Communicating Validity Information to Differentially Experienced Audiences: The Effects of Numeracy and Nontraditional Metrics

Nathaniel M. Voss
Kansas State University, nmvoss@ksu.edu

Christopher J. Lake
Kansas State University, lakec@ksu.edu

COMMUNICATING VALIDITY INFORMATION TO DIFFERENTIALLY EXPERIENCED AUDIENCES: THE EFFECTS OF NUMERACY AND NONTRADITIONAL METRICS

Nathaniel M. Voss[1] and Christopher J. Lake[1]

1. Kansas State University

ABSTRACT

One of the biggest challenges facing organizational researchers is convincing practitioners to adopt evidence-based personnel selection practices such as the structured interview. In this study, we examined the effects of nontraditional validity metrics and numeracy by presenting validity information about the structured interview to audiences with differing amounts of interview experience (students, working adults, and hiring managers). The results indicated that nontraditional metrics were associated with higher understanding, more positive attitudes, and greater perceptions of the usefulness of the structured interview. These effects were constant across differing levels of numeracy. Additionally, the results revealed that nontraditional metrics result in more positive perceptions because they facilitate greater understanding. Nontraditional metrics were, however, less effective when audiences had interview experience. These results can be leveraged by practitioners and researchers who are interested in more effectively communicating validity information about the structured interview.

KEYWORDS

structured interview, validity, judgement and decision making, numeracy, nontraditional validity metrics

One of the biggest challenges facing organizational researchers continues to be convincing practitioners to adopt evidence-based practices, especially in the domain of personnel selection (Rynes, 2012; Rynes et al., 2002). Whereas structured interviews are more valid predictors of job performance, less susceptible to bias, and more legally defensible than unstructured interviews (Gatewood et al., 2010; Huffcutt & Arthur, 1994), practitioners often maintain a preference for unstructured interviews (Highhouse, 2008; Nolan et al., 2016; Rynes et al., 2002). The use of ineffective selection methods such as the unstructured interview limits the ability to select optimal job applicants. Failure to adopt structured interviews is partially driven by poor validity communication (Giluk & Rynes, 2012). If validity evidence is not conveyed in a compelling manner, hiring managers will not be persuaded to use the structured interview.

In this study, we examine how nontraditional validity metrics affect understanding of validity information, attitudes toward the validity metric, and the perceived usefulness of the structured interview for differentially experienced audiences. This study adds to the existing literature (e.g., Brooks et al., 2014) by (a) examining a larger number of validity metrics, (b) determining if the effects of nontra-

ditional metrics depend on one's numeracy, (c) specifying why nontraditional metrics result in positive judgements toward the structured interview, and (d) comparing validity communication across audiences with differing amounts of interview experience. Better understanding the factors that lead to effective validity communication represents an important step in promoting evidence-based selection practices.

Nontraditional Validity Metrics

For validity information to be properly evaluated, it is important for this information to be well-understood and meaningful. Unfortunately, the traditional validity metrics (e.g., r and R^2) that are commonly used by researchers are neither well-understood nor meaningful for most audiences (Bridgeman et al., 2009; Brooks et al., 2014; Krasikova et al., 2018; Kuncel & Rigdon, 2012). To form positive judgments of numerical information, people must be able to evaluate the "goodness" of a number (Peters et al., 2009;

Corresponding author:
Nathaniel M. Voss
Email: nmvoss@ksu.edu

Slovic et al., 2007). If it is unclear whether a particular numerical value can be considered good, numbers cannot be accurately evaluated (e.g., Hsee & Zhang, 2010). Subsequently, people will be unable to understand or form positive judgments of the numerical information.

Conceivably, many of the problems that stem from the use of traditional validity metrics concern the evaluability of the numbers that are employed. For instance, because correlation coefficients (r) and coefficients of determination (R^2) rely on numerical formats (e.g., decimal r values ranging from -1 to +1) and concepts (e.g., shared variance) that are less likely to be encountered by non-researchers, they are not readily evaluable. To address this, alternative validity metrics that rely on more evaluable numerical formats, such as probabilities, have been developed. As an example, the Common Language Effect Size (CLES) metric is created by transforming correlation coefficients into a probability-based metric, with values ranging from 0% to 100%.[1] An array of nontraditional metrics have been developed to aid evaluability. These nontraditional metrics tend to be better understood and judged more favorably than traditional validity metrics (Brooks et al., 2014; Zhang, 2018).

For example, both Brooks et al. (2014) and Zhang (2018) found that the CLES, and an additional nontraditional metric, the Binomial Effect Size Display (BESD), were better understood than either the r and R^2 metrics. Further, Brooks et al. (2014) found that participants perceived interventions as more effective and were willing to pay more for such interventions when nontraditional metrics were used. Ultimately, it is the differences in the language of these various metrics that is impacting people's perceptions of the information that they are presented with such that the language employed by nontraditional metrics is more familiar and straightforward than the language employed by traditional metrics (see May, 2004). We extend the results of previous research (e.g., Brooks et al., 2014) by examining a larger number of validity metrics across different audiences and determining both when and why nontraditional metrics are effective validity communication tools (described below).

Hypothesis 1: Use of nontraditional validity metrics will be positively related to (a) understanding of the validity information, (b) attitudes toward the validity metric, and (c) the perceived usefulness of the structured interview.

Numeracy

Numeracy is an important concept within the judgment and decision making literature that has been implicated in effective decision making across domains (Peters, 2012). Numeracy is an individual difference referring to an "ability to comprehend, use, and attach meaning to numbers" (Nelson et al., 2008, p. 262). People with high levels of numeracy are able to engage in mathematical operations, use numbers in their decision making, are less susceptible to irrelevant, non-numeric information, and are better able to evaluate the meaning of numbers (Cokely et al., 2012; Lipkus & Peters, 2009; Peters et al., 2006). Accordingly, people with higher levels of numeracy should better understand validity information presented to them, form more positive attitudes toward the metric used to communicate the validity information, and also form more positive impressions of the content of the message (perceive the structured interview as more usefulness in this case). Because numeracy represents and individual difference variable, it is conceivable that numeracy will interact with the presentation format of the validity information to jointly influence people's understanding and perceptions of the information (see Zhang et al., 2018). In this study, participants are shown validity information about the structured and unstructured interview, with the structured interview having stronger validity. Regardless of the metric (traditional vs. nontraditional), those with high numeracy should find the validity information more convincing, leading to high understanding and positive perceptions of the higher validity structured interview. Those with low numeracy, however, are more likely to have higher understanding, more positive attitudes, and perceive the structured interview as usefulness when nontraditional metrics are used.

Hypothesis 2: Numeracy moderates the relationship between nontraditional validity metrics and (a) understanding of the validity information, (b) attitudes toward the validity metric, and (c) the perceived usefulness of the structured interview such that nontraditional metrics will be effective across all levels of numeracy (though slightly more effective for those with high numeracy) but that traditional metrics will only be effective for those with high numeracy.

Mediation Effects of Understanding

The above considerations indicate that nontraditional validity metrics enable greater understanding of numerical information and facilitate more positive numerical judgements. Because understanding should temporally precede the judgments that a person forms (i.e., it is unlikely that low understanding results in positive evaluations; Hibbard & Peters, 2003; Peters et al., 2009), understanding should mediate the relationship between nontraditional validity metrics and people's attitudes toward the validity metric and the perceived usefulness of the structured interview.

---

1 See Table S5 in the supplemental materials for examples of other nontraditional metrics.

In other words, the reason nontraditional validity metrics result in more positive judgments is because they facilitate greater numerical evaluability. Furthermore, consistent with the anticipated moderating effect of numeracy, this indirect effect between nontraditional metrics and people's attitudes of the validity metric and perceived usefulness of the structured interview (through understanding) is expected to be highest when people also have high numeracy. This expectation is consistent with the view that higher, relative to lower, levels of numeracy can substantially enhance the understanding of numerical information (Reyna et al., 2009; Zhang et al., 2018).

> **Hypothesis 3a:** *Nontraditional validity metrics lead to higher levels of understanding which will in turn lead to (a) more positive attitudes toward the validity metric (nontraditional metric → understanding → validity metric attitudes) and (b) greater perceived usefulness of the structured interview (nontraditional metric → understanding → structured interview usefulness).*

> **Hypothesis 3b:** *The indirect effect between nontraditional validity metrics and people's attitudes toward the validity metric and perceived usefulness of the structured interview is conditional on numeracy such that the indirect effect is highest when numeracy is also high.*

Validity Communication for Audiences With Differing Amounts of Interview Experience

Validity communication can be directed at both lay (e.g., Brooks et al., 2014) and experienced (e.g., Highhouse et al., 2017) audiences, but these audiences may differ in their receptivity to validity messages. People with first-hand experience with a given topic tend to rely heavily on their personal experience when forming evaluations and insufficiently adjust toward new educational information (Hoch, 2002). Hiring managers – those who have conducted interviews – will almost assuredly have used the unstructured interview, which many interviewers view quite favorably (Highhouse, 2008). Likewise, working adults – those who have experience being an interviewee – will likely have participated in the unstructured interview, which job applicants also view favorably (Diab et al., 2011). Interview experience can thus be conceptualized as consisting of both interviewer (i.e., hiring managers) and interviewee (i.e., working adults) experience. If such personal experiences and pre-existing viewpoints conflict with the validity information presented, experienced audiences should find it minimally persuasive. In contrast, lay audiences – those without first-hand experience conducting interviews and little experience participating in interviews (e.g., university students) – should be relatively open to considering evidence about the structured interview as they will not rely on their experiences as extensively.

> **Hypothesis 4:** *The relationship between nontraditional metrics and the perceived usefulness of structured interview is moderated by experience such that for participants with less interview experience, nontraditional metrics will result in greater perceptions of the usefulness of the structured interview than those with more interview experience.*

METHOD

Participants

Students. A total of 267 students from psychology courses participated in this study for course credit. Of these, 21 either did not provide complete data or failed to pass the self-report attention checks and were thus were not included in the analyses. Of the remaining participants, 44% identified as male, 83% identified as White/Caucasian, and the average age was 19 ($SD = 2.19$).

Working adults. A total of 274 working adults without any interviewing experience were recruited from Amazon's Mechanical Turk and were compensated $1.50. A carefully designed, disguised-purpose screening survey administered to 2,000 people on Mechanical Turk allowed us to identify and recruit 274 working adults without hiring experience.[2] The disguised-purpose nature of this screening survey (i.e., the purpose of this study was not explicitly stated) ensured that participants did not falsify information in hopes of being eligible for future studies. Participants completed the present study approximately 6 months after completing this screening survey. Of these working adults, 29 failed to provide complete data and were not included in the analyses. Of the remaining participants, 45% identified as male, 72% identified as White/Caucasian, 76% were employed full time, and the average age was 34 ($SD = 10.33$).

Hiring managers. The same carefully designed, disguised-purpose screening survey described above also allowed us to identify and recruit 155 people with hiring experience (i.e., people who have interviewed at least one to five people). Participants completed this study approximately 10 months after completing the screening survey and were compensated $1.50. Of the participants, 11 either did not provide complete data or failed to pass the self-report attention checks and were thus not included in the analyses. Of the remaining participants, 44% identified as male, 78% identified as White/Caucasian, 89% were employed full time, 96% have hired at least six people throughout their career, and the average age was 43 ($SD = 10.77$).

---

2   See Figures S1 and S2 in the supplemental materials for a screenshot of this screening survey.

Design and Procedure

In this between-subjects experiment, participants were randomly assigned to one of eight possible conditions, each containing a unique validity metric. The metrics used in the present study were: Common Language Effect Size (Dunlap, 1994; McGraw & Wong, 1992), Binomial Effect Size Display (Rosenthal & Rubin, 1982), correlation coefficient (r), coefficient of determination (R^2), risk ratio, odds ratio, natural frequencies (Hoffrage et al., 2000), and Taylor Russell Table values (Taylor & Russell, 1939). We utilized validity coefficients from Huffcutt and Arthur's (1994) meta-analysis for the unstructured ($r = .20$) and structured ($r = .57$) interview as baseline values for the computation of the seven remaining metrics, which were computed to be mathematically equivalent across all eight conditions.

Prior to viewing the validity metric, participants viewed a basic definition of unstructured and structured interviews and a definition of the validity metric they would be shown (see Brooks et al., 2014). The definition of the structured interview emphasized the consistency in question content, question order, and the use of a detailed scoring method. The definition of the unstructured interview emphasized the lack of consistency in question content, question order, and the use of intuition (see Levashina et al., 2014). Details about the definitions of the validity metrics that were provided can be found in the supplemental materials. Participants then viewed a statement concerning the effectiveness of the unstructured and structured interview using the randomly assigned validity metric. Participants then answered a self-reported, attention-check question (i.e., we asked if participants read the information). They then completed a series of self-report measures concerning their understanding of the information, attitudes toward the validity metric, and perceptions of the usefulness of the structured interview. Finally, participants completed two unique numeracy measures. All samples followed this same procedure.

Measures

Similar to prior research (Brooks et al., 2014), we used a series of 3-item measures to assess understanding (e.g., "I easily understood the information about the effectiveness of the structured and unstructured interviews"), attitudes toward the validity metric (e.g., "I liked the metric used to present the information about the effectiveness of the structured and unstructured interview"), and perceptions of the usefulness of the structured interview (e.g., "I believe that the structured interview is an effective technique for hiring job candidates"). To ensure a comprehensive assessment of the numeracy construct, we included two measures of numeracy that differ in their operationalization: The Berlin Numeracy Test (BNT; Cokely et al., 2012) and the Subjective Numeracy Scale (SNS; Fagerlin et al., 2007). The BNT requires participants to complete a series of mathematical problems and employs an adaptive structure to classify peoples' level of numeracy to a percentile ranging from 1 (bottom 25%) to 4 (top 25%), whereas the SNS is a self-report measure of people's numeracy.

RESULTS

Information regarding the correlations between study variables can be found in the supplemental materials. Descriptive statistics for the three outcomes variables across all eight conditions can be seen in Table 1. Prior to testing our main hypotheses, we first looked for differences in effectiveness among the nontraditional metrics (Table 1). Finding no evidence of differing effectiveness, we followed Brooks et al. (2014) and collapsed across the six nontraditional metrics and the two traditional metrics to form a new variable with two levels (e.g., Level 1 = nontraditional metric, Level 2 = traditional metric). This reduces the number of statistical tests conducted while focusing on differences between traditional and nontraditional metrics.

Next, to test Hypothesis 1, which concerned the effectiveness of nontraditional validity metrics, we conducted a series of regression analyses on the pooled sample using validity metric, numeracy, experience, the metric-by-numeracy, and metric-by-experience interactions as predictors. As Table 2 shows, nontraditional metrics led to enhanced understanding ($\beta = .20$), more positive attitudes toward the metric ($\beta = .16$), and greater perceptions of the usefulness of the structured interview ($\beta = .09$), supporting Hypotheses 1a, 1b, and 1c, respectively. Hypothesis 2 concerned the moderating effect of numeracy. Although there was a main effect of numeracy (SNS) across the three outcome variables, there were no significant metric-by-numeracy interactions, thereby indicating that the effects of the nontraditional validity metrics are equally effective across differing levels of numeracy (see Table 2). Hypothesis 2 was therefore not supported.

To test Hypotheses 3, which concerned the mediating role of understanding, we next conducted a series of mediation analyses using ordinary least squares path analysis (Hayes, 2013) with nontraditional metric as the predictor, understanding as the mediator, numeracy (SNS) as the moderator between nontraditional metric and understanding, and (a) attitudes toward the metric and (b) the perceived usefulness of the structured interview as the outcomes (two models; see Table 3). Because we did not observe any effects of numeracy when the BNT was used, we excluded this variable in this analysis. For all indirect effects, confidence intervals (CIs) were calculated using a bias-corrected bootstrap with 10,000 resamples (Preacher & Hayes, 2004). As shown in Table 3, across all samples, understanding mediated the relationship (i.e., the CIs did not contain 0) between nontraditional metrics and (a) attitudes toward the

TABLE 1.

Summary of the Reliability Estimates and Means for Understanding, Metric Attitudes, and the Perceived Usefulness of the Structured Interview Across All Eight Conditions

Outcome variable	α	Model F	η^2	Mean	Traditional metrics		Nontraditional metrics					
					r	R^2	CLES	BESD	RR	OR	NF	TRT
Pooled samples												
Understanding	.83	5.96**	.06	4.20	4.12$_a$	3.67$_b$	4.28$_a$	4.26$_a$	4.37$_a$	4.34$_a$	4.09$_{ab}$	4.43$_a$
Metric attitudes	.88	3.95**	.04	3.47	3.42$_{ab}$	2.99$_b$	3.48$_a$	3.50$_a$	3.60$_a$	3.60$_a$	3.43$_{ab}$	3.76$_a$
Structured interview usefulness	.85	1.58	.02	3.95	3.90	3.70	3.98	3.95	4.06	4.00	3.93	4.10
Students												
Understanding	.83	3.49**	.09	3.85	3.52$_{bc}$	3.29$_c$	3.94$_{abc}$	3.90$_{abc}$	4.13$_{ab}$	3.93$_{abc}$	3.73$_{abc}$	4.35$_a$
Metric attitudes	.84	2.43*	.07	3.19	3.08$_{ab}$	2.67$_b$	3.13$_{ab}$	3.24$_{ab}$	3.43$_{ab}$	3.32$_{ab}$	3.02$_{ab}$	3.62$_a$
Structured interview usefulness	.77	1.40	.04	3.76	3.52	3.47	3.79	3.82	4.05	3.79	3.72	3.94
Working adults												
Understanding	.84	5.59**	.14	4.38	4.53$_a$	3.77$_b$	4.34$_a$	4.44$_a$	4.55$_a$	4.73$_a$	4.20$_{ab}$	4.53$_a$
Metric attitudes	.87	1.71	.05	3.65	3.75	3.25	3.60	3.51	3.84	3.68	3.69	3.94
Structured interview usefulness	.87	1.26	.04	4.11	4.24	3.85	4.04	3.99	4.26	4.14	4.12	4.28
Hiring managers												
Understanding	.83	1.81	.08	4.48	4.57	3.96	4.70	4.71	4.50	4.39	4.56	4.39
Metric attitudes	.93	1.48	.07	3.64	3.51	2.98	3.81	3.96	3.50	3.89	3.75	3.70
Structured interview usefulness	.86	0.57	.03	3.99	4.06	3.83	4.16	4.06	3.72	4.10	3.98	4.05

Note. *p < .05 **p < .01. α = coefficient alpha. η^2 = eta-squared. r = correlation. R^2 = coefficient of determination. *CLES* = Common Language Effect Size. *BESD* = Binomial Effect Size Display. *RR* = risk ratio. *OR* = odds ratio. *NF* = natural frequencies. *TRT* = Taylor Russell Table. One-way ANOVAs were conducted for each outcome. For significant models (i.e., the individual rows), mean values not connected by the same letter are significantly different (p < .05; Tukey post hoc test).

validity metric (indirect effect = .11) and (b) the perceived usefulness of the structured interview (indirect effect = .08). There were no significant interactions between nontraditional metrics and numeracy on understanding for either model, thereby indicating that the indirect effect between nontraditional metrics and the outcome variables (through understanding) is constant across all levels of numeracy. These results provide support for Hypothesis 3a but not Hypothesis 3b.

Finally, to test Hypothesis 4, which concerned the effect of interview experience, we conducted a regression analysis and computed the interactions between experience and metric for the three outcome variables (as described above; see Table 2). There was a significant metric-by-experience interaction for students on both understanding and the perceived usefulness of the structured interview such that students had a higher understanding and perceived the structured interview as more useful when nontraditional metrics, compared to traditional metrics, were used. Despite hiring managers reporting a higher understanding of the information (β = .20), their perceptions of the usefulness of the structured interview were not affected by nontraditional validity metrics (as evidenced by the null metric-by-experience interaction for hiring managers). To further illustrate

TABLE 2.

Summary of Regression Parameter Estimates for Understanding, Metric Attitudes, and the Perceived Usefulness of the Structured Interview Across All (Pooled) Samples

Variable	Understanding				Metric attitudes				Structured interview usefulness			
	t-ratio	β	B	B SE	t-ratio	β	B	B SE	t-ratio	β	B	B SE
Metric (1 = nontraditional)	5.17	.20**	.20**	.04	3.97	.16**	.18**	.05	2.21	.09*	.08*	.04
Numeracy (BNT)	0.79	.04	.03	.03	0.45	.02	.02	.04	1.50	.07	.05	.03
Numeracy (SNS)	2.66	.12**	.09**	.03	2.62	.13**	.11**	.04	3.03	.15**	.10**	.03
Students	-7.36	-.40**	-.40**	.05	-4.31	-.25**	-.28**	.06	-3.72	-.21**	-.20**	.05
Hiring managers	3.62	.20**	.23**	.06	0.93	.05	.07	.07	0.08	.00	.00	.06
Metric x Numeracy (BNT)	-0.26	-.01	-.01	.03	-0.27	-.01	-.02	.04	-0.77	-.04	-.02	.03
Metric x Numeracy (SNS)	0.96	.04	.03	.03	1.03	.05	.04	.04	0.67	.03	.02	.03
Metric x Students	2.14	.12*	.12*	.05	0.66	.04	.04	.06	2.07	.12*	.11*	.05
Metric x Hiring Managers	-1.24	-.06	-.08	.06	0.49	.03	.04	.07	-1.05	-.06	-.07	.06

Note. *p < .05 **p < .01. BNT = Berlin Numeracy Test. SNS = Subjective Numeracy Scale. Metric was effect coded so nontraditional metric = 1. The student and hiring manager samples were effect coded as 1 with the working adult sample serving as the baseline group.

TABLE 3.

Summary of the Mediation Results

	Mediation Model 1		Mediation Model 2	
Predictor	Nontraditional metric		Nontraditional metric	
Moderator	Numeracy (SNS)		Numeracy (SNS)	
Mediator	Understanding		Understanding	
Outcome	Metric attitudes		Structured interview usefulness	
	Indirect effect	95% CI	Indirect effect	95% CI
Pooled samples	**.11**	.06, .17	**.08**	.05, .12
Students	**.14**	.07, .23	**.12**	.06, .20
Working adults	**.09**	.03, .16	**.06**	.02, .10
Hiring managers	.08	-.02, .19	.04	-.01, .10

Note. SNS = Subjective Numeracy Scale. Metric was effect coded so nontraditional metric = 1. 95% CI = bootstrapped 95% confidence interval. The first CI value represents the lower limit while the second value represents the upper limit. Indirect effects represent the mean indirect effect across different levels of numeracy. Indirect effects where the CI does not contain zero appear in bold. Numeracy was specified as a moderator for the first path of the mediation model (nontraditional metric → understanding).

the effects of experience, we conducted a subgroup analysis for each sample separately.[3] In general, smaller effect sizes were observed as experience increased. For example, the effects of nontraditional metrics on the perceived usefulness

of the structured interview were greater for students (β = .20) than hiring managers (β = .03). A similar trend was also observed for the indirect effects of the mediation analyses (Table 3). Collectively, these results provide support for Hypothesis 4 such that validity communication was less effective for those with more interview experience.

---

3   See the supplemental materials for the parameters for each sample (Table S4).

DISCUSSION

These findings align with other research showing the advantages of nontraditional metrics (e.g., Brooks et al., 2014). Nontraditional metrics were better received than traditional metrics (e.g., participants had more positive attitudes toward nontraditional metrics) and facilitated greater understanding of validity information. Nontraditional metrics also increased participant's perceptions of the usefulness of the structured interview, both directly and indirectly, by increasing understanding. Furthermore, the nontraditional metrics were equally effective for participants with differing levels of numeracy.

This work helps explain the processes by which nontraditional metrics are effective. For instance, nontraditional metrics enable understanding, and this understanding leads to greater perceptions of the usefulness of the structured interview. This fits with the organizational change literature, which indicates that understanding facilitates willingness to change (Erwin & Garman, 2010). Finally, consistent with our expectation that people may rely on their previous interview experience when forming their perceptions of the structured interview, we found it difficult to improve hiring managers' perceptions of the usefulness of the structured interview. Even though they preferred and better understood nontraditional metrics, this did not seem to translate into stronger preferences for the structured interview.

Although the effects of experience are largely consistent with our expectations, the specific mechanisms by which experience operates cannot be derived from this study. We assumed that hiring managers would rely on their personal experience with unstructured interviews, which they view favorably (Highhouse, 2008), when judging information about the structured interview. It is likely, however, that hiring managers consider more than just validity evidence when evaluating the utility of hiring methods. For example, hiring managers may acknowledge that structured interviews have higher validity than unstructured interviews but still not believe they are useful due to the practical challenges (e.g., lack of knowledge) of implementing structured interviews. Alternatively, hiring managers could again acknowledge the higher validity of structured interviews but not perceive them as useful due to fears of not receiving credit for their decisions when using structured interviews (Nolan et al., 2016). This, coupled with the negative reactions to structured hiring methods that employees often have (Diab et al., 2011), suggests that experience may be a particularly difficult barrier to overcome when communicating validity information.

Implications

These findings suggest that scholars aim their validity communication efforts toward enhancing understanding. For scholars, this means seeking feedback and checking with recipients to ensure understanding. Pedagogically, this is similar to an assessment conversation, whereby teachers engage in discussions with an audience to gauge comprehension of a topic (see Duschl & Gitomer, 1997). A conversational approach may be of particular benefit to those with low numeracy preferences, for whom a statistical report would not be very engaging.

This work bolsters the conclusion that nontraditional validity metrics are a better means of communication than traditional metrics (e.g., Brooks et al., 2014) and expands previous research by examining a larger number of validity metrics and showing that nontraditional metrics are equally effective for people with differing levels of numeracy. Taylor Russell Table metrics (Taylor & Russell, 1939) were generally liked and understood. Interestingly, this metric was not especially liked or understood by hiring managers, implying that specific metrics could vary in effectiveness depending on the audience. More to this point, one of the main findings from this study is that nontraditional metrics were less effective for more experienced audiences. Although this does not necessarily limit the utility of nontraditional metrics, it does suggest that additional strategies for conveying the usefulness of the structured interview will need to be employed when audiences have interview experience.

Limitations and Future Research

Experience with interviews was a determinant of participant's perceptions of the usefulness of the structured interview. A potential limitation was that interview experience was inferred from sample membership. Additional research is needed to investigate how one's personal interview experiences (e.g., number of unstructured/structured interviews conducted) relate to rigidity in perceptions of the structured interview. It is also important to note that hiring managers tended to be older compared to the other samples. Future research should ensure that it is experience, and not age, that is responsible for the results report here.

Also, although numeracy did not function as a moderating variable, a main effect of numeracy was observed, but only for the SNS. This was somewhat surprising because both the SNS and BNT are measures of the numeracy construct. It may be the case that it is one's self-reported numerical preferences (e.g., self-perceived confidence with numbers), rather than a person's actual numerical ability, that determines one's perceptions of validity information. Although we are unable to account for this inconsistency in this study, disentangling the ways in which different operationalizations of numeracy (see Reyna et al., 2009) are differentially related to perceptions of validity information represents a fruitful area for future research.

Additionally, although our decision to focus on the employment interview as the primary stimulus was driven by our desire to address the research–practice gap in this

area, it is possible that the results reported here do not generalized to other stimuli (e.g., other selection tools). Future research is needed to explore this possibility. Finally, we should also note that because understanding was a self-report measure, future research is needed to determine ways to assess objective understanding.

REFERENCES

Bridgeman, B., Burton, N., & Cline, F. (2009). A note on presenting what predictive validity numbers mean. Applied Measurement in Education, 22, 109-119.

Brooks, M., Dalal, D., & Nolan, K. (2014). Are common language effect sizes easier to understand than traditional effect sizes? Journal of Applied Psychology, 99, 332-340.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. Judgment and Decision Making, 7, 25-57.

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. International Journal of Selection and Assessment, 19, 209-216.

Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. Psychological Bulletin, 116, 509-511.

Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. Educational Assessment, 4, 37-73.

Erwin, D. G., & Garman, A. N. (2010). Resistance to organizational change: Linking research and practice. Leadership and Organizational Development Journal, 31, 39-56.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. Medical Decision Making, 27, 672-680.

Gatewood, R. D., Feild, H. S., & Barrick, M. (2010). Human resource selection (7th Edition). Mason, OH: Thompson South-Western.

Giluk, T. L., & Rynes, S. L. (2012). Research findings practitioners resist: Lessons for management academics from evidence-based medicine. In D. M. Rousseau (Ed.) The Oxford handbook of evidence-based management, (pp. 130-164). New York, NY: Oxford University Press.

Hayes, A. F. (2013). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. New York, NY: Guilford Press.

Hibbard, J. H., & Peters, E. (2003). Supporting informed consumer health care decisions: Data presentation approaches that facilitate the use of information in choice. Annual Review of Public Health, 24, 413-433.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. Industrial and Organizational Psychology: Perspectives on Science and Practice, 1, 333-342.

Highhouse, S., Brooks, M. E., Nesnidol, S., & Sim, S. (2017). Is a .51 validity coefficient good? Value sensitivity for interview validity. International Journal of Selection and Assessment, 25, 383-389.

Hoch, S. J. (2002). Product experience is seductive. Journal of Consumer Research, 29, 448-454.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. Science 290, 2261-2262.

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. Perspectives on Psychological Science, 5, 343-355.

Huffcutt, A., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. Journal of Applied Psychology, 79, 184-190.

Krasikova, D. V., Le, H., & Bachura, E. (2018). Toward customer-centric organizational science: A common language effect size indicator for multiple linear regressions and regressions with higher-order terms. Journal of Applied Psychology, 103, 659-675.

Kuncel, N. R., & Rigdon, J. (2012). Communicating research findings. In N. W. Schmitt & S. Highhouse (Eds.), Handbook of psychology: Vol 12. Industrial and organizational psychology (2nd ed., pp. 43–58). New York, NY: Wiley.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. Personnel Psychology, 67, 241-293.

Lipkus, I. M., & Peters, E. (2009). Understanding the role of numeracy in health: Proposed theoretical framework and practical insights. Health Education & Behavior, 36, 1065-1081.

May, H. (2004). Making statistics more meaningful for policy research and program evaluation. American Journal of Evaluation, 25, 525–540.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. Psychological Bulletin, 111, 361-365.

Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I., & Peters, E. (2008). Clinical implications of numeracy: Theory and practice. Annals of Behavioral Medicine, 35, 261-274.

Nolan, K. P., Carter, N. T., & Dalal, D. K. (2016). Threat of technological unemployment: Are hiring managers discounted for using standardized employee selection practices? Personnel Assessment and Decisions, 2, 30-47.

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. Current Directions in Psychological Science, 21, 31-35.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. Psychological Science, 17, 407-413.

Peters, E., Dieckmann, N. F., Västfjäll, D., Mertz, C. K., Slovic, P., & Hibbard, J. H. (2009). Bringing meaning to numbers: The impact of evaluative categories on decisions. Journal of Experimental Psychology: Applied, 15, 213-227.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behavior Research Methods, Instruments, & Computers, 36, 717-731.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. Psychological Bulletin, 135, 943-973.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74, 166-169.

Rynes, S. (2012). The research-practice gap in industrial-organizational psychology and related fields: Challenges and potential solutions. In S. W. J. Kozlowski (Ed.) The Oxford handbook of organizational psychology, volume 1 (pp. 409-452). New York, NY: Oxford University Press.

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. Human Resource Management, 41, 149-174

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. European Journal of Operational Research, 177, 1333-1352.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. Journal of Applied Psychology, 23, 565-578.

Zhang, D. C. (2018). Utility of alternative effect size statistics and the development of a web-based calculator: Shiny-AESC. Frontiers in Psychology, 9, 1221.

Zhang, D. C., Highhouse, S., Brooks, M. E., & Zhang, Y. (2018). Communicating the validity of structured job interviews with graphical visual aids. International Journal of Selection and Assessment, 26, 93-108.