (b) current overall and facet ratings of job satisfaction of the federal agencies contained within FEVS from Glassdoor.com as scraped from the Glassdoor application programming interface (API) within a month of the FEVS survey's administration. Using these data, we examined convergent validity, discriminant validity, and methods effects for the measurement of both overall and facet-level job satisfaction by analyzing a multitrait-multimethod matrix (MTMM). Most centrally, we provide evidence that overall Glassdoor ratings of satisfaction within US federal agencies correlate moderately with aggregated FEVS overall ratings ($r = .516$), supporting the validity of the overall Glassdoor rating as a measure of overall job satisfaction aggregated to the organizational level. In contrast, the validity of facet-level measurement was not well-supported. Overall, given varying strengths and weaknesses with both Glassdoor and survey data, we recommend the combined use of both traditional and crowdsourced data on organizational characteristics for both research and practice.

Crowdsourced information about traditionally private aspects of organizational and employee functioning is becoming an increasingly common source of data for not only academic researchers but also researchers practicing in human resources (Landers, Brusso, Cavanaugh, & Collmus, 2016). Workers often voluntarily and without compensation provide a significant amount of information about their workplace activities to public websites, like Glassdoor, Twitter, LinkedIn, and Indeed. For example, since 2008, Glassdoor has acquired over 45 million employee reviews across 830 thousand employers (Glassdoor, 2019). Through either web scraping or programmatic access of online data portals (i.e., application programming interfaces; APIs), datasets can be curated from such websites providing access to otherwise difficult-to-obtain organizational information from a diverse group of organizational members.

For the assessment of job satisfaction, Glassdoor, whose business model focuses upon maintaining a database of crowdsourced employer reviews, has been an attractive source of data for researchers, practitioners, and job

seekers. In one study, Luo, Zhou, and Shon (2016) found correlations between the content of textual Glassdoor reviews and company performance, observing that across 257,454 reviews of 425 organizations across industries, the presence of certain themes, such as teamwork, innovation, and respect, were associated with performance as assessed via traditional financial metrics. In other studies exploring different criteria, improvements in Glassdoor numeric ratings over time were found to be associated with increases in corporate performance (Green, Huang, Wen, & Zhou, 2018; Melian-Gonzalez, Bulchand-Gidumal, & Lopez-Valcarcel, 2015) and job interview difficulty (Chamberlain & Chen-Zion, 2015). Practitioners writing in

Corresponding author:

Richard N. Landers

Department of Psychology, University of Minnesota,
75 E River Road, Minneapolis, MN 55455

Email: rlanders@umn.edu

Harvard Business Review frequently rely upon Glassdoor data when drawing conclusions about workforce trends in satisfaction and appropriate business strategies in response (e.g., Dattner, 2016), with pointed mentions (e.g., Carucci, 2016) of Glassdoor's published "Best Places to Work" rankings based upon Glassdoor numeric ratings. Researchers at Glassdoor itself have been creating press releases based upon internal research on their database, for example, identifying key factors predicting employee turnover (Smart & Chamberlain, 2017) and driving American workers to relocate for work (Chamberlain, 2018). Within the industrial-organizational psychology literature, a mixture of academic and practicing researchers, Chamorro-Premuzic, Winsborough, Sherman, and Hogan (2016), went so far as to suggest that "organizations can effectively crowdsource their evaluations of leadership" (p. 631) using Glassdoor data, although they neither provided nor referenced empirical work to support this claim. Among job seekers in one industry survey, 48% of respondents (2201 of $N = 4633$) reported having used Glassdoor at some point during their most recent job search (Westfall, n.d.), and Glassdoor itself reports 67 million monthly visitors (Glassdoor, 2019).

Despite this popularity and seemingly high level of trust in Glassdoor data, evidence that these ratings are psychometrically valid measures of global job satisfaction is limited; at the facet level (e.g., satisfaction with compensation, work-life balance, etc.), it is entirely absent. DeKay (2013) conducted a content analysis of Glassdoor reviews, coding for indicators of motivator and hygiene factors, finding a correspondence between these codes and overall numeric Glassdoor ratings. This provides some support for the construct validity of the overall Glassdoor score as theoretical factors consistently defined in the motivation research literature appear to correspond with ratings, to some degree. However, we could identify no published research that spoke to the psychometric properties of Glassdoor ratings in terms of internal structure or correspondence with other more well-known satisfaction measures, which are crucial types of validity evidence (Cronbach & Meehl, 1955).

There are two core challenges faced by researchers wishing to develop this type of validity evidence on websites like Glassdoor that likely challenged past researchers and also inspired the development of the present study. First, a cross-organizational dataset containing valid ratings of job satisfaction from employees within each organization must be curated to compare job satisfaction to Glassdoor scores at the organizational level of analysis. Second, Glassdoor's databases must be referenced for data related to those organizations in a timely fashion; specifically, Glassdoor does not currently allow for accessing historical rating levels, such as at a specific point in the last year.¹ Instead, Glassdoor only allows interested users to access current ratings and content. After addressing these challenges, the purposes of the present study were twofold.

First, we sought to validate overall Glassdoor organization ratings using a publicly available database of job satisfaction ratings collected at roughly the same time as the Glassdoor ratings were collected, and we hypothesized that these ratings would converge. Second, we set out to explore the construct validity of the facet ratings acquired in the same fashion. To do this, we collected publicly available job satisfaction data from the results of the United States (US) Office of Personnel Management's (OPM) Federal Employee Viewpoint Survey (USOPM, 2017), a survey measure administered annually to employees within the US federal government. FEVS contains a facet-level job satisfaction measure commonly used in archival research on US federal employees (Fernandez, Resh, Moldogaziev, & Oberfield, 2015). Second, we collected ratings from Glassdoor on each of the federal agencies represented in FEVS. Third, we recoded the FEVS items into the job satisfaction facets assessed by Glassdoor. Last, we applied a multitrait-multimethod matrix (MTMM) analytic framework to assess the construct validity of the Glassdoor ratings against the FEVS ratings of the same organizations. Specifically, MTMM enabled us to examine three aspects of construct validity: convergent validity, discriminant validity, and method effects attributable to rating source. We can formally state our investigation as such:

Hypothesis: Overall job satisfaction ratings, aggregated to the organizational level, will correlate with Glassdoor.com overall ratings.

Research Question: What is the construct validity of Glassdoor.com ratings as measures of job satisfaction when contrasted with traditional job satisfaction ratings, at both the global and facet level?

METHOD

Sampling and Data Collection

We selected US federal agencies as our sampling frame in this study for two reasons. First, facet-level job satisfaction data are made publicly available in FEVS for all federal agencies, split by year, and these data are released after vetting by OPM. This permits the assessment of facet-level responses with a high level of trustworthiness; construct validity can be assessed in this dataset using commonly accepted psychometric approaches. Second, because Glassdoor data exist at the organization level of analysis (i.e., there is no way to link spe-

¹ Because Glassdoor.com is a commercial website hosted by a private organization, access restrictions in relation to its databases may change at any time and have indeed changed several times since this study was conducted. Researchers seeking to download Glassdoor data at this time should consult contemporary online discussions regarding these issues or contact Glassdoor directly, instead of relying upon the technical approach described herein.

cific people across datasets, and individual ratings are only even available when attached to a narrative review, which is historically a subset of all such ratings), Glassdoor ratings need to be validated at the organizational level (Klein, Dansereau, & Hall, 1994). Given that, we did not identify any other way to collect a large cross-organizational dataset of job satisfaction data.

US federal job satisfaction data. FEVS data contains satisfaction ratings from all US federal agencies. In the FEVS 2017 data, which assessed 407,789 federal employees in late 2016, 40 agencies were individually identified in addition to an “other” category. Below the agency level, 219 distinct codes were used to distinguish between agency units. For example, the Department of Agriculture is an agency and was specified as containing eight units, including Farm and Foreign Agriculture Services, Food Safety, and Rural Development. In our initial exploration of the corresponding Glassdoor data, we discovered that units were not typically distinctly rated. For example, although the Department of Agriculture appears in Glassdoor, Farm and Foreign Agriculture Services does not. Additionally, OPM collapses unit data into agency-level data or otherwise anonymizes it when unit-level data sample sizes are less than 10. Thus, we decided to focus our analysis at the agency level.

Glassdoor ratings. Glassdoor is somewhat unusual in the universe of public-facing crowdsourced websites in that it implements a “give-to-get” policy requiring users to submit a review of an employer after viewing three reviews left by others. This is intended to combat polarizing ratings biases associated with self-selection that can lead to bimodal ratings distributions (Hu, Pavlou, & Zhang, 2017; Li & Hitt, 2008). Incentivized reviewing has been found to reduce bias in ratings on Glassdoor, although this research was conducted by a team that included a Glassdoor researcher (Marinescu, Klein, Chamberlain, & Smart, 2018). Thus, Glassdoor may have some design characteristics that could improve its rating quality at the cost of slightly less prototypicality related to crowdsourced worker data in general. Given interest by the assessment community in the Glassdoor platform specifically, we decided this was a worthwhile tradeoff.

Collecting Glassdoor data was a multistep process. To match up the FEVS data collection effort as closely as possible to the timeliness of data collected from Glassdoor, we curated our ratings database in September 2016 by scraping Glassdoor’s public API (Landers et al., 2016). To do this, undergraduate research assistants first hand-coded agencies to Glassdoor identifiers by searching on the Glassdoor webpage; they were able to locate 37 of the 40 agencies in the FEVS data. Once a list of identifiers was developed, the Glassdoor API was queried for each identifier, from which current Glassdoor ratings were downloaded. This included overall ratings as well as ratings of Culture and Values,

Senior Leadership, Compensation and Benefits, Career Opportunities, and Work-life Balance, all to a precision level of one decimal place. Because reported Glassdoor ratings include all individual ratings in the database, some degree of temporal error is likely represented in Glassdoor ratings relative to the FEVS data; whereas FEVS ratings reflect job satisfaction at a narrowly defined point in time, Glassdoor ratings represent job satisfaction aggregated across time, throughout the history of the organization’s existence on Glassdoor up until the date the score is accessed. Although this may be limiting from a true-score validity estimation point of view, it increases generalizability to realistic use cases of Glassdoor: people consulting Glassdoor today to estimate current job satisfaction within the company.

Scale Development and Dataset Validation

Next, the individual-level item data in the FEVS dataset needed to be converted into a dataset containing scale scores at the organizational level to enable comparisons with the Glassdoor data. To do this, we followed four general steps modifying the FEVS dataset: initial item coding, evaluating and addressing missingness, scale validation and revision, and aggregation.

FEVS content coding. In the first step, we used a content coding approach to determine which items within the FEVS dataset assessed the same constructs as those in the Glassdoor dataset. To do this, one of the present authors of this paper first hand coded the 71 items in the FEVS to the six Glassdoor categories, discarding any items that did not cleanly map onto a Glassdoor category. After this initial hand coding, consensus judgments were reached with each item with another author, which eliminated several items. The resulting list of Glassdoor-relevant FEVS items after content coding appears in Table 1.

FEVS missingness. In the second step, we quantified and explored missingness (Newman, 2009). First, to account for missingness not at random, we eliminated any cases with greater than 75% missingness across all variables, which decreased the working sample size from 407,789 to 401,846. Within the remaining cases, we found that the variables identified in the first step contained between 0.19% and 9.98% missing data ($M = 2.35\%$; median = 3.23%). Although this is not substantial missingness on a per-item basis, case-wise missingness was 26.18%, which would cause a listwise deletion strategy to eliminate a large proportion of the dataset. Thus, we imputed missing values via expectation-maximization using Amelia II (Honaker, King, & Blackwell, 2011; Horton & Kleinman, 2012).

FEVS scale validation and revision. Third, we evaluated the validity of our constructed scales with a two-level confirmatory factor analysis, loading each of the five facet constructs onto the overall satisfaction construct, evaluating the χ^2 test and CFI, RMSEA, and SRMR fit indices associated with each model against generally accepted standards

TABLE 1.
FEVS Items and Glassdoor Categorizations

Glassdoor category	FEVS item number and text
Overall rating	39. My agency is successful at accomplishing its mission. 69. Considering everything, how satisfied are you with your job? 71. Considering everything, how satisfied are you with your organization?
Career Opportunities	1. I am given a real opportunity to improve my skills in my organization. 47. Supervisors in my work unit support employee development. 67. How satisfied are you with your opportunity to get a better job in your organization?
Compensation and Benefits	70. Considering everything, how satisfied are you with your pay?
Culture and Values	30. Employees have a feeling of personal empowerment with respect to work processes. 32. Creativity and innovation are rewarded. <i>37. Arbitrary action, personal favoritism and coercion for partisan political purposes are not tolerated.</i> 38. Prohibited personnel practices (for example, illegally discriminating for or against any employee/applicant, obstructing a person's right to compete for employment, knowingly violating veterans' preference requirements) are not tolerated.
Senior Leadership	53. In my organization, senior leaders generate high levels of motivation and commitment in the workforce. 60. Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor? 61. I have a high level of respect for my organization's senior leaders. 66. How satisfied are you with the policies and practices of your senior leaders?
Work–Life Balance	10. My workload is reasonable. <i>42. My supervisor supports my need to balance work and other life issues.</i>

Note. Items in list represent initial content coding. Italicized items were dropped after examining results of confirmatory factor analysis.

(Bagozzi & Yi, 1988; Hu & Bentler, 1999) using lavaan (Rosseel, 2012). Because Compensation and Benefits had only one item, it was modelled as perfectly reliable. This model fit somewhat well ($\chi^2[115] = 415666.17$, $p < .001$, CFI = .917, RMSEA = .095, SRMR = .042). Upon examination of modification indices, it was discovered that items 37 and 38 within the Culture and Values scale were more highly correlated with each other than other culture items tended to be, although item 38 had a greater loading on the Culture and Values construct than did item 37. Additionally, item 42 correlated more strongly with Career Opportunities (both at the construct and item level) than with either the other Work–Life Balance item or the Work–Life Balance latent construct. Thus, items 37 and 42 were eliminated from further analyses. When the confirmatory factor analysis was conducted without these items, all relative fit indices were within generally accepted standards ($\chi^2[87] = 206557.85$, $p < .001$, CFI = .952, RMSEA = .078, SRMR

= .033). Although fit likely could have been improved by dropping additional items, we decided that this point was a reasonable balance in modeling decisions combining our a priori content coding with data-driven decision making. Using this set of scales and indicators, the coefficient alpha reliabilities of the mean composite scores were also assessed, which were all above generally accepted thresholds ($\alpha_{\text{overall}} = .86$; $\alpha_{\text{career}} = .81$; $\alpha_{\text{culture}} = .82$; $\alpha_{\text{leaders}} = .92$). Thus, after this step, each agency was represented by only six construct score estimates.

FEVS aggregation. Fourth, we aggregated to the agency level by calculating mean scores within agencies, resulting in a final aggregated, imputed dataset of $N_{\text{agencies}} = 40$ representing between 320 and 46,991 responses per agency ($N_{\text{mean}} = 10059.50$; $N_{\text{median}} = 5360$). Simultaneously, we assessed three aggregation metrics, ICC(1), ICC(2), and r_{wg} (Bliese, 2000; James, Demaree, & Wolf, 1984). Although these statistics are often used to determine if aggregation is

justifiable versus explicit multilevel modeling, in the present study, we had no option except aggregation given the goal of examining construct validity at the agency level for our later MTMM validation. Thus, this analysis was intended as an assessment of reliability and agreement rather than a justification for aggregation. Each of these metrics provides a different type of information given slightly different assumptions. As shown in Table 2, ICC(1) results were universally low, ranging from .014 to .034, suggesting that the group mean of any dimension of satisfaction does not reflect individual job satisfaction very well (i.e., individual raters are not reliable estimates of the agency mean). This was expected; across all positions in an entire agency, there is likely to be substantial variance in job satisfaction. In contrast, ICC(2) results were universally high, suggesting that the overall sample size was sufficient to get a reliable estimate of group satisfaction means. In contrast to ICC, r_{wg} is an estimate of within-group agreement and in the present context, assesses the degree to which raters agree with each other within agencies; thus, each agency has its own r_{wg} . Minimum, means, and maximums are shown in Table 2. In general, agreement was moderate, with means ranging from .311 to .591.

Glassdoor reliability and validity. Turning next to the Glassdoor dataset, there was relatively little information available in the API output to evaluate ratings quality. The sole exception was sample size; Glassdoor reports the number of ratings represented by each agency's overall rating score, although sample sizes for facet scales may be smaller. In these data, we noted significant range in the quantity of information available; across the 37 agencies located, rating counts varied from 0 to 11852 ($k_{Mean} = 880$, $k_{Median} = 105$). Because extremely low rating counts are likely to negatively affect the validity of mean ratings, we eliminat-

ed 10 additional agencies from the dataset with outlying low rating counts ($k < 15$). Thus, the final Glassdoor dataset contained 27 agencies, reflecting a sacrifice of sample size in exchange for increased reliability. This dataset was merged with the aggregated, imputed FEVS dataset to create the final focal dataset for this study.

RESULTS

Descriptive statistics comparing sources by agency appear in Table 3. Because Glassdoor ratings are only shared to a single decimal place both on the Glassdoor website and through its API, FEVS means have been displayed at the same level of precision. In general, means in both datasets fell in the same general range across sources within their respective five-point scales.

The MTMM used to assess the validity of Glassdoor ratings appears in Table 4, which summarizes correlations across the six traits and two methods targeted by this study. MTMM analysis is a classic technique for assessing convergent and discriminant validity when multiple sources assess the same construct (Campbell & Fiske, 1959). Although confirmatory factor analysis is now generally used to conduct MTMM analyses in the modern research literature (Koch, Schultze, Burrus, Roberts, & Eid, 2015), this approach is more prone to error when sample sizes are small (operationally defined as $N < 125$ by Marsh & Bailey, 1991). Given the small sample size in the final agency-level dataset, we thus chose to interpret the MTMM in the classical fashion by examining patterns of correlations within the MTMM correlation matrix. For interpretation, we relied upon Schmitt, Coyle, and Saari's (1977) outline of the Campbell-Fiske criteria, which describes specific patterns of expected relationships.

TABLE 2.
Reliability Estimates for Aggregation

Construct	ICC		r_{wg}		
	(1)	(2)	Min	Mean	Max
Overall rating	.026	.996	.420	.591	.746
Career Opportunities	.014	.993	.354	.512	.670
Compensation and Benefits	.037	.997	.144	.332	.453
Culture and Values	.034	.997	.389	.512	.658
Senior Leadership	.032	.997	.290	.437	.634
Work-Life Balance	.024	.996	.162	.311	.449

TABLE 3. Mean Scores and Sample Sizes for FEVS and Glassdoor (GD) Samples on Study Constructs Across FEVS Agency Codes

Agency	Rating counts		Overall rating		Career		Comp/Benefits		Culture/Values		Senior leaders		Work-Life	
	FEVS	GD	FEVS	GD	FEVS	GD	FEVS	GD	FEVS	GD	FEVS	GD	FEVS	GD
AF	15165	7619	3.8	4.1	3.4	4.1	3.5	4.3	3.5	4.1	3.5	3.2	3.4	3.2
AG	22682	105	3.7	3.8	3.5	3.2	3.5	3.8	3.4	3.5	3.3	3.1	3.2	4.1
AM	2254	203	3.7	3.7	3.6	3.5	3.5	3.5	3.4	3.9	3.4	3.1	3.0	3.5
AR	16799	11852	3.7	4.0	3.4	4.1	3.5	4.2	3.3	4.0	3.4	3.1	3.3	2.8
CM	9677	145	3.8	3.9	3.6	3.2	3.6	2.8	3.5	3.7	3.4	3.4	3.3	4.1
DJ	16056	256	3.9	3.8	3.5	3.5	3.6	3.4	3.4	3.6	3.5	3.3	3.5	3.9
DL	11194	173	3.8	2.9	3.5	2.5	3.6	3.4	3.4	2.9	3.5	2.3	3.4	4.1
DN	7998	153	3.7	3.5	3.6	3.0	3.6	3.3	3.4	3.0	3.3	2.9	3.4	3.9
ED	2844	65	3.7	3.8	3.5	3.2	3.6	3.6	3.3	3.4	3.3	3.1	3.2	3.8
EP	10081	270	3.7	3.8	3.5	3.2	3.6	3.8	3.3	3.6	3.3	2.8	3.2	4.3
FC	638	35	3.7	3.6	3.5	3.0	3.4	3.0	3.3	3.5	3.4	3.3	3.5	3.9
GS	6991	173	3.8	3.4	3.7	2.9	3.7	3.6	3.5	3.0	3.5	2.5	3.5	4.2
HE	39844	152	3.8	3.6	3.6	3.2	3.6	3.6	3.5	3.2	3.5	2.9	3.4	3.6
HS	46296	374	3.4	3.1	3.2	3.0	3.4	3.6	3.1	2.6	3.0	2.1	3.2	3.0
HU	5437	76	3.6	3.0	3.5	2.8	3.6	3.8	3.3	2.8	3.3	2.7	3.2	4.2
IN	22764	70	3.6	3.6	3.5	3.0	3.5	3.4	3.3	3.6	3.2	3.1	3.1	4.0
NN	11119	279	4.1	4.3	4.0	4.0	3.8	4.0	3.9	4.2	3.8	3.7	3.6	4.3
NU	2135	60	3.9	3.9	3.6	3.6	3.7	4.1	3.5	4.1	3.6	3.6	3.7	4.6
NV	12121	7020	3.7	3.9	3.4	4.0	3.5	4.2	3.4	3.8	3.4	3.0	3.4	2.6
OM	3169	55	3.7	3.4	3.6	3.2	3.7	4.2	3.4	3.4	3.4	3.2	3.4	4.1
SE	3185	139	4.0	3.9	3.6	3.5	3.9	4.0	3.5	3.7	3.6	3.4	3.7	4.5
ST	5119	658	3.8	3.7	3.6	3.6	3.7	3.6	3.5	3.5	3.5	3.1	3.2	3.5
SZ	8699	457	3.7	3.3	3.5	2.8	3.6	3.8	3.4	3.0	3.4	2.8	3.2	3.9
TD	14712	77	3.8	3.6	3.5	3.1	3.5	3.7	3.4	3.6	3.4	2.9	3.4	3.8
TR	44885	168	3.6	3.8	3.4	3.3	3.3	3.5	3.3	3.5	3.3	3.3	3.3	4.0
VA	29644	1841	3.6	3.3	3.3	3.2	3.3	3.8	3.2	3.1	3.2	2.6	3.3	3.5

TABLE 4.

Multitrait-Multimethod Matrix Crossing Rating Source and Construct and Correlation Matrix Including Overall Scores

	FEVS						Glassdoor					
	Overall	Career	Comp/B	Cul/Val	Leaders	WL	Overall	Career	Comp/B	Cul/Val	Leaders	WL
FEVS												
Overall	(.86)											
Career	.848	(.81)										
Comp/B	.711	.739	(-)									
Cul/Val	.923	.916	.674	(.82)								
Leaders	.925	.836	.648	.897	(.92)							
Work–Life	.696	.446	.392	.524	.639	(-)						
Glassdoor												
Overall	.516	.425	.168	.569	.451	.317	(-)					
Career	.353	.191	.063	.435	.380	.254	.831	(-)				
Comp/B	.132	-.057	.208	.176	.187	.177	.260	.564	(-)			
Cul/Val	.521	.385	.135	.576	.497	.298	.902	.818	.336	(-)		
Leaders	.592	.542	.288	.577	.578	.423	.813	.593	.103	.820	(-)	
Work–Life	.437	.519	.565	.344	.344	.350	-.042	-.414	-.214	-.020	.295	

Note. Correlations are statistically significant at $\alpha = .05$ where $r \geq |.39|$. Correlations on the overall diagonal are coefficient alpha reliabilities before aggregation. **Bolded** correlations are MTMM monotrait-heteromethod estimates (i.e., validity diagonals). Light grey cells are MTMM heterotrait-monomethod estimates. Dark grey cells are MTMM heterotrait-heteromethod estimates. $N = 27$ federal agencies.

In brief, MTMM can be conceptualized as combining three types of information: convergent validity, method variance, and discriminant validity. First, convergent validity is estimated directly in the MTMM and is represented as monotrait-heteromethod correlations, so named because they assess the degree to which scores from the two methods, when intended to assess the same construct, in fact do so. In the present study, we would expect these estimates to be high if FEVS aggregated agency scores are accurately predicted by Glassdoor rating means. Second, method effects are reflected in the degree to which scores of different constructs assessed by the same method nevertheless correlate with each other; in the language of MTMM, the absence of method effects should be reflected in values within the monotrait-monomethod triangles approximately equal to corresponding values in the heterotrait-heteromethod triangles. In the present study, we would expect monotrait-monomethod correlations to be higher in the presence of common method variance (Spector, 2006) or other halo effects within either the FEVS or Glassdoor data; for example, if employees were using Glassdoor to vent their frustration with their employers rather than to make honest ratings, we would expect large method effects. Third, dis-

criminant validity is reflected by the degree to which different methods assessing different constructs in fact do not correlate. In MTMM, this type of validity is demonstrated when each monotrait-heteromethod estimate is greater than values in its respective rows and columns. Importantly, the presence of method variance may suppress this pattern.

To parse the results of this analysis, we will discuss each of these types of estimates displayed in the MTMM. Table 4 contains the facet-level MTMM but also intercorrelations between facets and Overall Job Satisfaction. Because nomologically speaking, Overall Job Satisfaction is a latent factor composed of facet scores, we would expect it a priori to be correlated with all other scores in the matrix. Thus, we did not include the overall scores in our MTMM analysis.

Evidence regarding convergent validity. Our hypothesis in this study was regarding convergent validity of the overall rating and stated that Glassdoor overall satisfaction ratings would correlate with FEVS overall satisfaction ratings. FEVS and Glassdoor overall satisfaction ratings did converge ($r = .516$, $p = .007$), with 26.7% of variance in FEVS mean scores explained by Glassdoor ratings, supporting our hypothesis. Although this is likely not sufficiently

high to merit replacement of traditional employee surveys with Glassdoor, it does suggest that overall Glassdoor ratings are reasonable but imperfect proxies for job satisfaction when globally measured using surveys.

Facet convergent validities varied widely. Career Opportunities, Compensation and Benefits, and Work–Life Balance were weakest ($r = .191, .208, .350$, respectively), whereas Culture and Values and Senior Leadership were larger and statistically significant ($r = .576, .578$, respectively). This fails to consistently achieve our decision criterion for evaluation via MTMM, which is that these validities would all be large and statistically significant. Thus, support for facet-level convergent validity was mixed but generally negative.

Evidence regarding method effects. To examine method effects, correlations within each heterotrait-monomethod triangle were compared with corresponding effects in the hetero-trait-monomethod triangles. This revealed an unexpected pattern; although method effects were present for both FEVS and Glassdoor data, the FEVS data appeared to have stronger common method variance. Correlations between facet scores within the FEVS data (min = .392, max = .916) were universally greater than their mirror correlations within the Glassdoor data (min = -.020, max = .820). Method effects were also more consistent in the FEVS data; whereas all correlations here were moderate to strong and positive, Glassdoor method effects were sometimes near zero. To draw conclusions regarding overall method effects in the facet measures consistently with the Campbell-Fiske approach, we counted these comparisons, finding evidence of method variance in both cases, although more so for FEVS (56.7% greater) than for Glassdoor (36.7% greater). Thus, we concluded method variance was present in both sources. However, we were unable to determine the degree to which this variance reflected authentic method variance attributable to rating source versus the high positive manifold of job satisfaction as a construct.

Evidence regarding discriminant validity. Because we found evidence of either method variance or problems with the positive manifold of job satisfaction, evidence regarding discriminant validities became more difficult to interpret. Specifically, the median heterotrait-monomethod correlation within FEVS was .661, which was greater than all observed convergent validities, and the median within Glassdoor was .315, which was greater than two observed convergent validities. Thus, before conducting this analysis, we knew that they would fail by the Campbell-Fiske criteria due to violations of underlying assumptions of MTMMs. In response, we altered our approach slightly to focus upon comparisons within the heterotrait-heteromethod triangles. Although this is a weaker test of discriminant validity overall, it still enabled us to determine if monotrait-heteromethod estimates were greater than the most difficult-to-interpret correla-

tions in the matrix. Within this more exploratory analytic framework, we found evidence supporting discriminant validity most positively for Senior Leadership (100%), less positively for Culture and Values (88%), Compensation and Benefits (75%), and Work–Life Balance (63%), and most poorly for Career Opportunities (25%). Thus, evidence for discriminant validity was mixed; our original confirmatory tests failed, and exploratory analyses revealed some patterns of interest for some constructs.

From these results, combined with those from the tests of convergent validity, we concluded that there is currently insufficient evidence to claim that high-quality facet measurement is generally available using Glassdoor data. However, we also concluded that high quality facet measurement was not available in the FEVS survey data. Thus, we were unable to demonstrate high-quality facet measurement in either dataset, which could reflect a more general problem with the measurement of facet-level job satisfaction in this study.

DISCUSSION

Overall, this study demonstrated that overall Glassdoor ratings can be used as an imperfect proxy for survey-based global job satisfaction. In support of our hypothesis, overall Glassdoor ratings did moderately correlate with traditional job satisfaction measures aggregated to the organizational level. However, we do not recommend simply replacing internal global job satisfaction surveys with the interpretation of Glassdoor data; the convergent validity coefficient was too low to support this as a general strategy, so whether this coefficient is of sufficient magnitude to justify decision making is revealed as context dependent. For practical purposes, for some organizations, an overall convergence of $r = .516$ may be “good enough,” given the resource expenditure necessary to deploy internal job satisfaction surveys. Substantial shifts in Glassdoor numbers, such as a 2 point drop over 3 months, may be sufficient for some types of decision making. In research contexts, however, this convergence is insufficiently strong to claim that the constructs assessed by Glassdoor are identical to those assessed by traditional satisfaction measures. Thus, for organizational decision making when small changes are substantively important or for any research purposes, Glassdoor data alone should not replace the use of survey-based global job satisfaction measures. Instead, Glassdoor itself should be considered a distinct source of information about job satisfaction, related to but distinct from traditional survey-based research, and an important research topic unto itself.

In exploration of our research question regarding facets, we concluded that evidence supporting the validity of facet-level measurement from crowdsourced job satisfaction data is currently limited. Although we found weak pos-

itive evidence of convergent validity for Culture and Values and Senior Leadership in the Glassdoor data, we also found evidence of method effects in both the FEVS and Glassdoor data, which made interpretation of discriminant validities difficult. It is unclear to what degree this occurred due to legitimate source effects or due to known high intercorrelations between facets of job satisfaction (Wanous & Lawler, 1972). When that common method variance was optimistically ignored, Culture and Values and Senior Leadership emerged as the highest quality facets of Glassdoor ratings. However, the use of these facets for decision making rests on several risky assumptions, so we do not recommend it at this time. Further research exploring facets with other datasets containing both crowdsourced and traditionally measured facet job satisfaction is needed.

Limitations

We identified three primary limitations to this study. First, unreliability may have attenuated observed validities. Within the FEVS heterotrait-monomethod triangles, the smallest correlations tend to be those associated with single-item measures (Compensation and Benefits, and Work–Life Balance) and the largest with multi-item measures with high reliability (all $\alpha > .80$ as demonstrated previously). All Glassdoor measures are essentially single-item scales, and single-item scales are generally associated with low reliabilities. Thus, all three types of estimates may have been attenuated when Glassdoor ratings were being assessed and to a greater degree for Glassdoor method effects estimates due to the multiplicative effects of unreliability on validity estimates. Given our research questions, this is analogous to a meta-analytic “operational validity,” in that unreliability in Glassdoor ratings is inherent to Glassdoor ratings; validities attenuated in this way accurately reflect the attenuation effects experienced by people drawing conclusions from the Glassdoor website. Nevertheless, it may also prevent us from estimating accurate true score estimates of validity.

Second, true score convergent validities between developed FEVS scales and Glassdoor category ratings were likely less than one due to differences in construct specification. Examining content validity in Table 1 demonstrates that our development process for FEVS scales still did not measure some Glassdoor facets very cleanly. Work–Life Balance was particularly problematic; there were no items that we thought excellently mapped onto Work–life Balance; Item 10 spoke to work–life balance only indirectly whereas Item 42 included a supervisory job performance dimension. Thus, our development process ultimately identified and included only one item that potentially has systematic error regarding measurement of Work–Life Balance. The relatively low convergent validity for that scale ($r = .350$), and for the other two lower validity facet measures,

may be attributable to limitations of the FEVS data in assessing the Glassdoor constructs rather than a low true score validity estimating Work–Life Balance. As much as we struggled mapping items, an even more fundamental problem is that the operational definitions of people making ratings on the Glassdoor website may be systematically different from ours. For example, although we as industrial-organizational psychologists have a research-informed definition of Work–Life Balance, it is highly doubtful that this is the definition that was used by all website visitors. Additionally, recall the temporal variance issue discussed earlier; although FEVS data are cross-sectional, Glassdoor data are aggregated and longitudinal. Although Glassdoor claims that its ratings include “an emphasis on recent reviews,” the precise weighting of timeliness is not publicly available. Thus, Glassdoor ratings may combine long-term swings in construct standing as well as short-term spikes; for example, the complete US federal government shutdown of 2013 and any lingering effects on job satisfaction may be represented in ratings. All these issues together speak to potential systematic differences in constructs and scope between the FEVS and Glassdoor data that likely decreased validity true scores a priori, reflecting a general limitation of this approach. If a true cross-organizational dataset could be developed with explicit parallels to Glassdoor construct labels, such as by asking employees to simply rate “Satisfaction with Work–Life Balance” with no further items, many of these issues could be addressed methodologically.

Third, although we had hoped that the use of FEVS data would provide us with high-quality psychometric measures of job satisfaction, it appeared that the FEVS data may suffer from common method variance making construct measurement difficult. As noted earlier, FEVS was not developed to assess the Glassdoor constructs, so some of the apparent method variance may be a side effect of our development process. Alternatively, because FEVS data are collected in a formal work environment, some employees may have believed that they could suffer unwanted consequences if responding negatively to this organization-sponsored survey (cf. Giacalone, Knouse, & Montagliana, 1997). When rating a company on Glassdoor, such suspicions are highly unlikely, because the provision of ratings on Glassdoor is initiated by the rater not requested by an entity affiliated with one’s employer. Thus, there is a possibility that Glassdoor true scores more accurately represent construct standing than do FEVS scores. With the current dataset, it is unclear which potential source of error is more problematic; a larger cross-organizational dataset with greater control of survey content, and preferably with criteria relevant to job satisfaction collected from a different source, such as supervisors, would be necessary to address this concern directly.

Practical Implications and Conclusion

We conclude with a direct recommendation for practice. Specifically, we recommend that overall Glassdoor ratings be taken seriously as informative regarding global job satisfaction. Employees do appear to add satisfaction-loaded ratings to Glassdoor, they do not merely use the website to vent their frustrations (which would have created a more skewed or even bimodal distribution), and they may even feel more unrestricted on Glassdoor to rate and comment honestly. That said, because the research literature regarding survey measures is much more established, and because the overall convergent validity was not high by a measurement standard ($r = .516$), we do not recommend using Glassdoor ratings to compare organizations directly on job satisfaction, especially when scores are similar. In the present study, the final rank orderings of organizations were compared, and they changed moderately between assessment methods. Although the top organization was the same across assessment methods, the correlation between rank orderings between methods was only 0.453. Thus, comparisons of dramatically higher or lower Glassdoor ratings may be useful, but differences of smaller magnitudes are likely uninterpretable. Furthermore, it is unclear which data source was the more accurate reflection of true score job satisfaction, so we recommend considering both as valid sources of information about job satisfaction but with different strengths and limitations.

Our recommendations for the use of facet scores are necessarily more nuanced. Although we found weak evidence in partial support of the construct validity of two facets, Senior Leadership as well as Culture and Values, we found essentially no evidence in support of the construct validity of the other three. This should not be interpreted to mean that Glassdoor facet scores are necessarily inferior or uninformative; instead, the current results are merely inconclusive, and there remain theoretical reasons to suspect Glassdoor facet ratings could be useful. Specifically, the greater positive manifold between facet measures of satisfaction from FEVS – in several cases, approaching 1.0 after correcting for attenuation due to unreliability – suggests that Glassdoor may provide more nuanced and less common method variance prone ratings. Given the much larger and more established research literature surrounding survey-based measures of facet job satisfaction, we currently recommend sticking to traditional methods for mission-critical facet measurement, especially when comparing results to past data collection efforts, but we also suggest careful consideration of the context of data collection. For an organization with employees worried that any negative satisfaction ratings could be used punitively, Glassdoor or other crowdsourced job satisfaction data might provide useful information diagnostic of specific satisfaction challenges within that organization, given reduced method variance

associated with that source, that would not be achievable with surveys. Additional research is needed to explore this further.

As a final note, we strongly recommend further research into ratings like these and other crowdsourced organizational data broadly. We have demonstrated that aggregated publicly available ratings can reflect organizational standing on constructs, at least under certain circumstances, on certain websites. Yet numeric ratings of job satisfaction are just the first breaking waves of a new era of publicly available, crowdsourced organizational data. Glassdoor also collects data on organizations' job availability and descriptions, interview processes and questions, work environment, and job-specific salaries in the form of employee numeric ratings, text responses, and even images. As trace data collection and analysis becomes even more commonplace and as Internet-enabled devices outside of organizational control increasingly enter the workplace, even finer organizational details will leak into public view, some accurate and some not. The ethics of this situation are irrelevant; it is inevitable given the course of modern technology. For example, one can only imagine the court of public opinion judging whether an organization's crowdsourced job performance and compensation data are appropriately correlated across race, sex, gender, religion, and other class memberships. Only through research-practitioner partnerships exploring the validity of such information can we contextualize such data for its public consumption. Without such research, we are no longer even part of the conversation.

REFERENCES

- Bagozzi, R. P. & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 16, 74-94.
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349-381). San Francisco, CA: Jossey-Bass, Inc.
- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Carucci, R. (2016). Big companies don't have to be soulless places to work. *Harvard Business Review*. Retrieved from <https://hbr.org/2016/03/big-companies-dont-have-to-be-soulless-places-to-work>
- Chamberlain, A. (2018). Metro movers: Where are Americans moving for jobs, and is it worth it? Retrieved from https://www.glassdoor.com/research/app/uploads/sites/2/2018/05/GD_ResearchReport_MetroMovers_Draft4.pdf

- Chamberlain, A. & Chen-Zion, A. (2015). Do difficult job interviews lead to more satisfied workers? Evidence from Glassdoor reviews. Retrieved from <https://www.glassdoor.com/research/studies/interview-difficulty/>
- Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 9(4), 621-640.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Dattner, B. (2016). Why you should interview people who turn down a job with your company. *Harvard Business Review*. Retrieved from <https://hbr.org/2016/08/why-you-should-interview-people-who-turn-down-a-job-with-your-company>
- DeKay, S. H. (2013). Peering through Glassdoor.com: What social media can tell us about employee satisfaction and engagement. In C. M Genest (Ed.), *Proceedings: CCI Conference on Corporate Communication 2013* (pp. 369-382). New York, NY: Corporate Communication International.
- Fernandez, S., Resh, W. G., Moldogaziev, T., & Oberfield, Z. W. (2015). Assessing the past and promise of the Federal Employee Viewpoint Survey for public management research: A research synthesis. *Public Administration Review*, 75, 382-394.
- Giacalone, R. A., Knouse, S. B., & Montagliani, A. (1997). Motivation for and prevention of honest responding in exit interviews and surveys. *Journal of Psychology*, 131, 438-448.
- Glassdoor. (2019). About us. Retrieved from <https://www.glassdoor.com/about-us/>
- Green, T. C., Huang, R., Wen, Q., & Zhou, D. (2018). Crowdsourced employer reviews and stock returns. Retrieved from http://faculty.georgetown.edu/qw50/Green,Huang,Wen,Zhou_EmpRatings.pdf
- Honaker, J, King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7). Retrieved from <https://www.jstatsoft.org/index.php/jss/article/view/v045i07/v45i07.pdf>
- Horton, N. J. & Kleinman, K. P. (2012). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, 61, 79-90.
- Hu, N., Pavlou, P. A., & Zhang, J. J. (2017). On Self-Selection Biases in Online Product Reviews. *MIS Quarterly*, 41(2), 449-471.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55.
- James, L.R., Demaree, R.G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85-98.
- Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review*, 19, 195-229.
- Koch, T., Schultze, M., Burrus, J., Roberts, R.D., & Eid, M. (2015). A multilevel CFA-MTMM model for nested structurally different methods. *Journal of Educational and Behavioral Statistics*, 40, 477-510.
- Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, 21, 475-492.
- Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4), 456-474.
- Luo, N., Zhou, Y., & Shon, J. J. (2016). Employee satisfaction and corporate performance: Mining employee reviews on Glassdoor.com. *Proceedings of the 37th International Conference on Information Systems*. Piscataway, NJ: IEEE.
- Marinescu, I., Klein, N., Chamberlain, A., & Smart, M. (2018). Incentives can reduce bias in online employer reviews. *Academy of Management Proceedings*, 2018,
- Marsh, H. W. & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, 15, 47-70.
- Melián-González, S., Bulchand-Gidumal, J., & González López-Valcárcel, B. (2015). New evidence of the relationship between employee satisfaction and firm economic performance. *Personnel Review*, 44, 906-929.
- Newman, D. A. (2009). Missing data techniques and low response rates: The role of systematic nonresponse parameters. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social science* (pp. 7-36). New York, NY: Routledge.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2). Retrieved from <https://www.jstatsoft.org/index.php/jss/article/view/v048i02>
- Schmitt, N., Coyle, B. W., & Saari, B. (1977). A review and critique of analyses of multitrait-multimethod matrices. *Multivariate Behavioral Research*, 12, 447-478.
- Smart, M. & Chamberlain, A. (2017). Why do workers quit? The factors that predict employee turnover. Retrieved from https://www.glassdoor.com/research/app/uploads/sites/2/2018/05/GD_ResearchReport_WhyWorkersQuit_Rebrand_Draft3.pdf
- Spector, P. E. (2006). Method variance in organizational research: Truth or urban legend? *Organizational Research Methods*, 9, 221-232.
- U.S. Office of Personnel Management. (2017). Federal Employee Viewpoint Survey results. In Washington, DC: Author. Retrieved from <https://www.opm.gov/fevs/public-data-file/>
- Wanous, J. P. & Lawler, E. E. (1972). Measurement and meaning of job satisfaction. *Journal of Applied Psychology*, 56, 95-105.
- Westfall, B. (n.d.). How job seekers use Glassdoor reviews. Retrieved from <https://www.softwareadvice.com/resources/job-seekers-use-glassdoor-reviews/>

RECEIVED 02/11/19 ACCEPTED 08/15/19