European Journal of Personality, *Eur. J. Pers.* **34**: 826–844 (2020)
Published online 16 July 2020 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2290

Psychometric and Validity Issues in Machine Learning Approaches to Personality Assessment: A Focus on Social Media Text Mining

LOUIS TAY[1]*, SANG EUN WOO[1], LOUIS HICKMAN[1] and RACHEL M. SAEF[2]

[1]*Department of Psychological Sciences, Purdue University, West Lafayette, IN USA*
[2]*Northern Illinois University, DeKalb, IL USA*

Abstract: In the age of big data, substantial research is now moving toward using digital footprints like social media text data to assess personality. Nevertheless, there are concerns and questions regarding the psychometric and validity evidence of such approaches. We seek to address this issue by focusing on social media text data and (i) conducting a review of psychometric validation efforts in social media text mining (SMTM) for personality assessment and discussing additional work that needs to be done; (ii) considering additional validity issues from the standpoint of reference (i.e. 'ground truth') and causality (i.e. how personality determines variations in scores derived from SMTM); and (iii) discussing the unique issues of generalizability when validating SMTM for personality assessment across different social media platforms and populations. In doing so, we explicate the key validity and validation issues that need to be considered as a field to advance SMTM for personality assessment, and, more generally, machine learning personality assessment methods. © 2020 European Association of Personality Psychology

Key words: machine learning; big data; social media; text mining; validity

Psychology has a longstanding interest in analysing lingual behaviours exhibited by individuals as a reflection of their unique personality characteristics (Pennebaker, Mehl, & Niederhoffer, 2003; Sanford, 1942). With fast-emerging technological advances in collecting, processing, and analysing high-volume text data (Srinivasan, 2020), language data extracted from various social media platforms (e.g. Facebook, Twitter, Snapchat, Instagram, YouTube, Pinterest, and LinkedIn) are gaining popularity as an alternative source of personality-relevant information over personality ratings made by human judges (Woo, Tay, & Proctor, 2020). To this end, researchers have developed machine learning algorithms to extract personality-relevant information from social media language (Kern et al., 2016; Schwartz et al., 2013; referred to as social media text data in this paper), along with other types of virtual data such as non-verbal activities (e.g. likes, follows, and profile picture; Kosinski, Stillwell, & Graepel, 2013; Wilson, Gosling, & Graham, 2012). This machine learning approach to personality assessment (MLPA) involves computers analysing the language used in text data to develop algorithms for predicting a person's standing on self-reported personality traits. In this article, we refer to this general method as the *social media text mining* (SMTM) approach to personality assessment (see Kern et al., 2016, for a detailed overview of this method). The promise of technology capable of analysing social media data as a way of enhancing (or even substituting) human judgement in personality assessment has drawn public attention and scholarly discussion, ranging from unbridled enthusiasm to strong scepticism. From the perspective of personality assessment, the *validity* of this new approach as a *measure* of personality (e.g. Bleidorn & Hopwood, 2019; Kern et al., 2016; Park et al., 2015; Schwartz et al., 2013; Woo, Tay, Jebb, Ford, & Kern, 2020) is key, which is the focus of the present article.

A recent review by Bleidorn and Hopwood (2019) has laid out a framework for organizing various issues related to the construct validation of MLPAs using digital footprints such as social media behaviour. Drawing from some of their key arguments, the current article expands their discussion of validity. While much of our discussion about validity can apply to other forms of data used for MLPA (e.g. video, voice, and non-verbal social media behaviours), we focus on SMTM for two reasons. First, at this point, the literature appears rather saturated with 'big picture' discussions of what big data and MLPA approaches can and cannot do *in general*. However, this has left the literature without a focused, in-depth discussion around validity issues associated with specific methods of data collection and analysis within the broader machine learning framework. For these methods to be applied both responsibly and broadly by researchers, they need guidance on how to validate MLPA methods. Second, SMTM has become an increasingly popular method for personality measurement. For example, a seminal paper on automated personality assessment using social media text data

*Correspondence to: Louis Tay, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907, USA.
E-mail: stay@purdue.edu
Louis Tay and Sang Eun Woo contributed equally to the paper.
This article earned Open Data and Open Materials badges through Open Practices Disclosure from the Center for Open Science: https://osf.io/tvyxz/wiki. The data and materials are permanently and openly accessible at https://osf.io/cgpmz/?view_only=4a56e3fb9aa6476bb2b9b27273b4124d. Author's disclosure form may also be found at the Supporting Information in the online version.

Handling editor: John Rauthmann
Received 17 July 2019
Revised 28 June 2020, Accepted 28 June 2020

(Schwartz et al., 2013) has been cited more than 1000 times on Google Scholar.

We seek to make three contributions to the literature. First, we expand the traditional focus on predictive accuracy in validating SMTM. Not only do we meta-analytically examine the predictive validity of SMTM, but we also discuss and review other aspects of psychometric evidence not typically considered, such as reliability, factorial structure, discriminant relationships with measures of purportedly different constructs, *and* predictive relationships with criterion variables. In so doing, we believe that we can provide the field with specific information on where we have more construct validity evidence—and psychometric evidence more broadly—and draw attention to areas that are potentially lacking. Second, we propose that, while conventional test validation frameworks for SMTM (and machine learning approaches, more generally) are useful, other aspects of validity are particularly pertinent in this context. Drawing from the work of Borsboom, Mellenbergh, and van Heerden (2004) on validity, we explore the issues of reference (i.e. ground truth about one's personality) and causality (i.e. how personality determines variations in social media language) as applied to SMTM. Through this, we seek to advance and spur new ideas for enhancing the validity and utility of SMTM personality assessments. Finally, we outline the opportunities and challenges in examining the generalizability of SMTM algorithms. Considering the substantial diversity in the types of users and social media platforms that exist, a wealth of social media text data can be collected and used for validation of SMTM algorithms.

REVIEW OF SOCIAL MEDIA TEXT MINING VALIDATION RESEARCH

We first review past research on the validity evidence for SMTM as a personality assessment method and then identify areas in which empirical evidence is currently lacking. Within the existing literature, there are two primary approaches to assessing personality using social media data: *manual* and *automated* approaches. In the manual approach (i.e. manual social media ratings), human raters review the content of a social media site (or the text extracted from it) and complete a questionnaire asking them about the personality of the target individual. In the automated approach (i.e. SMTM), computers analyse the text from social media sites and relate the information to external sources of personality such as self-reports.

While the focus of our article is on automated (i.e. SMTM) approaches, we also review and discuss findings related to the manual approaches as a point of comparison in thinking about the issues of validation. Both approaches start with the same kind of *behavioural* information (i.e. social media text), but they are also significantly different from a validity standpoint. Validity issues related to the manual approach closely resemble those of zero acquaintance ratings based on behavioural observation. The only difference is that the former utilizes social media behaviour as a data source, whereas the zero acquaintance ratings can be done in either online or offline settings. In other words, zero acquaintance ratings can be considered a broader methodological concept that subsumes the manual approach to assessing personality using social media data. The automated approach, however, does not fall into any existing categories of human-based personality assessments. It starts with behavioural observations (e.g. language use on social media) as a data source but then uses machine learning algorithms to score these observations to infer traits on the basis of human ratings of personality such as self-reports or other-reports (i.e. machine learning-generated personality scores).

Conceptually, Brunswik's (1956) lens model has been proposed as a way to understand how behavioural observations or cues—in this case, social media text data—are used by an observer to judge someone's personality (e.g. Hall, Pennington, & Lueders, 2014; Hinds & Joinson, 2019). This is illustrated in Figure 1. In the manual approach, observers use available cues (i.e. words or phrases in social media text data) to make personality judgements, or ratings on a particular trait. In the automated approach, it is proposed that machine learning algorithms take the role of the observer (although later we discuss the limitations of this perspective), as the algorithms assign weights (sometimes in very complex ways) to the cues (i.e. *features*, in SMTM parlance). The appeal of Brunswik's lens model is in its simplicity and scope in illustrating both manual and automated approaches. In the context of validation efforts, it has also led researchers to focus primarily on predictive accuracy—or the extent to which the algorithms can predict personality judgements (i.e. self-ratings or other-ratings). In other words, researchers are typically interested in the extent to which cues in social media text data *accurately* predict personality trait ratings. This is the primary piece of validity evidence that we examine through meta-analysis and then conceptually expand upon by discussing other aspects of construct validity.

We note here that a more sophisticated way of understanding the accuracy of personality trait judgements comes from the realistic accuracy model (RAM; Funder, 1995). The RAM posits that perceiver accuracy is a function of the environment (i.e. social media platform) and the

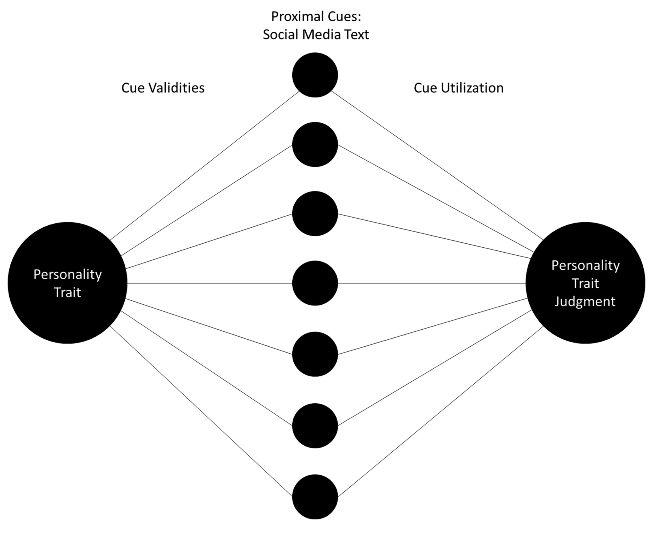

Figure 1. Brunswik's lens model applied to personality trait judgements.

DOI: 10.1002/per

perceiver (i.e. an entity like a person or computer). With regard to the environment, it assumes that personality traits can produce *relevant* behavioural cues (i.e. social media text) that are made *available* to the observer. This corresponds to cue validities in Figure 1. With regard to the perceiver, when relevant behavioural cues are made available, the extent that judgements are accurate will depend on the *detection* of those cues and the subsequent *utilization* of them for personality judgement. This corresponds to cue utilization in Figure 1. One reason that the RAM is more sophisticated is because it is a process approach and explicitly posits a formulaic representation for accuracy as a function of the following: relevance of behavioural cues to trait × availability of behavioural cues when making a judgement × extent cues are detected × the way cues are used (Funder, 1995). This RAM multiplicative formulation has also been empirically demonstrated (Rogers & Biesanz, 2019).

Meta-analysis of convergence with self-reports

Owing to the lack of primary studies available to date, our meta-analysis focused only on one type of validity evidence: convergence of machine learning algorithms with self-reports of personality, which, as mentioned, may also be referred to as *predictive accuracy* in this context. This is sometimes viewed as convergent validity evidence for MLPA in SMTM. However, as we explain later, there are limitations to this perspective. Thus, we use the term convergence with self-reports rather than convergent validity.

Method

We conducted a search for papers on PsycINFO and Google Scholar by pairing the keyword *social media* with *personality*. We sought published papers, and we excluded all papers that did not utilize text (e.g. only used profile pictures) to assess at least one of the Big Five personality traits or that turned out not to be situated in social media (e.g. were situated in essays or used self-reported social media activities). Second, we reviewed the reference sections of the papers our search uncovered as well as two recent reviews of similar topics (Azucar, Marengo, & Settanni, 2018; Tskhay & Rule, 2014) for additional relevant articles. Our search yielded 25 published articles. Table 1 provides descriptive statistics for the studies uncovered in our search.

Table 1. Summary of published studies assessing personality on social media

Assessment type	Sample size			
	k	Median	Range	*SD*
Manual	17	103	6–416	111
Automated				
Closed vocabulary[a]	5	253.5	142–74 941	37 362
Open vocabulary[a]	12	5135	44–74 941	32 075

Note: *k* refers to the number of published studies found that assessed personality from social media.
[a]Sample size includes training samples.

Our inclusion criterion for the meta-analysis was that studies needed to report at least one correlation between self-reported Big Five personality traits and the same trait assessed either by (i) a human rater (i.e. manual) or (ii) a machine learning model (i.e. automated). This enabled us to compare the accuracy of different approaches to personality assessment. Human raters often (but not always) utilized more than only textual information. For the automated approaches, we excluded studies that used content other than text.

Prior to meta-analysis, studies were removed if they were not independent from each other (e.g. multiple researchers conducting similar procedures on MyPersonality dataset). This left 15 publications with k = 21 independent samples for meta-analysis. The third author of this paper coded each study in these articles for (i) sample size, (ii) information source (e.g. Facebook and LinkedIn), (iii) sample (e.g. university students and MyPersonality dataset), (iv) self-report personality scale, (v) type of social media text data personality assessment (i.e. automated vs. manual), (vi) the correlations between traditional self-reported personality ratings and other-reported personality traits, (vii) the cross-validation strategy used if they used SMTM, and (viii) whether discriminant and/or criterion validity information was provided. In addition, a research assistant was trained on the meta-analytic coding procedures and coded 11 of these samples for sample size, information source, sample information, the scale used for self-report personality, type of personality assessment, and the correlations between self-reported and judged traits. The agreement for the first five categories was 100%. Agreement for the correlations between self-reported and judged traits was 88%, and the disagreement was caused by transcription errors by the research assistant. The agreement was 100% after a discussion of coding by the two coders. Table 2 presents the 21 independent samples, coded characteristics, and effect sizes.

We used the psychmeta R package (Dahlke & Wiernik, 2019) to calculate meta-analytic effect sizes, standard deviations, and confidence intervals. The final meta-analytic coding sheet with all coded information for the 21 included samples and the R code used to conduct the meta-analysis are available on OSF: https://osf.io/cgpmz/?view_only=8e959947f234438dbad834bdd905d3ae

Results and discussion

Table 3 presents the results of our meta-analysis examining the convergence of self-reported personality assessments and social media-based personality assessment (b, as well as the results from an earlier meta-analysis of zero acquaintance personality ratings for comparison (Connolly, Kavanagh, & Viswesvaran, 2007). Forest plots of effect sizes for each personality dimension are shown in Figures 2–6. We discuss our findings subsequently, organized by the two major approaches—manual and automated (i.e. SMTM) methods.

When manually rating personality, whether using Facebook pages, LinkedIn profiles, or other social media, interrater reliability consistently reaches acceptable levels with as few as three raters for all traits except neuroticism (Kluemper & Rosen, 2009; Roulin & Levashina, 2019). Therefore, raters appear capable of general agreement about

Table 2. Studies and samples included in meta-analysis

Publication/study	*N*	Source	Sample	Scale	Type	E	A	C	ES	O
Back et al. (2010)	236	Facebook and StudiVZ	Univ. students	TIPI; BFI-10/NEO hybrid	Manual	.39	.22	.27	.13	.41
Garcia and Sikström (2014)	304	Facebook	MTurks	EPQ-R-S	Open voc.	.10			.10	
Gill et al. (2006)	18	Emails	British adults	EPQ-R-S	Manual	.89			−.38	
Golbeck (2016), Dataset 1	127	Facebook	MyPersonality	IPIP variants	Open voc.	.37	.41	.25	.38	.36
Golbeck (2016), Dataset 2	8569	Facebook	MyPersonality	IPIP variants	Open voc.	.22	.24	.20	.18	.20
Golbeck (2016), Dataset 3	69	Facebook	Facebook users	BFI-44	Open voc.	.24	−.35	−.07	−.18	−.35
Hall et al. (2014)	100	Facebook	Univ. students	BFI-44	Manual	.23	.32	.20	.16	.15
Kluemper et al. (2012), Study 1	274	Facebook	Univ. students	IPIP-50	Manual	.44	.40	.30	.23	.42
Kluemper et al. (2012), Study 2	244	Facebook	Univ. students	IPIP-50	Manual	.28	.26	.19	.21	.16
Marcus et al. (2006), Study 1	274	Personal site	Site makers	BFI-44	Manual	.23	.01	.18	.20	.36
Marcus et al. (2006), Study 2	72	Personal site	Site makers	BFI-44	Manual	.21	.37	.32	.48	.53
Park et al. (2015)	4824	Facebook	MyPersonality	IPIP variants	Open voc.	.42	.35	.37	.35	.43
Qiu, Lin, Ramsay, and Yang (2012)	142	Twitter	Mixed	BFI-44	Manual	.05	.32	.05	.23	.09
Roulin and Levashina (2019), Study 1	133	LinkedIn	Univ. students	IPIP-20	Manual	.20	.10	.08	−.02	.06
Roulin and Levashina (2019), Study 2	118	LinkedIn	Univ. students	IPIP-20	Manual	.31	.09	.14	.07	.04
Schwartz et al. (2013)	18 735	Facebook	MyPersonality	IPIP variants	Closed voc.	.27	.25	.29	.21	.29
Stopfer, Egloff, Nestler, and Back (2014)	103	SchuelerVZ ans StudiVZ	Univ. students	BFI-10	Manual	.31	.21	.31	.10	.44
van de Ven, Bogaert, Serlie, Brandt, and Denissen (2017), Study 1	178	LinkedIn	Univ. students and workers	TIPI	Manual	.29	.24	.11	.03	.30
van de Ven et al. (2017), Study 2	97	LinkedIn	Workers	G5-R	Manual	.29	−.07	.00	−.09	.07
Van Iddekinge et al. (2016)	416	Facebook	Univ. students	IPIP-50	Manual			.08		
Vazire and Gosling (2004)	89	Personal site	Yahoo users	BFI-44	Manual	.26	.31	.35	.21	.42

Note: Univ. students, University students; MTurks, Amazon Mechanical Turks; Personal site, personal website; Site makers, personal website makers; TIPI, Ten Item Personality Inventory; BFI, Big Five Inventory; EPQ, Eysenck Personality Questionnaire; IPIP, International Personality Item Pool; Open voc., open-vocabulary text mining; Closed voc., closed-vocabulary text mining; E, extraversion; A, agreeableness; C, conscientiousness; ES, emotional stability; O, openness.

Table 3. Convergence of social media assessments with self-reports of personality

		Personality trait	k	N	r	SD_r	95% CI
Assessment approach							
Manual		Extraversion	14	2078	.29	0.12	[0.22, 0.36]
		Agreeableness	13	2060	.22	0.14	[0.13, 0.30]
		Conscientiousness	14	2476	.18	0.10	[0.12, 0.24]
		Emotional stability	14	2078	.15	0.13	[0.08, 0.22]
		Openness to experience	13	2060	.28	0.16	[0.18, 0.37]
Automated							
	Closed vocabulary[a]	Extraversion	1	18 177	.27		
		Agreeableness	1	18 193	.25		
		Conscientiousness	1	18 195	.29		
		Emotional stability	1	18 177	.21		
		Openness to experience	1	18 202	.29		
	Open vocabulary[b]	Extraversion	5	13 893	.29	0.11	[0.15, 0.43]
		Agreeableness	4	13 589	.28	0.08	[0.15, 0.41]
		Conscientiousness	4	13 589	.26	0.10	[0.10, 0.42]
		Emotional stability	5	13 893	.24	0.09	[0.11, 0.36]
		Openness to experience	4	13 589	.28	0.14	[0.06, 0.50]
Connolly et al. (2007)							
Zero acquaintance[c]		Extraversion	8	724	.29	0.14	[0.18, 0.40]
		Agreeableness	8	724	−.01	0.16	[−0.13, 0.11]
		Conscientiousness	8	724	.23	0.22	[0.06, 0.40]
		Emotional stability	8	715	.05	0.22	[−0.12, 0.22]
		Openness to experience	7	609	.14	0.24	[−0.06, 0.34]

Note: k, the number of studies included in the meta-analytic effect size estimate; N, the total validation sample size; r, the uncorrected, sample size weighted mean observed correlations between self-reports and raters' mean ratings for manual, stranger rating for zero acquaintance, and automatic assessment for closed-vocabulary and open-vocabulary approaches, respectively; SD_r, the standard deviation of the effect size estimate; 95% CI, 95% confidence interval around r.

[a]Closed-vocabulary approach was not included in the meta-analysis because only one study was found in that category (Schwartz et al., 2013); we still include the correlation values from Schwartz et al. (2013) for easy comparison.

[b]Park et al. (2015) was included but Schwartz et al. (2013) was excluded owing to using the same dataset and similar methods but Park et al. (2015) reported higher convergence.

[c]From Connolly et al. (2007), table 4.

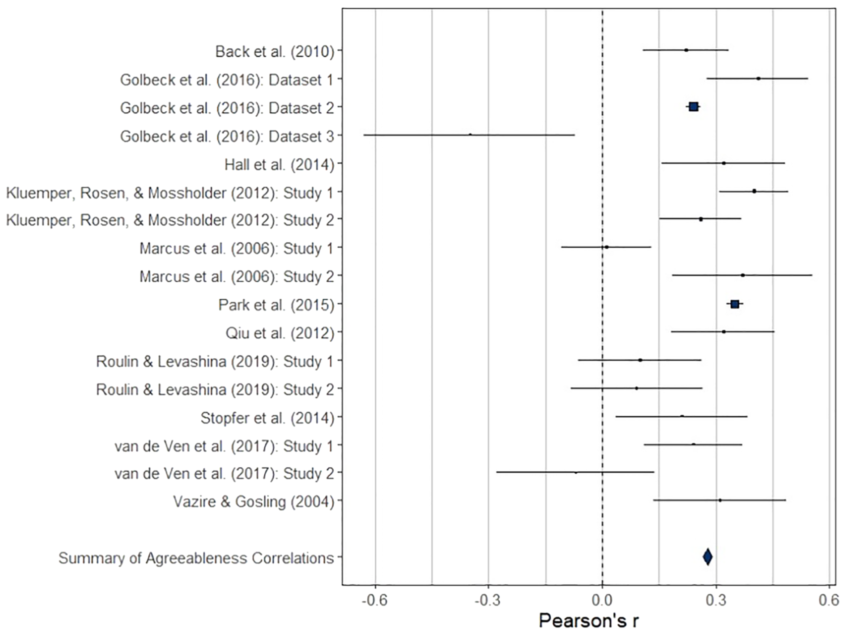

Figure 2. Forest plot of agreeableness effect sizes. [Colour figure can be viewed at wileyonlinelibrary.com]

people's personalities when viewing their social media profiles, and some evidence suggests such assessments can be valid as well. As shown in Table 3, observer personality ratings show some convergence with self-reported personality. However, this convergence varies substantially, with some studies finding extremely strong evidence of convergence (e.g. extraversion r = .89 in lab-produced emails; N = 18; Gill, Oberlander, & Austin, 2006), and other studies finding little to no evidence of convergence (e.g. neuroticism r = −.02; openness r = .06; conscientiousness r = .08; Study

DOI: 10.1002/per

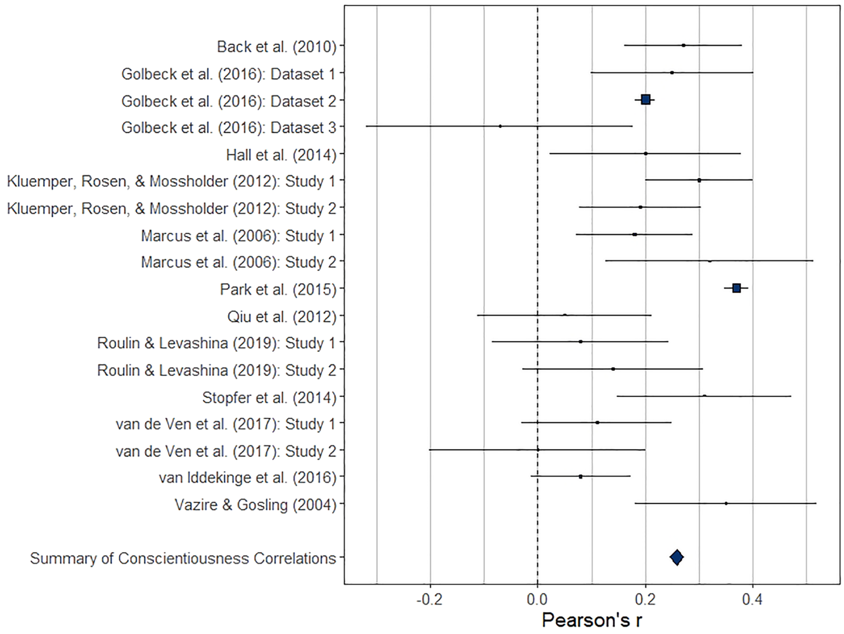

Figure 3. Forest plot of conscientiousness effect sizes. [Colour figure can be viewed at wileyonlinelibrary.com]

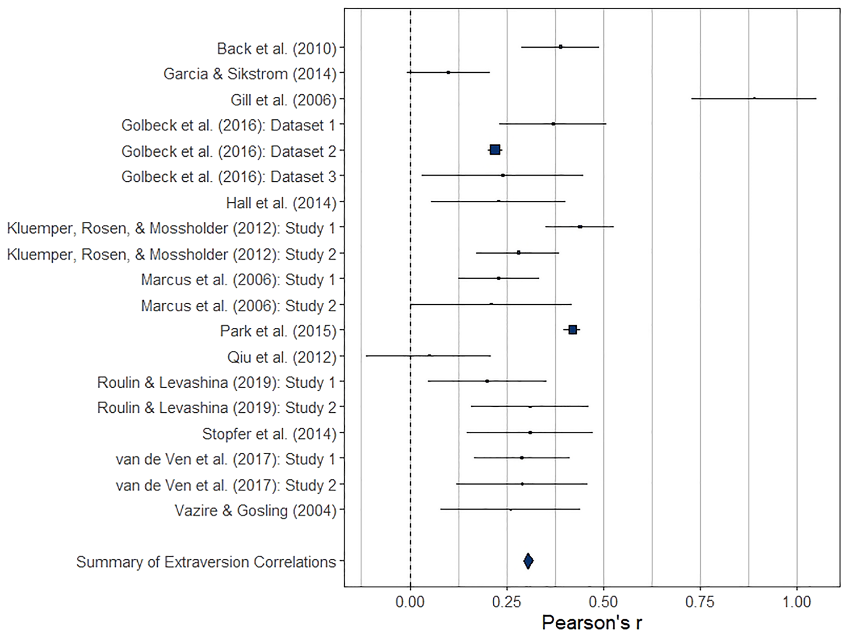

Figure 4. Forest plot of extraversion effect sizes. [Colour figure can be viewed at wileyonlinelibrary.com]

1 LinkedIn profiles in Roulin & Levashina, 2019). Overall, the sample size weighted mean convergence across all five traits was .22. Although researchers have utilized frame-of-reference training in hopes of improving personality judgement accuracy (Kluemper & Rosen, 2009; Kluemper, Rosen, & Mossholder, 2012), to our knowledge, no research has examined whether such training improves reliability or validity of social media-based personality judgements. Overall, the evidence suggests that manual personality ratings using social media data capture some variance in self-reported personality.

In the automated approaches, text data from a social media page (or other types of virtual data; cf. Kosinski et al., 2013) are analysed ('mined') by a computer, and then a machine learning algorithm uses the features extracted from the text as predictors of other established (usually self-reported) measures of personality. Then, the resulting model (e.g. regression weights assigned to extracted features) for predicting self-reported personality is tested on a separate validation sample. The automated approaches typically use either closed-vocabulary or open-vocabulary text mining (Kern et al., 2016). In closed-vocabulary text mining, text data are processed through a pre-determined dictionary that specifies meaningful categories of words (i.e. top-down), whereas open-vocabulary approaches do not specify *a priori*

DOI: 10.1002/per

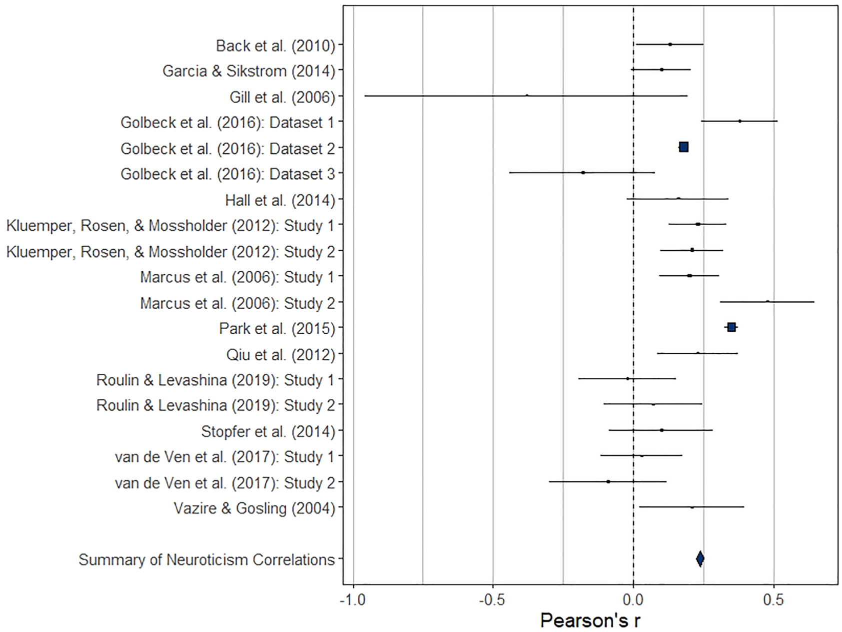

Figure 5. Forest plot of neuroticism effect sizes. [Colour figure can be viewed at wileyonlinelibrary.com]

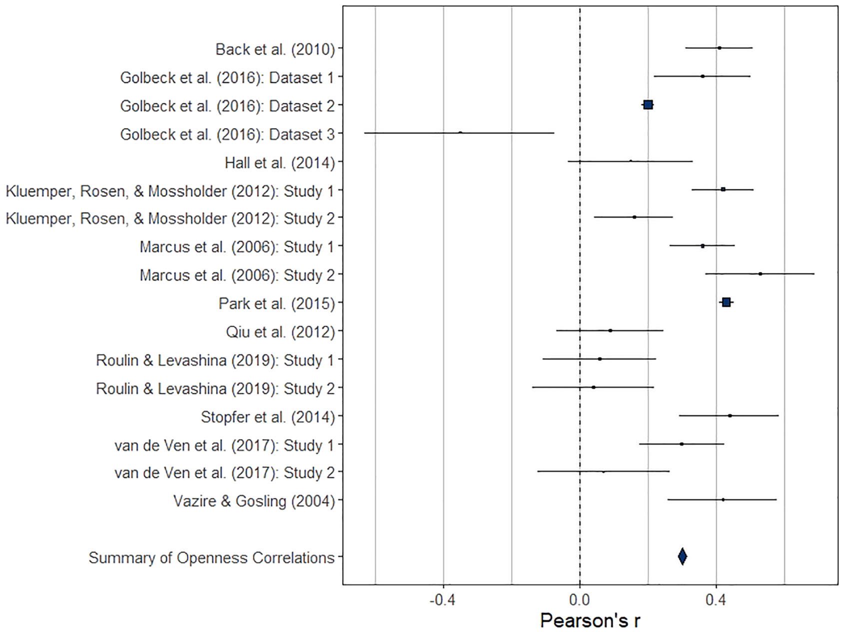

Figure 6. Forest plot of openness effect sizes. [Colour figure can be viewed at wileyonlinelibrary.com]

relationships between features and instead allow meaningful words and phrases to emerge from data (i.e. bottom-up; see Srinivasan, 2020, for a detailed overview of these text mining methods). Kern et al. (2016) have provided a detailed description of these methods. One of the most widely used closed-vocabulary approaches introduced to date is the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2003). Using the LIWC approach, research has powerfully demonstrated that the language we tend to use on social media is correlated with how people view themselves and how they are perceived by others (Qiu et al., 2012). However, only one study cross-validated[1] closed-vocabulary approaches with *a priori* weights: Schwartz et al. (2013) used 25% of their 136 000 Facebook profiles as a holdout/test sample for both open-vocabulary and closed-vocabulary approaches. In their study,

[1]We planned to examine the moderating effect of cross-validation strategy on convergent validity, yet the only automatic study we uncovered that did not use a separate test set was weak methodologically in terms of using computers to capture behaviour, and had lowest convergent validity of automatic studies (Garcia & Sikström, 2014).

cross-validated LIWC categories showed significant correlations with self-reported personality traits (Table 3) but still had lower correlations with self-reported traits than open-vocabulary approaches did and did not incrementally predict self-reports beyond open-vocabulary approaches.

For open-vocabulary approaches, the sample weighted mean convergence across all five traits was .27. The meta-analytic results for open-vocabulary approaches in Table 3 are composed primarily of Golbeck's (2016) three studies and the Park et al. (2015) validation of open-vocabulary approaches. Golbeck cross-validated *Receptiviti*, a commercially available product that combines closed-vocabulary and open-vocabulary approaches to assess personality from language use. Park et al., on the other hand, developed a text mining approach for assessing personality on the wall posts of nearly 67 000 Facebook profiles and then cross-validated it on 4824 profiles. The Park et al. study, as far as our search uncovered, represents the gold standard for evaluating convergence between traditional self-report personality assessment and SMTM personality assessment, with correlations, as follows: openness r = .43, conscientiousness r = .37, extraversion r = .42, agreeableness r = .35, and neuroticism r = .35. These correlations outperform a well-known language-based personality assessment tool, IBM Watson Personality Insights, which reports an average convergence with self-reports across the Big Five r = .33 (IBM, 2018).

Psychometric and validity evidence beyond convergence with self-reports: reliability, factorial structure, discriminant relationships, and criterion predictions

In addition to the meta-analytic findings on the convergence with self-reported personality ratings, we review and discuss the literature on other types of information as related to psychometric and validity evidence of SMTM personality assessment (AERA, APA, and NCME, 2014)—namely, reliability, factorial structure, discriminant relationships with measures of purportedly different constructs, and predictive relationships with criterion variables.

Reliability

The reliability of a measure is a necessary condition for measurement validity (Lord & Novick, 1968). When using other-reported personality ratings (from multiple 'others'), reliability is most often indexed via interrater reliability—if judges agree in their assessments, it suggests that construct relevant variance is being captured (Funder, 1995). ICC(2, k) is the most frequently reported index of interrater reliability, and it tends to exceed .50 for all traits except neuroticism. Interrater reliability is almost always reported in studies of manual personality ratings, although one paper reported Cronbach's alpha to index the reliability of observer ratings (e.g. Marcus, Machilek, & Schütz, 2006), which does not inform us as to whether observers agreed in their judgements. It is important to note here that individual observers may be reliable (e.g. over time or across multiple ratings), but their ratings may not always converge with other raters. Indeed, unique observer effects were shown to account for equal or larger variances compared with common observer effects across the different Big Five traits (McAbee & Connelly, 2016). It is beyond the scope of this article to discuss whether interrater reliability is an appropriate metric for measurement reliability, but suffice to say, this will be a helpful discussion to have in future work assessing the psychometric properties of SMTM.

By contrast, with SMTM, the reliability of MLPA is less often addressed. Park et al. (2015) is the only study we uncovered that addressed measurement reliability. Specifically, they examined test–retest reliability by splitting the available Facebook posts of each user in 6-month intervals, retaining users with at least 1000 words in each interval. Test–retest reliability averaged r = .70 across all five traits in consecutive 6-month intervals, comparable with test–retest reliabilities of self-reported traits. We believe that test–retest reliability is a particularly important aspect of validity-related evidence, especially for personality assessments because personality traits are relatively enduring patterns of thinking, feeling, and behaving (McCrae, Kurtz, Yamagata, & Terracciano, 2011). We encourage researchers to assess and report test–retest reliability in future studies.

Another type of reliability that is particularly relevant in the current context is consistency across different scoring algorithms created using the same text data, which is largely lacking in the SMTM literature. From the perspective of Brunswik's lens model where computer algorithms are construed as raters, we can imagine a scenario in which researchers are interested in the inter-'rater' reliability of different computer algorithms. Indeed, Sajjadiani, Sojourner, Kammeyer-Mueller, and Mykerezi (2019) recently coined the term 'inter-algorithm' reliability to reflect correspondence between different algorithms used. This was applied to the context of machine learning for job application form data and not personality assessments. By assessing the level of convergence when applying different algorithmic approaches in scoring personality, it enables researchers to also understand to what extent derived scores are algorithm specific or algorithm agnostic. We believe that this concept should be considered and potentially incorporated when assessing the reliability of SMTM approaches for personality assessment.

Factorial structure

In the context of social media data, the concept of factorial validity evidence would imply that the behavioural cues (i.e. social media text) are aligned with the posited underlying structure of the personality construct. As shown in Figure 1, the cue validities in Brunswik's lens model reflect the relation between social media text data and the underlying personality construct. The cues (i.e. specific language in social media text) are also often posited as correlated in the lens model, and we can imagine that they could form a unidimensional factor structure in Figure 1. In other words, social media text language (e.g. 'love' and 'like'; 'great' and 'awesome') are correlated and may be indicators of the same underlying factor. This parallels factor analytic approaches (e.g. exploratory factor analysis and confirmatory factor analysis) to assess factorial validity in personality measures (Hopwood & Donnellan, 2010). Apart from a

DOI: 10.1002/per

unidimensional structure, as shown in Figure 1, we can also imagine cases where there are multiple underlying personality constructs with corresponding behavioural cues, leading to a multidimensional structure. There may also be more complex structures when the cues (i.e. specific language in social media text) may index multiple personality constructs (e.g. the phrase ‘I love you!’ may index the aspect of enthusiasm in extraversion and the aspect of compassion in agreeableness).

However, factorial validity evidence is often not examined in the context of SMTM. The closest *analog* to factorial validity evidence is word clouds, where words and phrases that most strongly correlate to (or predict) the personality outcome are visualized as being larger and more frequent words are darker in colour (Schwartz et al., 2013). In a sense, these are visualizations of how the different cues (i.e. specific language in social media texts) load onto the personality construct. Yet there are substantial caveats. These visualizations actually reflect not how the different cues load onto the personality *construct per se* but onto the personality outcome of interest. That is, we are not modelling the cue validities (on the left side of Figure 1) but the cue utilization (on the right side of Figure 1). Another caveat brought up by a reviewer is that word clouds are based on the frequency of word usage rather than their co-occurrence as typically understood in a factor analysis sense and may belong in the realm of *face validity*. Instead, if we are interested in factorial validity evidence, factor analytic approaches should be applied to determine how these different cues (i.e. specific language in social media texts) hang together. In other words, this may require a different methodological approach (e.g. clustering; and factor analysis on text data) than seeking to maximize convergence on self-reports or other-reports of personality.

While we encourage the use of factor analysis, we recognize that there are also challenges. MLPAs (e.g. SMTM) often adopt an empirical keying approach, such that the machine learning algorithms seek to maximize prediction accuracy and so select social media texts (i.e. features) that are going to have high predictive value. This is unlike conventional approaches where we seek to retain personality survey items that most converge with each other (i.e. form a single factor structure) for a single personality dimension. Therefore, there may be some discrepancy between the factor analysis of social media texts and personality survey items. Peripheral evidence that conducted factor analyses on Google search words suggests that the factor structures may not directly map onto factor structures from survey instruments (Ford, Jebb, Tay, & Diener, 2018). In addition, researchers likely will encounter issues of low base rates of social media texts affecting the factorial structure as well. Some corrections to the correlations (Pennebaker, Boyd, Jordan, & Blackburn, 2015) from which the factorial structure is derived can be applied. Or less sophisticated forms of factor analytic approaches such as principal components analysis to determine the presence of a general factor may be unobjectionable for now. Conceptually, we posit that factor structures of text data may not veridically ‘measure’ a specific personality construct given that text data likely emerge out of multiple situational and personality factors; therefore, ‘pure’ loadings of text data onto a specific underlying personality construct may not occur with any reasonable frequency—or at all. Clearly, more validation research and research on new methods for understanding factorial validity are needed in this area.

Discriminant relationships

Discriminant validity evidence refers to the extent to which a measure of a given construct (e.g. extraversion) is not related to measures of other, distinct constructs (e.g. openness) to which it should not be related. Although self-reports tend to exhibit good discriminant evidence, the research uncovered by our search suggests that both manual and automated judgements of personality from social media tend to have relatively poor discriminant validity evidence. For instance, Van Iddekinge, Lanivich, Roth, and Junco (2016) had raters judge conscientiousness from Facebook profiles. The heterotrait–monomethod correlations for conscientiousness judgements ranged from $r = .57$ [contextual aspects of work ratings (i.e. interpersonal skills, adaptability, and creativity) *and* conscientiousness] to $r = .92$ (having levels of KSAOs suitable for the job *and* conscientiousness), suggesting a lack of discrimination among constructs by raters in this scenario. To our knowledge, Park et al. (2015) is the only MLPA study we uncovered that addressed discriminant validity evidence for Big Five traits when applying SMTM. Big Five traits had heterotrait–heteromethod correlations (between self-reports and language-based scores, $|r| = .00$ to $.17$) that were low, which were also relatively small compared with the monotrait–heteromethod correlations (i.e. convergent r of traits across self-reports and language-based scores = .35 to .42). This implies that there is some level of discriminant validity evidence for Big Five traits.

With SMTM approaches, one of the key challenges to assessing discriminant relationships is that the computer algorithms are seeking to maximize convergence (with self-reports or other-reports of the trait of interest; i.e. empirical keying). This is different from the creation of conventional self-report and other-report personality scales where scale items are selected on the basis of whether they sufficiently load onto the intended construct and do not load onto non-relevant constructs. Analogous to scale creation, we envision the development of new methods or algorithms that not only account for maximizing convergence between MLPAs for predicting the same trait but also minimize convergence with MLPAs for predicting non-relevant traits. For example, we need to consider multi-objective optimization approaches (Jin, 2006). Further, we need to develop ways to apply multi-trait (multi-method) examinations of SMTM personality scores (i.e. machine learning-generated personality scores). Similar to factorial validity evidence, we expect that more validation research will be needed.

Criterion predictions

If a personality measure is designed for the purpose of decision making, such as selection or clinical diagnosis, another important type of validity evidence is required: criterion-related validity (i.e. the predictive relationships with external variables of interest). For example, in order to use

DOI: 10.1002/per

MLPA scores on the basis of social media text data to index individuals' suitability for employment, the MLPA scores should predict job performance. Using social media text data manual approaches, some researchers have found that Facebook-based agreeableness, conscientiousness, and neuroticism judgements exhibit meaningful relationships with academic outcomes and work performance (Ivcevic & Ambady, 2012; Kluemper et al., 2012). However, contrary to the robust positive relationship with performance found when using traditional personality measures (e.g. Barrick & Mount, 1991), conscientiousness had a *negative* relationship with future workplace performance when corporate recruiters rated applicant conscientiousness on the basis of their Facebook profiles (van Iddekinge et al., 2016). Park et al. (2015) was, again, the only SMTM study to provide correlations with external variables—including what might be considered important life outcomes (e.g. life satisfaction and number of sick days taken). On the whole, SMTM tended to be less strongly related to the external criteria than self-reported traits, yet the SMTM assessments did exhibit meaningful correlations with many outcomes that mirrored the pattern of self-reported trait relationships (e.g. positive relationship between extraversion and job satisfaction).

We suggest that future validation research seeks to examine the relation of SMTM and other MLPAs with known criteria. This will be an important litmus test to determine whether personality scores derived from SMTM (i.e. machine learning-generated personality scores) predict outcomes, whether they predict outcomes beyond traditional self-report and/or other-report measures of personality, and whether these predictive relations hold over time as we would expect of individual difference variables.

Summary

So far we have reviewed and discussed findings from SMTM research that takes 'conventional' approaches to validating a measure—that is, studies following the recommendations given in the literature on test validity and construct/scale validation pertaining to reliability and relationships with other variables (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). While initial results have shown promise (Park et al., 2015), as presented and discussed above, we believe that more validation work is needed, and newer methodological approaches will be required to examine different types of psychometric and validity evidence fully. Empirical data gathered through such investigations are undoubtedly valuable, and we strongly encourage more research efforts to be devoted to the accumulation of multiple types of validity evidence, as discussed above.

REFLECTIONS ON ADDITIONAL VALIDITY ISSUES: REFERENCE AND CAUSALITY

In this section, we raise other important questions related to validity that have not received much scholarly attention thus far. These questions come from appreciating the fact that personality scores generated by SMTM are profoundly different from those made by human judges from the perspective of scoring and response processes, as illustrated using Brunswik's lens model schematic. Such differences highlight the unique validity considerations for MLPAs, compelling the field of personality measurement to move beyond what is traditionally considered as best-practice recommendations for validation. Subsequently, we reflect on two major ways in which validating MLPA approaches, such as SMTM, may diverge from traditional validation practices for personality assessment on the basis of human raters.

First and foremost, SMTM is not an independent source of personality judgement. Machine learning algorithms used for personality assessment are best understood as scoring techniques that represent artificial intelligence, which is designed to 'mimic' human intelligence. To do so, machine learning algorithms are trained on inputs from human judges (i.e. traditional self-report or other-report personality ratings), accepting the human judgements as ground truth. We note that these automated approaches are different from the manual approaches where humans rate the social media text themselves. The SMTM approach seeks to obtain from the same individual their social media text data and their personality (self-reports or other-reports) and to use MLPA to derive a score of their personality (mimicking self-reports or other-reports) from the social media text data. In other words, the purpose of SMTM is to automate the social–cognitive processes in which humans make (imperfect) judgements about someone's personality on the basis of the (limited) data made available to them. While machine learning algorithms can utilize a greater range of behavioural data than humans when judging personality and do so with greater consistency across all target individuals, because they utilize human judgements as ground truth, they are fundamentally subject to human intelligence and all cognitive fallacies, perceptual inadequacies, and response biases that are associated with it. This brings us back to a crucial validity concern about 'what is true' about personality (i.e. *reference*; Borsboom et al., 2004) and how to evaluate the validity of personality measures without relying on other (faulty) measures.

Second, to establish *causality* in measurement validity (i.e. how personality determines variations in scores derived from SMTM; Borsboom et al., 2004), there needs to be an additional step of validation that maps out psychological processes in which personality leads to the specific social media language behaviours. Whether exhibited online or offline, human behaviours result from a combination of multiple person-level psychological constructs (e.g. interests, goals, and attitudes) in addition to personality, as well as various situational factors, independent of personality. Personality is correlated with a variety of behaviours such as health-risk behaviours (Caspi et al., 1997) and alcohol abuse (Cloninger, Sigvardsson, & Bohman, 1988); these behaviours are considered as outcomes that are predicted by personality, not as direct consequences of personality. Validation efforts involving calculating correlations with other variables serve epistemological purposes, but validity

DOI: 10.1002/per

claims are made as strong as the degree to which the accompanying theoretical and empirical evidence can point to causality beyond mere predictions.

In sum, we believe that the perspectives by Borsboom et al. (2004) on validity are important to consider for SMTM approaches of personality, beyond conventional validation approaches (Bleidorn & Hopwood, 2019). In the following, we explore and discuss the issues of reference (i.e. ground truth) and causality (i.e. how personality determines variations in social media language) as applied to SMTM.

Reference: what is the ground truth?

We call for more explicit and thoughtful discussions around what is considered the 'ground truth' about one's personality (cf. Funder, 2012) throughout the scientific process of designing, conducting, and interpreting research regarding SMTM's validity (or any other MLPA studies for that matter). To facilitate such discussions, Table 4 delineates different sources of personality-relevant information that capture different aspects of one's personality trait. Self-reports and other-reports of personality are often considered to reflect two important aspects of personality: identity and reputation, respectively (Solomon & Vazire, 2016). The question is where would SMTM fall within such a framework? One may argue that it is more aligned with reputational conceptualizations of personality, as it utilizes external (and socially coded) online language behaviours as an input, rather than introspective accounts of the persons themselves. Following this logic, she or he might also conclude that social media behaviours are inadequate in capturing a self-constructed version of one's personality (Hogan, 1983; McAdams, 1996). However, we suggest that it would be a faulty (or at best premature) conclusion. In automated personality assessment approaches like SMTM, the target person's language behaviour data are coded and scored by an algorithm that is specifically trained to maximize their predictive relationships with self-ratings and/or other-ratings of personality—whichever is used as ground truth during the training of scoring algorithms.

Consistent with the perspective of trait realism (Funder, 2012; Tellegen, 1991)—and scientific realism more broadly (Borsboom et al., 2004; Haig, 2020), our premise is that there is an objective truth ('reality') about one's personality independent of observations and that all empirical approaches to capturing such a reality (i.e. measurements) would fall short in one way or another. This is also aligned with the fundamental assumption of Brunswik's lens model, shown in Figure 1, where there is an underlying personality trait construct that we seek to assess. Given that self-reports and other-reports form the basis of the machine learning algorithms, we need to recognize that we are assessing a proxy (i.e. self-reports and other-reports) of a personality trait. Therefore, we inherit both the strengths and weaknesses of the proxy measure. In other words, it is important to recognize that self-reported and other-reported personality scores are both inadequate and fallible in capturing the entire truth about personality and machine algorithms built to predict such scores inherit their inadequacies and, therefore, have additional challenges in validity.

To illustrate these issues, we extend Brunswik's lens model to show the potential gaps in validity for human

Table 4. An overview of validity considerations for four major personality measurement approaches

Sources and measurement approach	Advantages	Limitations and challenges
Self-report	Best in capturing 'identity'	
• Direct	Captures internal, less visible traits very well (e.g. affective tendencies, motives, preferences, and inclinations that are less behavioural)	Highly subjective Less effective in capturing personality as perceived by others (reputation)
• Indirect (e.g. TAT and conditional reasoning)	Less subjective than direct self-ratings	Developing reliable and valid scoring keys
Acquaintances observers (e.g. family and friends)	Best in capturing 'reputation'	
• Direct	Less subjective than direct self-reports	Contaminated by social relationship dynamics and rater idiosyncrasies (e.g. leniency/severity biases and projection)
• Behavioural (e.g. parent ratings of children personality-relevant behaviours)	Less subjective than direct acquaintance ratings	Only capable of capturing personality states—or behavioural expressions; cannot capture the explanatory side of personality (i.e. cognitive–affective mechanisms)
Zero acquaintance observers	Free from biases driven by social relationships	Contaminated by rater idiosyncrasies
• Offline behaviours (e.g. lab)	Less subjective than self-ratings or acquaintance ratings	Contextualized; may not be generalizable to other offline (as well as online) settings
• Online behaviours (e.g. social media)	Naturalistic	Contextualized; may not be generalizable to offline settings (and across online platforms)
Machine learning algorithms (e.g. SMTM)	Largely disinterested (free from human rater biases other than what is included in the model training)	Reflects the biases already present in the training data (i.e. human-sourced data on personality used as 'the ground truth')
	Does not suffer from cognitive limitations (e.g. heuristics, carelessness; order and contrast effects)	

Note: SMTM, social media text mining.

DOI: 10.1002/per

judgements and SMTM approaches that are built on these human judgements. A typical personality judgement rating is valid (i.e. provide good information) to the extent that (i) the personality trait produces valid cues as observed in behaviours and (ii) that these valid cues are appropriately utilized in the judgement. Therefore, the extent that human raters (i.e. self-ratings and other-ratings) can produce valid ratings will depend on whether observed behaviours are valid for the personality trait and whether human raters use the appropriate social media text.

As shown in Figure 7, we present valid judgements for Scenarios 1 through 4 (corresponding to each cue number). In Scenario 1, observed behaviours are not related to the personality trait and are not used in the judgement (i.e. true negative on available behaviours). In Scenario 2, observed behaviours are not only observed but also unrelated to the personality trait (i.e. true negative on unavailable behaviours). In Scenarios 3 and 4, the behaviours are related to the personality trait, and they are also used as cues in the personality trait judgement (i.e. true positive on observed behaviours). There may be multiple cases, however, when invalid judgements can be made by human raters, as shown in Scenarios 5 through 7 (corresponding to each cue number). In Scenario 5, observed behaviours are related to the personality trait but not used in the personality trait judgement (i.e. false negative on observed behaviours). In Scenario 6, observed behaviours are unrelated to the personality trait but are incorrectly used in the personality trait judgement (i.e. false positive on observed behaviours). In Scenario 7, observed behaviours are related to the personality trait but are not available for judgement (i.e. false negative on unavailable behaviours). This illustrates that the personality trait judgements are imperfect proxies of personality traits and represents the first layer of validity issues to consider.

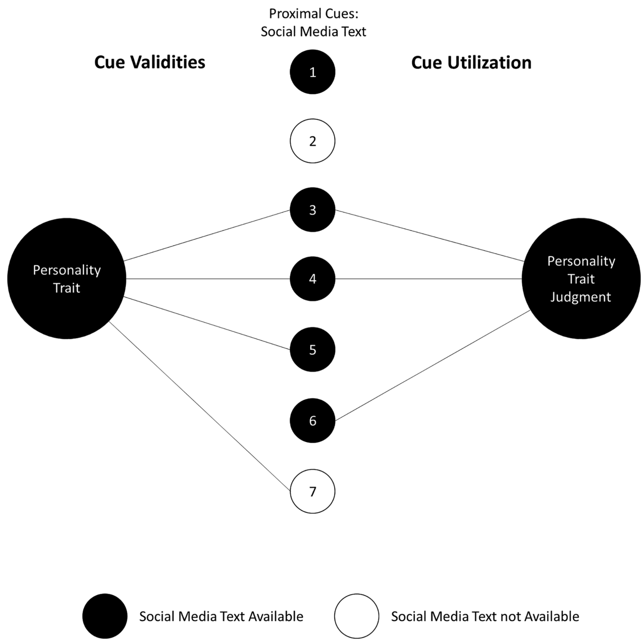

Figure 7. Brunswik's lens model of personality trait judgements on social media texts: illustrating seven potential scenarios of validity and invalidity.

The second layer of validity issues is encountered when one uses SMTM approaches to predict personality traits from these fallible human judgements. As shown from Figure 7, the potential scenarios of validity and invalidity can lead to the self-reported or other-reported personality to contain trait-relevant and trait-irrelevant variance in the scores. To the extent that the self-reports or other-reports of personality contain trait-relevant variance of the targeted personality trait will determine the extent that we have more valid inferences of the targeted personality trait in SMTM approaches.

Figure 8A shows the case for when variance in ratings is trait *relevant*. In this case, because the variance in the ratings is trait relevant, machine learning algorithms will utilize all the personality trait-relevant cues. However, the utilization of these cues will still be weaker to the extent that the trait-relevant ratings fail to cover the entire construct space. Further, they will be weaker to the extent that text mining procedures implemented capture only partial information.

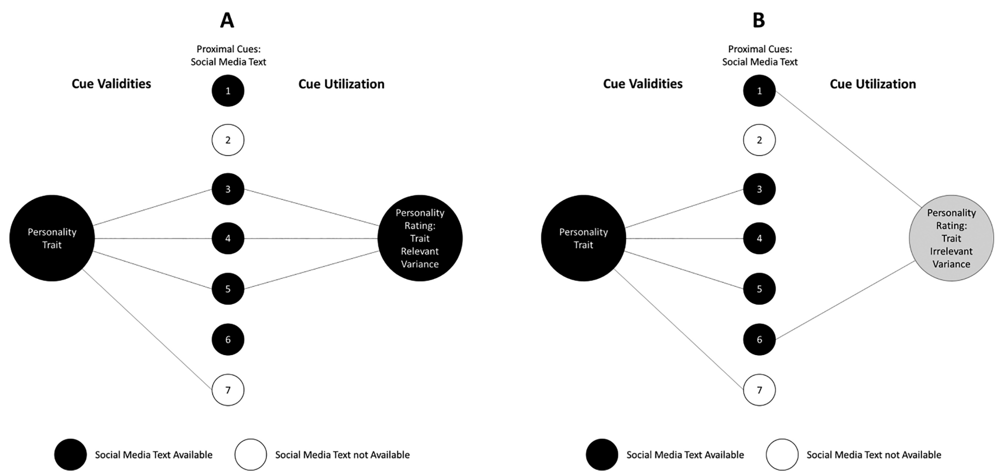

Figure 8. Brunswik's lens model of personality trait ratings on social media texts: the cases of trait-relevant and trait-irrelevant variance in trait ratings.

DOI: 10.1002/per

This may occur because of text preprocessing that reduces the informativeness of the text or because of the machine learning algorithms implemented. As shown in Scenario 7, there can be social media text data that are unavailable but related to the personality trait. This will also reduce the validity of the machine learning algorithms applied to social media text data.

Figure 8B shows the case for when variance in the ratings is trait *irrelevant*. Trait-irrelevant variance may arise from random error or systematic variance (e.g. variance associated with other constructs rather than the targeted one such as social desirability; specific method variance). In this case, because the variance in the ratings is trait irrelevant, machine learning algorithms will utilize all the personality trait-irrelevant cues (i.e. Scenarios 1 and 6). In Scenario 2, social media text data are unavailable and actually unrelated to the personality trait. In this case, the problem of trait-irrelevant variance in self-reports or other-reports of personality leading the misuse of irrelevant social media text data may be reduced.

These illustrations are intended to show that the validity of SMTM approaches depends centrally on the personality trait ratings themselves and specific scenarios where invalidity can occur. At the same time, what is also important to note from Figure 8 is that the cue utilization (i.e. relations between social media text data and scores from self-ratings or other-ratings of personality) *does not* actually reveal information about cue validities (i.e. relations between social media text data and underlying personality trait) themselves. Interestingly, using a particular measure of personality (e.g. self-reports) as ground truth in developing a predictive algorithm and then using the same information as a reference to validate the algorithm as a way of measuring personality is *circulus in probando*. Consider that this is quite different from typical personality scale creation where scale items are generated based on theory and independent of other personality scales that measure the same construct. Convergent validity evidence can be obtained because these independent measures converge. Contrast this with the fact that personality scores obtained from SMTM are not independent of (but inductively derived to maximize convergence with) the ground truth personality measure. This is one of the key reasons why not all researchers (ourselves included) view MLPA convergence with the personality trait ratings (self-reports or other-reports) as convergent validity evidence; at best, this would reflect predictive accuracy.

As shown in Figure 8, the cue validities are posited as invariant regardless of whether self-reported or other-reported personalities have trait relevant or irrelevant variance. In seeking to assess whether SMTM approaches are valid, it is important to focus on these cue validities. In thinking about validity and researchers' efforts toward validation, we need to address questions that arise from the concept of reference. To what extent do these social media language features clearly map onto the construct of interest? To what extent do these social media language features hang together to reflect the underlying personality factor? To what extent are the algorithms geared toward cue utilizations actually capable of identifying valid cues?

Causality: how does personality affect social media language behaviour?

Most (if not all) widely used validation efforts only offer peripheral information regarding MLPA validity by focusing on its ability to *predict* other variables (e.g. self-reported personality and future behaviour), but they do not provide direct evidence for MLPA's validity as a measure of personality. A scientific realist's confidence in the validity of SMTM will be established if she or he is convinced that scores derived from SMTM accurately reflect the purported construct of personality. In this regard, the best validation strategy would be to demonstrate a *causal* link between one's personality and the behavioural data extracted from social media. In other words, validation research should aim to theorize and verify the *response processes* in which personality causes the way one engages in social media language behaviours (e.g. frequency of using particular words, and patterns of language use over time and across platforms).

To this end, it is critical to examine person-specific psychological (cognitive and affective) mechanisms through which one uniquely encodes a situation and generates behavioural responses in the form of language on social media (Mischel & Shoda, 1995). Any systematic inter-individual (i.e. between-person) differences found in such intra-individual (i.e. within-person) patterns of encoding, affect, cognition, and behaviour can then be linked with known measures of personality as well as with other key psychological constructs. This type of validation research will significantly advance our understanding of how personality, together with other person and situational factors, affects the way social media language behaviours are generated. In the following paragraphs, we elaborate on two interrelated areas where future research will be informative.

Differing motives for engaging in social media

Social media, as the name suggests, is social (outward-facing) in nature and, as such, captures a unique sphere of an individual's behavioural repertoire (identity vs. reputation; Hogan, 1983). Most salient functions of social media include getting in contact with new people, keeping in touch with friends, sharing experiences/opinions, and socializing (Brandtzæg & Heim, 2009). Indeed, a significant reason for why individuals start and continue using social media is because of the number of other peers who are using it (Lin & Lu, 2011), and many features of social media seek to create significant social interactions such as co-viewing of videos (Haridakis & Hanson, 2009). Another significant motivation behind social media use appears to be informational. Brandtzæg and Heim (2009) found that access to information, including current events, trends, fashion, and so forth, was the fourth most important reason for social media use. Analyses of Instagram motives also found that knowing more about others was one of the main reasons (Sheldon & Bryant, 2016). In summarizing the reasons why people use social media sites, a multidimensional scaling study found that the two dimensions of fun-related (social oriented) and content-related (information oriented) emerged (Luchman, Bergstrom, & Krulikowski, 2014).

DOI: 10.1002/per

These socio-informational motives are likely the driving force behind individuals' behavioural patterns on social media, and according to the social–cognitive approach to personality, these motives are defined as part of the personality system (Saef, Woo, Carpenter, & Tay, 2018). More specifically, the Big Five personality traits capture specific patterns of activation, accessibility, interrelationships, and organization of social–cognitive mechanisms that guide the way in which individuals encode and interpret salient and consequential situational cues, which translate into behaviour (cf. DeYoung, 2015; Fleeson & Jayawickreme, 2015; Mischel & Shoda, 1995). An individual's unique web of social–cognitive mechanisms manifest individual differences in patterned responses to situational features, described as if–then behavioural signatures (e.g. If X is present, then I will exert more effort at work today; Mischel & Shoda, 1995). Social–cognitive mechanisms are information-processing units that define intra-individual consistencies in how people encode situational cues into their internal (psychological) system, which subsequently trigger cognitive and affective mechanisms such as affect, goals, and regulatory strategies that are closely interconnected.

Differential manifestation of personality-relevant behaviours on social media (vs. offline)

Previous research looking at how traditional conceptualizations of personality (mainly the Big Five personality traits) relate to online behaviour has provided initial insights into how these traits are expressed in online settings (e.g. Blackwell, Leaman, Tramposch, Osborne, & Liss, 2017; Gosling, Augustine, Vazire, Holtzman, & Gaddis, 2011; Saef et al., 2018). However, psychological research on how established personality traits predict specific patterns of online behaviour is still in its infancy. Not much empirical or theoretical insight into the existence of 'online personality expression' has been gained, and the majority of relevant studies come from a computer science rather than a psychological perspective. In addition, a large portion of existing studies have used self-reported, rather than actual, online behaviours (e.g. Blackwell et al., 2017; Gosling et al., 2011), which has limited our understanding of this relationship. However, the existing research has illustrated that the way in which certain personality traits manifest in online settings may differ from offline settings. For instance, Seidman (2013) found that neuroticism predicted sharing or venting on Facebook even though neuroticism has not been linked to socializing or information sharing in offline settings. Differences in how personality is expressed online versus offline could be due to differences in the structure and features of online social media contexts (as compared with offline contexts; McFarland & Ployhart, 2015), which means the prescribed encoding and social–cognitive responses will vary. These differences may call for empirical efforts toward expanding person × situation frameworks to incorporate features of online contexts.

Behavioural manifestations of personality are, at least in part, a function of contextual characteristics (Tett & Guterman, 2000). We suggest that whole trait theory (WTT; Fleeson & Jayawickreme, 2015) of personality may be used to outline why research must incorporate specific features of social media context to understand better how personality manifests online versus offline. WTT is an integrative framework of personality that describes the mechanisms through which personality traits explain differences in responses to situations. According to WTT, the *whole* of personality is made up of two halves: descriptive and explanatory. The descriptive part is defined by the density distribution of expressions (i.e. behaviour) characteristic of each personality trait, and the explanatory half encompasses the social–cognitive mechanisms that define individual differences in patterned responses to situational features. Therefore, given the systematic difference in the features of, and social dynamics within online versus offline contexts (McFarland & Ployhart, 2015), behavioural expressions of personality established in previous research may not translate to online settings. Rather, research must consider how the unique features of online social media platforms elicit certain behavioural patterns reflecting different personality traits. The explanatory side of personality can be expanded by empirically examining patterns of social–cognitive responses to unique features of online social media platforms. For instance, personality theory must consider the different modes of communication across different online platforms. While all social media sites facilitate social interactions, the nature and structure of these interactions often differ from those in offline settings. For instance, online social interactions are often one-sided and asynchronous in nature, whereas there is a higher expectation of reciprocal social interactions in offline settings. Further, some platforms have higher levels of synchronicity or anonymity than others, which may differentially elicit extraversion-related or agreeable-related behaviours, just as an example. Systematically taking into account these unique features of social media contexts will be crucial for understanding how personality traits function to influence behaviours on social media (see McFarland & Ployhart, 2015, for a fuller discussion of the differences between social media and offline contexts).

GENERALIZABILITY: ARE SOCIAL MEDIA TEXT MINING APPROACHES FOR ASSESSING PERSONALITY APPLICABLE ACROSS PLATFORMS, USERS, AND ALGORITHMS?

In this section, we discuss generalizability as an important dimension of SMTM and other machine learning approaches for assessing personality, given the unique context of social media and opportunities for access to such data. In validation research for personality assessments, it is typical to consider issues of generalizability across various contexts (e.g. Nye, Roberts, Saucier, & Zhou, 2008). However, there is a surprising absence of this work in SMTM approaches. For example, the most successful studies in showing convergence with self-reports, such as Park et al. (2015), utilized the MyPersonality dataset (Celli, Pianesi, Stillwell, & Kosinski, 2013). The MyPersonality dataset was generated with a Facebook application wherein hundreds of thousands of people voluntarily completed self-reports of personality

DOI: 10.1002/per

and provided access to their Facebook profiles. For one thing, large datasets are necessary to ensure that any patterns identified as useful for predicting personality successfully cross-validate. However, the issue of generalizability across different social media platforms is not addressed even with the best of these studies. This issue is critical because there is a proliferation of different types of social media platforms and user populations. As discussed above, context is important for the expression of personality traits. We believe that with the growing ease of technologically enabled data collection on various social media platforms, more empirical research can be conducted to accumulate an estimation of true effect size correlations between human-generated personality judgement and machine learning-generated judgements and what contexts may moderate such correlations.

Differences across platforms: do they elicit different personality-relevant behaviours?

Are social media data sufficiently reliable and consistent across platforms in accessing personality-relevant information? Social media platforms are not unitary, and there are often substantial differences in the features among different types of platforms resulting in non-equivalent data (i.e. behaviour). Therefore, at the surface level, the different types of data and features that can be captured will vary across different platforms. For example, Instagram emphasizes image data as a means for communication, whereas LinkedIn emphasizes text data. While an exhaustive list of the features of different social media platforms is outside the scope of this paper, we direct interested readers to the Meyer Foundation (2014), which details some of the major differences between social media platforms. However, it is important to note that features of social media platforms do change over time (and sometimes quickly), which presents unique opportunities *and* challenges *when* working with social media platforms. For example, Twitter famously increased its character limit from 140 to 280 in late 2017, so while this provides the opportunity to collect more behavioural data, research conducted before this upgrade affected the amount of data researchers could collect. We simply note here that the prevalence of the types of text, visual, audio, and meta-data is different and that such differences need to be accounted for, as they are associated with differences in the types of data, available, and analyses that can be used to infer personality. This affects both the strength of validity evidence within each social media platform for assessing personality and the generalizability of findings across different types of platforms.

From the validity (and causality) standpoint, as discussed in the previous section, it is important to recognize that each social media platform imposes unique restrictions on personality expression. This means that the way in which personality traits function to cause behaviours on one social media platform may differ from another. For instance, Facebook allows users to provide biographical information (e.g. current employer and educational institution) and share timeline events and long posts, while Twitter allows users to 'follow' others (without them following you back) and only allows short posts limited to 280 characters. Differences in features will shape not only the language used by people but also how personality is expressed on these online platforms more generally. Similar to how situations create boundaries on the way in which traits are able to be expressed (Fleeson, 2007), social media platforms create restrictions on the way in which users can express their traits. As each social media platform has different features available for consuming and sharing content, they create different situational boundaries on how people can express their personality traits. In other words, because different social media platforms restrict the possible behaviours, the situations that they create differ in opportunities for personality expression. For instance, because Twitter has a character limit and Facebook does not, proper grammar may not be an indicator of conscientiousness on Twitter, while we may see that on Facebook. Therefore, the way that conscientiousness manifests in grammar on the two platforms may not converge. Further, there may also be differences in the likelihood of using the different features of social media platforms depending on people's personality, which affect researchers' ability to access trait-relevant behaviours online. We further discuss this issue in the following section.

Differences in users: who is included?

While there is an assumption that social media data may be valid for assessing personality in general populations because of how many people are on social media, this may not necessarily be the case and needs to be examined through more validation research. The reason for this is that the demographics of individuals using social media are not representative of national populations. For example, according to Pew, while a majority of Americans utilize some form of social media, the demographics utilizing these technologies are generally young, urban, more educated, and wealthier. On the most popular social media platform Facebook, for example, an estimated 81% of individuals aged 18–29 use it compared with 41% of people aged 65 and older. These trends generalize to other countries as well (Pew Research Center, 2018). These differences are helpful to consider as estimating personality distributions of populations may be limited for generalization without considering sampling and sample weights. This implies that there is a need to capture demographic information apart from personality-relevant behaviours for population inferences and population-level subgroup comparisons to enhance validity.

Another dimension to consider is technology accessibility and social media usage. Individuals who have smartphone technology have greater access to and usage of social media apps (Villanti et al., 2017). This implies that individuals who have smartphones likely have more posts and greater engagement. This results in differences in the reliability of assessing personality in systematic ways between individuals who have technology access as compared with those who do not. Personality appears to play into the use of technology and the amount of usage in the first place. For example, extraversion and openness to experience have been shown to be positively associated with technology acceptance (Svendsen, Johnsen, Almås-Sørensen, & Vittersø, 2013). More specifically, individuals who are extraverted,

DOI: 10.1002/per

agreeable, and neurotic reported more time spent on Facebook; extraversion and agreeableness also predicted more postings and photos (Moore & McElroy, 2012). Relying only on social media language in posts to assess personality does not account for initial base rate distributions in personality. For example, extraverted individuals may be more likely to have social media accounts and activity, and there will be fewer data points and representation for individuals who are less extraverted. In other words, these uneven distributions may need to be considered in at least three instances: inferring population-level means and distributions; range restriction when examining convergence with other established personality reports of the same trait; use of multiple features predicting personality where each feature may have differing underlying personality distributions (or influences); and even accuracies in predicting individual-level personalities when the amount of data (or missingness) may be confounded with the personality attribute one seeks to infer.

There are also demographic differences in individuals using different types of social media platforms (Pew Research Center, 2018). For example, the average LinkedIn user is older and more educated than Twitter users. These demographic differences are likely linked to personality differences. On average, for instance, individuals who are older (Roberts, Walton, & Viechtbauer, 2006) and more educated (Hampson, Goldberg, Vogt, & Dubanoski, 2007) are more conscientious. While the mean level differences in personality that arise from demographic differences across platforms may not directly affect validity, one should be aware of these differences as they may have indirect effects. For example, research has shown that college-age narcissists prefer using Twitter, whereas adult-age narcissists prefer using Facebook (Davenport, Bergman, Bergman, & Fearrington, 2014), implying that the presence of some personality dimensions may be more readily captured on specific platforms owing to demographic-based preferences in usage. Other researchers have similarly found differential relations between personality and social media usage between Facebook and Twitter (Hughes, Rowe, Batey, & Lee, 2012).

Differences in algorithms and data access: are all predictions made equal?

The final issue is that the process of creating machine learning models is often unstandardized across different researchers or research groups. This is unlike the creation of standardized personality instruments that can be administered and scored in a similar fashion in different contexts. There are several reasons for this. Foremost, the level of data access that researchers have is often different due to differences in resources. For instance, it is possible to purchase a full set of Twitter posts as opposed to using the Twitter firehose, where researchers may only have a sliver of personality-relevant social media language. This is further complicated given that what social media companies choose to share (e.g. firehose) or block from access (e.g. features) often changes over time. Apart from data access, researchers often use different preprocessing techniques, extracted features, and machine learning models. Different analytic choices can substantially affect results even when seeking to address the same research question (Silberzahn et al., 2018). For example, our meta-analysis reveals that the congruence between SMTM and self-reports can vary substantially in different studies, as seen from the large 95% confidence intervals across the different personality traits.

In all these validation efforts for generalizability and beyond, we believe that there is a critical need for more openness and transparency in MLPA. Personality science will be better served by an ecosystem that encourages replicable algorithms across different research groups. We also propose that a best-practice reporting standards be developed and widely shared, similar to JARS/MARS (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). The development of frameworks and software to create more open science around MLPA (and more generally, machine learning models in psychology) will be required to standardize, evaluate, and enhance the scientific quality of MLPA.

CLOSING THOUGHTS

The proliferation of social media use across different segments of the population and cultures enables us to overcome some of the concerns about the lack of inclusivity and representation of typically WEIRD (Western, Educated, Industrialized, Rich, Democratic) research participants that are used in psychology (Henrich, Heine, & Norenzayan, 2010). Further, the regularity of social media usage (e.g. majority of Facebook individuals visit the site once or more a day; Pew Research Center, 2018) provides a window for researchers to capture language and other behaviour in the context of everyday life which complements other modes of data for understanding personality (Cattell, 1957). In short, mining of social media text data represents an exciting possibility for advancing behavioural personality measurement techniques beyond existing methods, such as self-reports and other-reports that are often viewed as limited by human perceptions (Woo, Tay, & Proctor, 2020). We urge personality psychologists—academics and practitioners alike—to actively engage in discussions about validity and engage in more validation research, which we believe will lead to significant scientific advancements of behavioural personality research.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *The American Psychologist*, *63*, 839–851. https://doi.org/10.1037/0003-066X.63.9.839.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, *21* (3), 372–374. https://doi.org/10.1177/0956797609360756.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26.

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C., & Liss, M. (2017). Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. *Personality and Individual Differences*, *116*, 69–72. https://doi.org/10.1016/j.paid.2017.04.039.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(), 1088868318772990, 190–203. https://doi.org/10.1177/1088868318772990.

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

Brandtzæg, P. B., & Heim, J. (2009). Why people use social networking sites. In A. A. Ozok, & P. Zaphiris (Eds.), *Online communities and social computing* (pp. 143–152). Berlin Heidelberg: Springer.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Caspi, A., Begg, D., Dickson, N., Harrington, H., Langley, J., Moffitt, T. E., & Silva, P. A. (1997). Personality differences predict health-risk behaviors in young adulthood: Evidence from a longitudinal study. *Journal of Personality and Social Psychology*, *73*, 1052–1063. https://doi.org/10.1037/0022-3514.73.5.1052.

Cattell, R. B. (1957). *Personality and motivation structure and measurement (1958-03918-000)*. Yonkers-on-Hudson, N.Y.: World Book Co.

Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013, June 28). *Workshop on computational personality recognition: Shared task*. *Seventh International AAAI Conference on Weblogs and Social Media*. Seventh International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6190.

Cloninger, C. R., Sigvardsson, S., & Bohman, M. (1988). Childhood personality predicts alcohol abuse in young adults. *Alcoholism, Clinical and Experimental Research*, *12*, 494–505. https://doi.org/10.1111/j.1530-0277.1988.tb00232.x.

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*, 110–117. https://doi.org/10.1111/j.1468-2389.2007.00371.x.

Dahlke, J. A., & Wiernik, B. M. (2019). Psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*, *43*, 415–416. https://doi.org/10.1177/0146621618795933.

Davenport, S. W., Bergman, S. M., Bergman, J. Z., & Fearrington, M. E. (2014). Twitter versus Facebook: Exploring the role of narcissism in the motives and usage of different social media platforms. *Computers in Human Behavior*, *32*, 212–220. https://doi.org/10.1016/j.chb.2013.12.011.

DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of Research in Personality*, *56*, 33–58. https://doi.org/10.1016/j.jrp.2014.07.004.

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestion in behavior. *Journal of Personality*, *75*, 823–861.

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *523-76*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009.

Ford, M. T., Jebb, A. T., Tay, L., & Diener, E. (2018). Internet searches for affect-related terms: An Indicator of subjective well-being and predictor of health outcomes across US states and metro areas. *Appled Psychology: Health and Well-Being*, *10*, 3–29. https://doi.org/10.1111/aphw.12123.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670. https://doi.org/10.1037/0033-295X.102.4.652.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*, 177–182. https://doi.org/10.1177/0963721412445309.

Garcia, D., & Sikström, S. (2014). The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. *Personality and Individual Differences*, *67*, 69–74. https://doi.org/10.1016/j.paid.2013.10.001.

Garcia, D., & Sikström, S. (2014). The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. *Personality and Individual Differences*, *67*, 92–96.

Gill, A. J., Oberlander, J., & Austin, E. (2006). Rating e-mail personality at zero acquaintance. *Personality and Individual Differences*, *40*, 497–507. https://doi.org/10.1016/j.paid.2005.06.027.

Golbeck, J. (2016). Predicting personality from social media text. *AIS Transactions on Replication Research*, *2*, 1–10. https://doi.org/10.17705/1atrr.00009.

Gosling, S. D., Augustine, A. A., Vazire, S., Holtzman, N., & Gaddis, S. (2011). Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. *CyberPsychology, Behavior & Social Networking*, *14*, 483–488. https://doi.org/10.1089/cyber.2010.0087.

Haig, B. D. (2020). Big data science: A philosophy of science perspective. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 15–33). Washington, DC: American Psychological Association. https://doi.org/10.1037/0000193-002.

Hall, J. A., Pennington, N., & Lueders, A. (2014). Impression management and formation on Facebook: A lens model approach. *New Media & Society*, *16*, 958–982. https://doi.org/10.1177/1461444813495166.

Hampson, S. E., Goldberg, L. R., Vogt, T. M., & Dubanoski, J. P. (2007). Mechanisms by which childhood personality traits influence adult health status: Educational attainment and healthy behaviors. *Health Psychology*, *26*, 121–125. https://doi.org/10.1037/0278-6133.26.1.121.

Haridakis, P., & Hanson, G. (2009). Social interaction and co-viewing with YouTube: Blending mass communication reception and social connection. *Journal of Broadcasting & Electronic Media*, *53*, 317–335. https://doi.org/10.1080/08838150902908270.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83. https://doi.org/10.1017/S0140525X0999152X.

Hinds, J., & Joinson, A. (2019). Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science*, *28*, 204–211. https://doi.org/10.1177/0963721419827849.

Hogan, R. (1983). A socioanalytic theory of personality. In M. M. Page (Ed.), *Nebraska symposium on motivation 1982. Personality: Current theory and research* (pp. 55–89). Lincoln: University of Nebraska press.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346. https://doi.org/10.1177/1088868310361240.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of

DOI: 10.1002/per

social media usage. *Computers in Human Behavior*, *28*, 561–569. https://doi.org/10.1016/j.chb.2011.11.001.

Ivcevic, Z., & Ambady, N. (2012). Personality impressions from identity claims on Facebook. *Psychology of Popular Media Culture*, *1*(1), 38–45. https://doi.org/10.1037/a0027329.

Jin, Y. (2006). *Multi-objective machine learning 16*. Berlin, Heidelberg: Springer Science & Business Media.

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: Methodologies and challenges. *Psychological Methods*, *21*, 507–525. https://doi.org/10.1037/met0000091.

Kluemper, D. H., & Rosen, P. A. (2009). Future employment selection methods: Evaluating social networking web sites. *Journal of Managerial Psychology*, *24*, 567–580. https://doi.org/10.1108/02683940910974134.

Kluemper, D. H., Rosen, P. A., & Mossholder, K. W. (2012). Social networking websites, personality ratings, and the organizational context: More than meets the eye? *Journal of Applied Social Psychology*, *42*, 1143–1172. https://doi.org/10.1111/j.1559-1816.2011.00881.x.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*, 5802–5805. https://doi.org/10.1073/pnas.1218772110.

Lin, K.-Y., & Lu, H.-P. (2011). Why people use social networking sites: An empirical study integrating network externalities and motivation theory. *Computers in Human Behavior*, *27*, 1152–1161. https://doi.org/10.1016/j.chb.2010.12.009.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Luchman, J. N., Bergstrom, J., & Krulikowski, C. (2014). A motives framework of social media website use: A survey of young Americans. *Computers in Human Behavior*, *38*, 136–141. https://doi.org/10.1016/j.chb.2014.05.016.

Marcus, B., Machilek, F., & Schütz, A. (2006). Personality in cyberspace: Personal web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology*, *90*, 1014–1031. https://doi.org/10.1037/0022-3514.90.6.1014.

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review*, *123*(5), 569–591. https://doi.org/10.1037/rev0000035.

McAdams, D. P. (1996). Personality, modernity, and the storied self: A contemporary framework for studying persons. *Psychological Inquiry*, *7*, 295–321. https://doi.org/10.1207/s15327965pli0704_1.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*, 28–50. https://doi.org/10.1177/1088868310366253.

McFarland, L. A., & Ployhart, R. E. (2015). Social media: A contextual framework to guide research and practice. *Journal of Applied Psychology*, *100*, 1653–1677. https://doi.org/10.1037/a0039244.

Meyer Foundation (2014). *Social media platform comparison: Key channels, trends & features to inform your story sharing process*. https://www.meyerfoundation.org/sites/default/files/files/SWT-Platform-Comparison-090414.pdf.

Mischel, W., & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268. https://doi.org/10.1037/0033-295X.102.2.246.

Moore, K., & McElroy, J. C. (2012). The influence of personality on Facebook usage, wall postings, and regret. *Computers in Human Behavior*, *28*, 267–274. https://doi.org/10.1016/j.chb.2011.09.009.

Nye, C. D., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality*, *42*, 1524–1536. https://doi.org/10.1016/j.jrp.2008.07.004.

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

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Austin Texas.

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

Pew Research Center. (2018). *Social Media Use in 2018*. Retrieved from https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/02/PI_2018.03.01_Social-Media_FINAL.pdf.

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*, 710–718. https://doi.org/10.1016/j.jrp.2012.08.008.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25. https://doi.org/10.1037/0033-2909.132.1.1.

Rogers, K. H., & Biesanz, J. C. (2019). Reassessing the good judge of personality. *Journal of Personality and Social Psychology*, *117*, 186–200. https://doi.org/10.1037/pspp0000197.

Roulin, N., & Levashina, J. (2019). LinkedIn as a new selection method: Psychometric properties and assessment approach. *Personnel Psychology*, *72*, 187–211. https://doi.org/10.1111/peps.12296.

Saef, R., Woo, S. E., Carpenter, J., & Tay, L. (2018). Fostering socio-informational behaviors online: The interactive effect of openness to experience and extraversion. *Personality and Individual Differences*, *122*, 93–98. https://doi.org/10.1016/j.paid.2017.10.009.

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *The Journal of Applied Psychology*, *104*, 1207–1225. https://doi.org/10.1037/apl0000405.

Sanford, F. H. (1942). Speech and personality. *Psychological Bulletin*, *39*, 811–845. https://doi.org/10.1037/h0060838.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*, e73791. https://doi.org/10.1371/journal.pone.0073791.

Seidman, G. (2013). Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and Individual Differences*, *54*, 402–407. https://doi.org/10.1016/j.paid.2012.10.009.

Sheldon, P., & Bryant, K. (2016). Instagram: Motives for its use and relationship to narcissism and contextual age. *Computers in Human Behavior*, *58*, 89–97. https://doi.org/10.1016/j.chb.2015.12.059.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356. https://doi.org/10.1177/2515245917747646.

Solomon, B. C., & Vazire, S. (2016). Knowledge of identity and reputation: Do people have knowledge of others' perceptions? *Journal of Personality and Social Psychology*, *111*, 341–366. https://doi.org/10.1037/pspi0000061.

Srinivasan, P. (2020). Text mining: A field of opportunities. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological* *research* (pp. 179–199). Washington, DC: American Psychological Association. https://doi.org/10.1037/0000193-009.

Stopfer, J. M., Egloff, B., Nestler, S., & Back, M. D. (2014). Personality expression and impression formation in online social networks: An integrative approach to understanding the processes of accuracy, impression management and meta-accuracy. *European Journal of Personality*, *28*(1), 73–94. https://doi.org/10.1002/per.1935.

Svendsen, G. B., Johnsen, J.-A. K., Almås-Sørensen, L., & Vittersø, J. (2013). Personality and technology acceptance: The influence of personality factors on the core constructs of the technology acceptance model. *Behaviour & Information Technology*, *32*, 323–334. https://doi.org/10.1080/0144929X.2011.553740.

Tellegen, A. (1991). Personality traits: Issues of definition, evidence, and assessment. In *Thinking clearly about psychology: Essays in honor of Paul E. Meehl, Vol. 1: Matters of public interest; Vol. 2: Personality and psychopathology* (pp. 10–35). Minneapolis: University of Minnesota Press.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, *34*, 397–423. https://doi.org/10.1006/jrpe.2000.2292.

Tskhay, K. O., & Rule, N. O. (2014). Perceptions of personality in text-based media and OSN: A meta-analysis. *Journal of Research in Personality*, *49*, 25–30. https://doi.org/10.1016/j.jrp.2013.12.004.

van de Ven, N., Bogaert, A., Serlie, A., Brandt, M. J., & Denissen, J. J. A. (2017). Personality perception based on LinkedIn profiles. *Journal of Managerial Psychology*, *32*(6), 418–429. https://doi.org/10.1108/JMP-07-2016-0220.

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*, 1811–1835. https://doi.org/10.1177/0149206313515524.

Vazire, S., & Gosling, S. D. (2004). e-Perceptions: Personality impressions based on personal websites. *Journal of Personality and Social Psychology*, *87*(1), 123–132. https://doi.org/10.1037/0022-3514.87.1.123.

Villanti, A. C., Johnson, A. L., Ilakkuvan, V., Jacobs, M. A., Graham, A. L., & Rath, J. M. (2017). Social media use and access to digital technology in US young adults in 2016. *Journal of Medical Internet Research*, *19*, e196. https://doi.org/10.2196/jmir.7303.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, *7*, 203–220. https://doi.org/10.1177/1745691612442904.

Woo, S. E., Tay, L., Jebb, A., Ford, M. T., & Kern, M. (2020). Big data for enhancing measurement quality. *Big Data in Psychological Research*, Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/0000193-004.

Woo, S. E., Tay, L., & Proctor, R. (2020). *Big Data in Psychological Research*, Washington, DC: American Psychological Association.