Personnel Assessment and Decisions

Volume 5
Issue 3 *Advanced Technologies in Assessment: A Science–Practice Concern*

Article 3

2019

Validity Evidence for Off-the-Shelf Language-Based Personality Assessment Using Video Interviews: Convergent and Discriminant Relationships with Self and Observer Ratings

Louis Hickman
Purdue University

Louis Tay
Purdue University

Sang Eun Woo
Purdue University

Recommended Citation

Hickman, Louis; Tay, Louis; and Woo, Sang Eun (2019) "Validity Evidence for Off-the-Shelf Language-Based Personality Assessment Using Video Interviews: Convergent and Discriminant Relationships with Self and Observer Ratings," *Personnel Assessment and Decisions*: Number 5 : Iss. 3 , Article 3.
DOI: https://doi.org/10.25035/pad.2019.03.003
Available at: https://scholarworks.bgsu.edu/pad/vol5/iss3/3

Validity Evidence for Off-the-Shelf Language-Based Personality Assessment Using Video Interviews: Convergent and Discriminant Relationships with Self and Observer Ratings

Louis Hickman[1], Louis Tay[1], and Sang Eun Woo[1]

1. Purdue University

KEYWORDS

automated personnel assessment, personality, text mining

ABSTRACT

Technological advances have led to the development of automated methods for personnel assessment that are purported to augment or outperform human judgment. However, empirical research providing validity evidence for such techniques in the selection context remains scarce. In addressing this void, this study focuses on language-based personality assessments using an off-the-shelf, commercially available product (i.e., IBM Watson Personality Insights) in the context of video-based interviews. The scores derived from the language-based assessment were compared to self and observer ratings of personality to examine convergent and discriminant relationships. The language-based assessment scores showed low convergence with self-ratings for openness, and with self- and observer ratings for agreeableness. No validity evidence was found for extraversion and conscientiousness. For neuroticism, the patterns of correlations were in the opposite of what was theoretically expected, which raised a significant concern. We suggest more validation work is needed to further improve emerging assessment techniques and to understand when and how such approaches can appropriately be applied in personnel assessment and selection.

Recent technological and analytical advances have enabled the development of new methods for personnel assessment and selection. Today's technological innovations have led to automated methods for assessing job applicants that are purported to augment or even outperform human judgment (Hoffman, Kahn, & Li, 2018). Although such approaches hold potential for improving decision making, some approaches have been found to perpetuate existing biases (e.g., Amazon's resume screening tool; Dastin, 2018). This implies that more work is needed to validate automated approaches to help researchers and practitioners understand when, where, and how they can be applied to improve personnel assessment and selection.

Given the prevalence of off-the-shelf applications for automating personnel assessment and selection (Zielinski, 2018), many organizations are now turning to these solutions in their hiring practices as they can lead to substantial cost savings compared to developing similar solutions in house or relying on manual ratings (e.g., interview raters). For example, low cost off-the-shelf simulations are available (Boyce, Corbet, & Adler, 2013), automated testing systems can improve selection outcomes (Hoffman et al., 2018), and statistical algorithms to rank candidates can outperform expert judgment (Kuncel, Klieger, Connelly, & Ones, 2013). Critically, the influx of off-the-shelf solutions requires careful validation, which includes determining whether assessment results from these solutions are aligned with those derived from more traditional measurement approaches that have established evidence of validity.

Automated Personality Assessment in Selection

Personality traits are widely recognized as key individual-level predictors of job performance (Dudley, Orvis, Lebiecki, & Cortina, 2006; Ones, Viswesvaran, & Dilchert,

Corresponding author:
Louis Hickman
Email: lchickma@purdue.edu

2005). Personality produces relatively lower adverse impact than general mental ability (Ryan, Ployhart, & Friedel, 1998) yet may have equivalent or superior predictive validity (e.g., conscientiousness; Connelly & Ones, 2010). As such, personality is increasingly used in personnel selection.

Although self-reports are the most common method of personality assessment, there are concerns about faking and self-presentation biases (see Hough & Oswald, 2008). In this vein, observer ratings based on observable behaviors (e.g., language) may be used to overcome undesirable response distortions in self-reported personality assessment. Indeed, personality ratings from coworkers, family, and friends have been found to have validities roughly double the magnitude of self-reports (Oh, Wang, & Mount, 2011).

Recently, researchers have sought to apply automated, language-based models as alternatives to self-reports for assessing personality (e.g., Kern et al., 2016; Park et al., 2015; Schwartz et al., 2013; Youyou, Kosinski, & Stillwell, 2015). These approaches have been imported into off-the-shelf applications for personnel assessment and selection. For example, a variety of language-based assessments have been integrated in platforms assessing applicant personality in video interviews (e.g., HireVue; Quantified Communications), and Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, Boyd, & Francis, 2015) is part of the Receptiviti language-based executive and leadership assessment system (Receptiviti, n.d.).

Despite the potential promise of language-based models for assessing personality in the selection context, most investigations to date have been validated in the context of social media language use (e.g., IBM, 2018). It is unknown whether language-based models trained on social media text are effective at predicting personality when applied to a selection setting. Models trained on social media text may translate poorly to workplace applications.

In the present article, we seek to address this gap to understand whether off-the-shelf language-based models for assessing personality can be effectively applied to the selection context. Specifically, we investigate the convergent-discriminant validity evidence of an off-the-shelf language-based personality assessment tool validated for assessing personality on social media in the context of video interviews. It is unknown whether language-based personality models trained on social media text can be validly applied in other contexts. A similar study did the reverse: They applied a language-based model of personality created on workplace emails to social media, finding that it underperformed compared to existing solutions (Golbeck, 2017). This follows past work validating off-the-shelf technological solutions for personnel assessment, such as the convergence between human ratings and automated algorithms for achievement record scoring (Campion, Campion, Campion, & Reider, 2016), the potential to identify deceptive impression management in employment interviews (Auer, 2018), and automatically assessing applicant interview performance (Naim, Tanveer, Gildea, & Hoque, 2018).

The Present Study

Our goal is to provide initial evidence regarding the convergence of an off-the-shelf, language-based personality assessment with self and observer ratings of personality. Specifically, we selected IBM Watson Personality Insights (PI) because of its prominence and usage in higher education, finance, and in other software solutions (HG Insights, 2018). To create IBM Watson PI, IBM recruited a set of active Twitter users to complete self-reports of personality (IBM, 2018). Their tweets were then analyzed via global vector for word representation (GloVe; Pennington, Socher, & Manning, 2014), which estimates the similarity of words using the frequency of their co-occurrence and their proximity in texts within the training corpus. The similarity is represented by a vector: Vectors close in value suggest words are similar in meaning, whereas vectors far apart in value suggest words are dissimilar in meaning. Those vectors are then fed into a machine learning algorithm to predict personality traits from language use. Although IBM Watson PI provides information in its documentation about its convergence with self-reports of personality when assessing social media language use, it is unknown whether IBM Watson PI can reliably and validly assess personality in selection contexts such as video interviews. IBM Watson PI's personality assessment has been applied to identify cyber bullies on Twitter (Balakrishnan, Khan, Fernandez, & Arabnia, 2019), but validity has not been assessed outside of social media. To the extent that it shows promise for mass, unproctored video interviews for personality assessment, it could lead to substantial cost savings for organizations. To the extent that it cannot, it would indicate that language-based models of personality trained on social media may be less useful for selection contexts.

Using automated approaches for assessment in selection represents a high level of structure. All candidates are asked the same questions, and all are judged using the same criteria. To assess personality, open-ended questions should be used because they provide fewer behavioral constraints (i.e., lower situational strength), increasing the variety of acceptable responses, and thereby obtaining the freest expression of behavior and personality-relevant information (Blackman, 2002). Participants recorded their responses to an open-ended prompt, and then we transcribed those responses and used IBM Watson PI to assess their personality. We assessed the extent to which personality scores converged with self-reports and reports from observers who rated participants' personality based on behaviors and speech observed in the videos. Doing so provided an initial investigation into the convergent and discriminant validity of off-the-shelf language-based personality assessment in interview settings.

METHOD

Participants and Procedure

We recruited 180 participants via Amazon Mechanical Turk (MTurk) who participated in exchange for $2. MTurks are generally demographically diverse, better representing the general population than college students and providing data at least as reliable as student samples (Woo, Keith, & Thornton, 2015). Participants recorded their responses to the following open-ended prompt:

> Talk about a topic or a story that you know and is personal to you. Do not hesitate to talk about your feelings and do not limit your answer to simple descriptions. Options include: 1. a personal experience (traveling, childhood memory, recent event). 2. your dreams (career, love, friends, hobbies). 3. your general views on a matter you feel strongly about.

Of the 180 videos submitted, two were removed from further analyses because the participants read content verbatim from a website, and one other was removed because it had no audio, leaving a final sample of 177 videos (61% female; 80% White). Participants were instructed to make their videos 2-4 minutes in length (*M* = 3 min 10 s; *SD* = 40 s; range 0 min 30 s-5 min 38 s). Participants also responded to a self-report personality questionnaire.

Three doctoral students first rated participants' personality for a set of sample videos, discussed behavioral cues and sources of agreement and disagreement, then independently watched the remaining videos and rated participants' personality. Then the videos were transcribed using Google Cloud Speech-to-Text, and the transcription was entered into IBM Watson PI to obtain its personality assessment.

Measures

Self-reports of personality. A 60-item personality survey comprised the 50-item International Personality Item Pool (IPIP; Goldberg, 1999) scale of markers for the FFM (Goldberg, 1992) and the Ten-Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003) was used to assess the FFM. The IPIP consists of 10 items for each FFM trait, whereas the TIPI consists of two items for each FFM trait. The combined scale consisted of 12 items for each FFM trait. All five scales showed acceptable internal consistency (see Cronbach's alpha values in Table 1).

Observer ratings on video data. Three doctoral students in industrial-organizational psychology watched the videos and assessed each participant's personality. The TIPI was adapted from a self-report to an other report by asking raters the extent to which the participant appeared to fulfill each of the TIPI's items (e.g., "Extraverted, enthusiastic"). The TIPI was chosen over other measures to reduce the time required to provide ratings. Ratings were averaged, and interrater reliabilities were adequate for all FFM traits (see Table 1). More visible traits such as extraversion and conscientiousness had higher intraclass correlations, whereas less visible traits such as openness and neuroticism (Allik, Realo, Mõttus, & Kuppens, 2010) had lower intraclass correlations.

Off-the-shelf language-based assessment through IBM Watson PI. IBM Watson PI originally used a closed vocabulary approach, including elements of LIWC (Pennebaker, Mehl, & Niederhoffer, 2003), to estimate personality. However, recent advances in text mining have led to the adoption of open vocabulary approaches that inductively associate words and/or phrases with outcomes of interest—in this case, personality traits. IBM Watson PI was developed and validated by using Twitter content to predict self-reports of personality (IBM, 2018). They do not provide specific trait correlations from the initial development work, but they do provide two summary scores describing the overall accuracy of the system's personality predictions from social media language. The mean absolute error (MAE) indexes the difference between self-reported and predicted personality scores, with 0 indicating no error and 1 indicating total error. The average correlation is the average correlation between self-reported and predicted personality scores across all the FFM traits. For the English language version, they reported that the average MAE was .12 and that the average correlation was .33 (IBM, 2018). By way of comparison, Schwartz et al.'s (2013) open vocabulary approaches for predicting personality from social media usage achieved a maximum average correlation of .35 across the FFM traits, and Park et al. (2015) achieved average correlation of .39, suggesting that IBM Watson PI has accuracy comparable to state-of-the-art text mining approaches. Additionally, IBM found that Watson PI's inferred personality traits predicted a variety of consumption preferences, suggesting it holds potential for predicting real world behavior. Of the 177 interview videos in the current study, IBM Watson PI was able to provide personality scores for 166 of the transcripts.

RESULTS

Table 1 displays the means, standard deviations, Cronbach's alpha coefficients (for the self-reported personality), interrater reliabilities (for the observer-rated personality), and correlations for the three sources of personality ratings. Intercorrelations for IBM Watson PI scores ranged from -.29 (between agreeableness and openness) to .70 (between conscientiousness and neuroticism). Some of these correlations did not conform to the expected patterns of association

TABLE 1.

Correlation matrix of IBM Watson PI, Self-, and Observer Ratings of FFM of Personality

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. PI Extraversion	4.03	.27	-														
2. PI Openness	5.68	.18	.12	-													
3. PI Conscientiousness	4.68	.25	*.50*	-.08	-												
4. PI Agreeableness	5.47	.23	.15	*-.29*	*.34*	-											
5. PI Neuroticism	3.70	.29	*.54*	.03	*.70*	-.01	-										
6. Self-report E	4.04	1.27	.06	.01	.03	.14	.01	(.92)									
7. Self-report O	5.48	.82	-.04	.18	-.03	-.01	-.06	*.34*	(.83)								
8. Self-report C	5.05	1.09	.02	.06	.05	-.05	.07	.09	*.26*	(.90)							
9. Self-report A	5.54	.91	-.07	.09	.07	.17	-.03	*.32*	*.47*	*.22*	(.86)						
10. Self-report N	3.75	1.35	-.06	-.13	-.15	.09	*-.20*	*-.32*	-.12	*-.51*	*-.26*	(.90)					
11. Observer E	4.30	1.14	.07	.07	.04	-.06	.08	*.22*	.19	.00	.14	-.13	(.78)				
12. Observer O	4.31	.84	.03	.08	-.14	-.11	-.13	.16	.12	-.15	.05	.06	.17	(.72)			
13. Observer C	4.65	.89	-.02	.15	.01	-.17	-.01	.08	.02	.07	.01	-.13	*.31*	*.35*	(.80)		
14. Observer A	4.72	.84	.03	-.14	.11	*.20*	-.02	.12	-.06	.04	.15	-.01	*.20*	*.27*	*.31*	(.76)	
15. Observer N	3.48	.99	*-.34*	-.10	*-.36*	.08	*-.41*	-.15	-.01	*-.20*	-.02	*.31*	*-.34*	*-.22*	*-.39*	*-.30*	(.66)

reliability	🟦
validity/MTHM	🟧
HTMM	🟩
HTHM	⬜

Note. The diagonal reports reliability in parentheses, using intraclass correlation for observer ratings and Cronbach's alpha for self-ratings. E=extraversion. O=openness to experience. C=conscientiousness. A=agreeableness. N=neuroticism. Underline indicates $p < .05$. *Italics* indicate $p < .01$. IBM Watson PI $N = 166$, self- and observer ratings $N = 177$. For IBM Watson PI correlations, $p < .05$ when $r > .15$ and $p < .01$ when $r > .19$, while for self- and observer rating correlations, $p < .05$ when $r > .14$ and $p < .01$ when $r > .19$

from the personality literature. In particular, neuroticism was highly positively correlated with both extraversion and conscientiousness (*r*s = .70 and .54, respectively). This unusual pattern was reflected in IBM Watson PI neuroticism scores having negative correlations with self and observer ratings of neuroticism, whereas the remaining traits had positive correlations with self- and observer ratings.

For the self-report scale, alpha coefficients ranged from .92 (Extraversion) to .83 (Openness), with mean of .88. These alphas meet or exceed those reported for the IPIP by Goldberg (1999) and Gow, Whiteman, Pattie, and Deary (2005). All factors were correlated with one another in a theoretically expected pattern, consistent with the literature (e.g., Ones, 1993). For example, neuroticism was negatively correlated with the other four FFM scales (*r*s ranging from -.51 to -.12), whereas the rest of the FFM scales showed modest to moderate correlations with one another (*r*s ranging from .09 to .47).

The observers had interrater reliabilities ranging from .66 (Neuroticism) to .80 (Conscientiousness), with mean of .74. Similar to self-reported personality scores, all factors were correlated with one another in a theoretically meaningful way consistent with what has been found in the literature (e.g., Ones, 1993). For example, neuroticism was negatively correlated with the other four FFM scales (*r*s ranging from -.39 to -.22), whereas the rest of the FFM scales showed modest to moderate correlations with one another (*r*s ranging from .17 to .35).

We present the analyses of the average heterotrait-monomethod (HTMM) correlations for each method, the heterotrait-heteromethod (HTHM) correlations for each pair of methods, and the monotrait-heteromethod (MTHM) correlations for each pair of methods in Table 2. For IBM Watson PI, monotrait correlations with self-reports range from -.20 (Neuroticism) to .18 (Openness), whereas monotrait correlations with observer reports range from -.41 (Neuroticism) to .20 (Agreeableness).

The HTHM correlations were lowest, as expected. However, the HTMM correlations were higher than MTHM correlations, indicating that methods, not traits, represent the major source of variance in the scores. In recent decades, researchers have utilized confirmatory factor analysis to objectively assess MTMM matrix (Kenny & Kashy, 1992). However, the number of estimated parameters in our model compared to our sample size led to model nonconvergence. Therefore, we analyzed the convergence/discrimination of these measurement methods using generalizability theory and ANOVA methods for partitioning the variance (Schmitt & Stults, 1986; Woehr, Putka, & Bowler, 2012). The bottom row of Table 2 presents these statistics. Specifically, these indices reveal that 15% of observed variance is attributable to shared variance specific to either trait or to person main effects (C1: average MTHM correlations). Only 7% of the trait-method units' observed variance is trait-specific variance (D1: average HTHM correlations). Contrasting D1 to C1 suggests over half of the convergence can be attributed to person main effects. Trait variance is 13 percentage points lower than the amount of variance attributable to a given method (D2; average MTHM correlations minus average HTMM correlations), and method accounts for 20% of the total variance (MV: average HTMM correlations minus average HTHM correlations). Overall, little

TABLE 2.
Multitrait Multimethod (MTMM) Analysis

	Self-reports	Observer reports	IBM Watson PI	Average
HTMM	.29	.29	.28	.29
	Self-observer	IBM-observer	Self-IBM	Average
HTHM	.09	.11	.06	.09
MTHM	.17	.15	.13	.15
	Convergence Index (C1)	Discrimination Index 1 (D1)	Discrimination Index 2 (D2)	Method Variance (MV)
Variance Partitioning	.15	.07	-.13	.20

Note. HTMM=heterotrait-monomethod. HTHM=heterotrait-heteromethod. MTHM=monotrait-heteromethod. C1: proportion of variance attributable to person main effects and shared variance specific to traits. D1: proportion of variance in trait-method units attributable to variance specific to traits. D2: difference in proportion of variance accounted for by traits vs. methods. MV: proportion of variance attributable to methods (Woehr et al., 2012).

trait variance is captured. Both analytical methods converge to suggest that convergent and discriminant evidence for construct validity is poor.

To evaluate whether automated solutions may be able to replace a single rater among multiple raters to save organizations money (e.g., Campion et al., 2016), we compared IBM Watson PI's and single observer rating's convergence to self-reports. We calculated single observer correlations with self-reports and averaged the correlations, then compared the average correlations to IBM Watson PI's convergence with self-reports. Compared to the average of single observer correlations, IBM Watson PI showed larger correlations with self-reports for agreeableness (r_{obs} = .13 vs. r_{PI} = .17) and openness (r_{obs} = .10 vs. r_{PI} = .18), similar correlation with conscientiousness (r_{obs} = .05 vs. r_{PI} = .05), and lower correlation with extraversion (r_{obs} = .18 vs. r_{PI} = .06). For neuroticism, the correlation between IBM Watson PI and self-report scores was negative, which was theoretically uninterpretable as mentioned above (r_{obs} = .22 vs. r_{PI} = - .20). This suggests that for agreeableness, openness, and conscientiousness, IBM Watson PI can function as well as a single observer in assessing self-reported personality. A critical caveat is that the magnitude of correlations are low despite performing better than a single observer.

We also assessed how IBM Watson PI's convergence with self-reports compares to personality ratings at zero-acquaintance. Table 3 displays this information, using zero acquaintance correlations from meta-analysis (Connolly, Kavanagh, & Viswesvaran, 2007). Although overall, IBM Watson PI does not outperform zero-acquaintance ratings, its performance was most promising for openness and agreeableness.

Last, we inspected whether demographic differences were observed in the IBM Watson PI and observer ratings of personality. Men received a higher score on IBM Watson PI neuroticism compared to women (t = 2.48, df = 128, p = .01, 95% confidence interval for difference = .02, .20). In contrast, women rated themselves higher on neuroticism than did men (t = -2.09, df = 156, p = .04, 95% confidence interval for difference = -.02, -.82). No other demographic differences were observed.

DISCUSSION

Technological advances afford researchers and practitioners the ability to supplement or even replace human judgment with objective assessments of job applicants. Such approaches hold potential to reduce appearance, gender, and race biases that influence selection decisions. Other researchers claim to have outperformed IBM Watson PI, Schwartz et al. (2013), and Park et al. (2015) in predicting personality from social media posts, but higher accuracy has only been achieved when language features were combined with self-reports of attitudes and behavior as predictors of self-reported personality (Hall & Caton, 2017). To our knowledge, this study is the first to examine the convergent and discriminant validity evidence of language-based personality assessment with self and observer ratings of personality in the context of a video interview. IBM Watson PI showed significant monotrait correlations with self and observer ratings of agreeableness. Additionally, self-reports of openness showed significant monotrait correlations with IBM Watson PI. However, these correlations were very low in magnitude, and no evidence supported convergence with conscientiousness or extraversion.

As noted by a reviewer, the low convergence may be emblematic of a larger concern: that research using language to estimate personality may suffer from a criterion problem (Boyd & Pennebaker, 2017). Specifically, because such approaches utilize self-reports as the gold standard for accuracy, they inherit and compound the known shortcomings of self-reports (i.e., constraints on self-knowledge and response biases). As such, these approaches for estimating personality do not advance our understanding of personality—rather, they can only advance our understanding of how language-use corresponds to people's perceptions of their own personality. Approaches that utilize more valid sources of personality, such as coworkers and family members, may be more useful than models built on self-reports.

The negative correlation between IBM Watson PI's neuroticism score and the self- and observer ratings was a persistent concern in our analyses. IBM Watson PI's neu-

TABLE 3.
Comparison to Zero Acquaintance Convergence With Self-Reports

	E	O	C	A	N
Zero acquaintance	.29	.14	.23	-.01	.05
Observers	.22	.12	.07	.15	.31
IBM Watson PI	.06	.18	.05	.17	-.20

Note. Source of zero acquaintance correlations: Connolly et al., 2007.

roticism score was related in the opposite way we would expect it to be with the monomethod extraversion and conscientiousness scores (*r*s = .54 and .70, respectively), as well as with the monotrait self- and observer ratings (*r*s = - .20 and -.41, respectively). Due to these unexpected results, we repeatedly inspected the system documentation to ensure we were interpreting the trait score correctly. The trait is labeled *emotional range* in their system, and they equate it with neuroticism. The various facet scores are all scored such that a high score indicates maladjustment, either through increased stress, anger, or depression. We searched for papers using IBM Watson PI that reported correlations among the trait scores. This search was unsuccessful for two reasons: The search was temporally restricted because of the recent change in IBM Watson PI from a closed vocabulary approach built on LIWC to an open vocabulary approach, and no recent papers we found utilizing IBM Watson PI reported trait correlations. Additionally, we contacted IBM directly to ask for the trait intercorrelations from validation studies, but they were unwilling to provide them, raising concerns that the trait intercorrelations do not match the accepted structure of the FFM. They stated that trait scores of neuroticism sometimes do not align with the other outputs in expected ways, but they plan to correct this in a forthcoming update. Off-the-shelf approaches require caution because they are often a “black box,” requiring users to assess the level of rigor in product documentation prior to use and following each update.

Limitations and Future Directions

Although the setting used here is more natural to the selection context than social media (Van Iddekinge, Lanivich, Roth, & Junco, 2016), the data examined here are not from an actual selection context. Using data from an actual selection decision would be ideal because it would allow for assessing the validity of the IBM Watson PI personality assessment for hiring decisions, job performance, and turnover. Relatedly, although the current study findings did not provide strong validity evidence based on convergent and discriminant relationships with self- and observer ratings, future research should investigate other types of validity evidence such as predictive relationships with important individual and organizational outcomes.

The average word count is another concern in our videos. The accuracy of IBM Watson PI’s personality scores asymptotes at 3,000 words. The average correlation across all traits is .21 at 600 words, and caps out at .26 at 3,000 words. Future studies of IBM Watson PI may see better convergence with traditional measures of personality by using a longer sample of speech. This suggests that longer interviews may be required to fully utilize this tool.

Finally, one conceptual concern is the trait activation potential of the video prompts. In assessment centers, exercises with higher trait activation potential elicit more accurate ratings of personality (Lievens, Chasteen, Day, & Christiansen, 2006; Speer, Christiansen, & Honts, 2015). Future investigations could benefit from using multiple prompts and assessing whether prompts with greater trait activation potential elicit more accurate personality estimates when using language-based models.

Conclusion

Technological advances hold potential for changing the way we assess and select job applicants. However, to date, little evidence exists to guide researchers and practitioners as to which approaches can accurately assess job applicants. This study took initial steps to fill this gap by analyzing an off-the-shelf language-based personality assessment tool, IBM Watson PI, that has been validated for assessing personality with social media data, in an interview context. The results showed that short video resumes, which are commonly used, apparently provide little personality-relevant information, and in that context, IBM Watson PI demonstrates little convergence with self- and observer ratings. More work is needed to understand whether this tool can be accurate in personnel assessment contexts.

REFERENCES

Allik, J., Realo, A., Mõttus, R., & Kuppens, P. (2010). Generalizability of self-other agreement from one personality trait to another. Personality and Individual Differences, 48(2), 128–132. https://doi.org/10.1016/j.paid.2009.09.008

Auer, E. M. L. (2018). Detecting deceptive impression management behaviors in interviews using natural language processing. Master of Science (MS), thesis, Psychology, Old Dominion University, DOI: 10.25777/yx69-dy97 https://digitalcommons.odu.edu/psychology_etds/70

Balakrishnan, V., Khan, S., Fernandez, T., & Arabnia, H. R. (2019). Cyberbullying detection on twitter using Big Five and Dark Triad features. Personality and Individual Differences, 141(January), 252–257. https://doi.org/10.1016/j.paid.2019.01.024

Blackman, M. C. (2002). The employment interview via the telephone: Are we sacrificing accurate personality judgments for cost efficiency? Journal of Research in Personality, 36(3), 208–223. https://doi.org/10.1006/jrpe.2001.2347

Boyce, A. S., Corbet, C. E., & Adler, S. (2013). Simulations in the selection context: Considerations, challenges, and opportunities. In M. Fetzer & K. Tuzinski (Eds.), Simulations for Personnel Selection (pp. 17–42). New York, NY: Springer.

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. Current Opinion in Behavioral Sciences, 18, 63–68. https://doi.org/10.1016/j.cobeha.2017.07.017

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). "Initial investigation into computer scoring of candidate essays for personnel selection": Correction to Campion et al. (2016). Journal of Applied Psychology, 101(7), 975. https://doi.org/10.1016/S0257-8972(98)00555-6

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin, 136(6), 1092–1122. https://doi.org/10.1037/a0021212

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. International Journal of Selection and Assessment, 15(1), 110–117. https://doi.org/10.1111/j.1468-2389.2007.00371.x

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. Journal of Applied Psychology, 91(1), 40–57. https://doi.org/10.1037/0021-9010.91.1.40

Golbeck, J. (2017). Predicting personality from social media text. AIS Transactions on Replication Research, 2(September), 1–10. https://doi.org/10.17705/1atrr.00009

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. Psychological Assessment, 4(1), 26–42.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. Personality Psychology in Europe, 7(1), 7–28.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Gow, A. J., Whiteman, M. C., Pattie, A., & Deary, I. J. (2005). Goldberg's "IPIP" Big-Five factor markers: Internal consistency and concurrent validation in Scotland. Personality and Individual Differences, 39(2), 317–329. https://doi.org/10.1016/j.paid.2005.01.011

Hall, M., & Caton, S. (2017). Am I who I say I am? Unobtrusive self-representation and personality recognition on Facebook. PloS One, 12(9), e0184417.

HG Insights. (2018). Companies using IBM Watson Personality Insights, market share, customers and competitors. Retrieved from https://discovery.hgdata.com/product/ibm-watson-personality-insights

Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in hiring. Quarterly Journal of Economics, 133(2), 765–800. https://doi.org/10.1093/qje/qjx042

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial–organizational psychology: Reflections, progress, and prospects. Industrial and Organizational Psychology: Perspectives on Science and Practice, 1(3), 272–290. https://doi.org/10.1111/j.1754-9434.2008.00048.x

IBM. (2018). Watson PI Doumentation. Retrieved from https://cloud.ibm.com/docs/services/personality-insights/science.html

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. Psychological Bulletin, 112(1), 165–172. https://doi.org/10.1037/0033-2909.112.1.165

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language. Psychological Methods, 21(4), 507–525. https://doi.org/10.1037/met0000091 T4 - Methodologies and challenges PM - 27505683 M4 - Citavi

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. Journal of Applied Psychology, 98(6), 1060–1072. https://doi.org/10.1037/a0034156

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. Journal of Applied Psychology, 91(2), 247–258. https://doi.org/10.1037/0021-9010.91.2.247

Naim, I., Tanveer, I., Gildea, D., & Hoque, M. E. (2018). Automated analysis and prediction of job interview performance. IEEE Transactions on Affective Computing, 9(2), 191–204. https://doi.org/10.1109/TAFFC.2016.2614299

Oh, I. S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. Journal of Applied Psychology, 96(4), 762–773. https://doi.org/10.1037/a0021832

Ones, D. S. (1993). The construct validity of integrity tests. Unpublished doctoral dissertation, University of Iowa.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Personality at work: Raising awareness and correcting misconceptions personality at work: Raising awareness and correcting misconceptions. Human Performance, 18(4), 389–404. https://doi.org/10.1207/s15327043hup1804

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. E. P. (2015). Automatic personality assessment through social media language. Journal of Personality and Social Psychology, 108(6), 934–952. https://doi.org/10.1037/pspp0000020

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic inquiry and word count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. Annual Review of Organizational Psychology and Organizational Behavior, 54, 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.

Receptiviti. (n.d.). Science. Retrieved from https://www.receptiviti.com/science/

Ryan, A. M., Ployhart, R. E., & Friedel, L. A. (1998). Using personality testing to reduce adverse impact: A cautionary note.

Journal of Applied Psychology, 83(2), 298–307. https://doi.org/10.1037/0021-9010.83.2.298

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10(1), 1–22. https://doi.org/10.1177/014662168601000101

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS ONE, 8(9). https://doi.org/10.1371/journal.pone.0073791

Speer, A. B., Christiansen, N., & Honts, C. (2015). Assessment of personality through behavioral observations in work simulations. Personnel Assessment and Decisions, 1(1). https://doi.org/10.25035/pad.2015.006

Van Iddekinge, C. H., Lanivich, S. E., Roth, P. L., & Junco, E. (2016). Social media for selection? Validity and adverse impact potential of a Facebook-based assessment. Journal of Management, 42(7), 1811–1835. https://doi.org/10.1177/0149206313515524

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of g-theory methods for modeling multitrait-multimethod data: Clarifying links to construct validity and confirmatory factor analysis. Organizational Research Methods, 15(1), 134–161. https://doi.org/10.1177/1094428111408616

Woo, S. E., Keith, M., & Thornton, M. A. (2015). Amazon Mechanical Turk for industrial and organizational psychology: Advantages, challenges, and practical recommendations. Industrial and Organizational Psychology: Perspectives on Science and Practice, 8(2), 171-179. http://dx.doi.org/10.1017/iop.2015.21

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zielinski, D. (2018). Predictive assessments give companies insight into candidates' potential. Society for Human Resource Management. Retrieved from https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/predictive-assessments-insight-candidates-potential.asp

RECEIVED 01/29/19 ACCEPTED 04/18/19