

Tilburg University

Exploring the application of a text-to-personality technique in job interviews

Holtrop, Djurre; Oostrom, Janneke K.; Breda, Ward R. J van; Koutsoumpis, Antonis; Vries, Reinout E. de

Published in:
European Journal of Work and Organizational Psychology

DOI:
[10.1080/1359432X.2022.2051484](https://doi.org/10.1080/1359432X.2022.2051484)

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Holtrop, D., Oostrom, J. K., Breda, W. R. J. V., Koutsoumpis, A., & Vries, R. E. D. (2022). Exploring the application of a text-to-personality technique in job interviews. *European Journal of Work and Organizational Psychology*, 31(6), 799-816. <https://doi.org/10.1080/1359432X.2022.2051484>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



Exploring the application of a text-to-personality technique in job interviews

Djurre Holtrop, Janneke K. Oostrom, Ward R. J van Breda, Antonis Koutsoumpis & Reinout E. de Vries

To cite this article: Djurre Holtrop, Janneke K. Oostrom, Ward R. J van Breda, Antonis Koutsoumpis & Reinout E. de Vries (2022) Exploring the application of a text-to-personality technique in job interviews, *European Journal of Work and Organizational Psychology*, 31:6, 799-816, DOI: [10.1080/1359432X.2022.2051484](https://doi.org/10.1080/1359432X.2022.2051484)

To link to this article: <https://doi.org/10.1080/1359432X.2022.2051484>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 06 Apr 2022.



Submit your article to this journal [↗](#)



Article views: 1022








View related articles [↗](#)



View Crossmark data [↗](#)

Exploring the application of a text-to-personality technique in job interviews

Djurre Holtrop ^{a,b}, Janneke K. Oostrom ^c, Ward R. J van Breda ^{d,e}, Antonis Koutsoumpis ^e
and Reinout E. de Vries ^e

^aTilburg University, Department of Social Psychology, Simon Building, Tilburg, The Netherlands; ^bCurtin University, The Future of Work Institute, Faculty of Business and Law, Australia; ^cSchool of Business & Economics, Department of Management & Organisation, Vrije Universiteit Amsterdam, The Netherlands; ^dNeedForward Research, The Netherlands; ^eVrije Universiteit Amsterdam, Department of Experimental and Applied Psychology, The Netherlands

ABSTRACT

This research's purpose was to develop a valid and transparent text-to-personality technique to fit the requirements for personnel selection assessments. In this research we developed an advanced word-counting technique, the HEXACO text-to-personality (HTTP) technique, based on prior lexical personality research to assess personality from job interviews. To evaluate the technique's construct and criterion-related validity we conducted three studies and analysed the transcripts of asynchronous ($n = 102$ and 72) and face-to-face ($n = 155$) interviews. These studies provided four key insights. First, the HTTP technique showed small to medium correlations with self-reported and interviewer-rated personality. Second, the technique showed mixed, but generally favourable, evidence for criterion-related validity. Third, the technique produced a more construct valid personality score when the interview questions activated the predicted personality trait. Fourth, the technique's additional features (i.e., having weighted keywords and adjusting the keywords' weight for adjacent quantifiers) did not improve its validity; unit-weighting was approximately equally effective. Altogether, the results show that a word-count text-analysis technique can discover traces of personality in interview transcripts. Still, significant improvements are needed before these types of automatically computed text-to-personality ratings can be used to replace or supplement interviewer ratings.

ARTICLE HISTORY

Received 22 October 2020
Accepted 2 March 2022

KEYWORDS

Personality; interviews; text-analysis; trait-activation; word count; personnel selection

Introduction

Organizations are rapidly adopting a wide range of new technologies to automatically evaluate job interviews (e.g., Woods et al., 2020). These technologies often involve algorithms, which are computational techniques with varying levels of complexity that automate the assessment of job-related characteristics and replace or assist the recruiter when making selection decisions (Mirowska, 2020; Parasuraman et al., 2000). Especially for large scale selection programmes, these techniques reduce costs and decision time. Furthermore, as research has consistently shown that computational data combination techniques outperform human judgements in terms of reliability and criterion-related validity (Grove et al., 2000; Kuncel et al., 2013), these techniques potentially allow for a more valid assessment of job-related characteristics. However, despite their popularity in practice, it is still unclear whether these techniques actually provide valid job interview ratings (Woods et al., 2020).

The most promising technology to automatically assess job-relevant psychological constructs in job interviews are text-analysis techniques (e.g., Hickman et al., 2021; Mairesse et al., 2007; Park et al., 2015; Schwartz et al., 2013; Tausczik & Pennebaker, 2011). Although many organizations have adopted this technique, academic research that has applied text-analysis techniques to job interviews is scarce. Therefore, the goal of this paper is to develop a text-

analysis technique based on past lexical personality research and apply this technique to transcripts of structured job interviews. Specifically, we developed the HEXACO text-to-personality word-counting technique (HTTP technique) to measure the HEXACO personality traits based on verbal utterances in job interviews and we conducted three studies to assess its construct and criterion-related validity. The HTTP technique uses weighted keyword lists and considers quantifiers (e.g., "very") adjacent to the keywords to improve its accuracy. We compared the validity of the HTTP technique to that of a traditional keyword-counting technique that counts all words equally. Furthermore, as the type of interview questions is likely to affect the validity of a candidate's response for the assessment of specific personality dimensions, we also investigated the effects of using trait activating (Tett & Burnett, 2003) versus trait irrelevant interview questions on the validity of the HTTP technique.

Our study contributes to the literature in a number of ways. First, we developed a technique that is specifically designed to assess personality traits as objectively and transparently as possible. Although there have been earlier attempts to assess personality traits based on verbal utterances, these approaches rely heavily on subjectivity in the keyword classifications (e.g., Mairesse et al., 2007; Tausczik & Pennebaker, 2011), or use black-box algorithms (e.g., Hickman et al., 2021; Naim et al.,

2018). Second, we specifically contribute to the selection and assessment research by applying such a text-to-personality technique to job interviews. Most text-to-personality techniques are built on written texts obtained from social network websites (e.g., Park et al., 2015; Schwartz et al., 2013). As there are substantial differences between *informal* – written – language and *formal* – spoken – language, thus most previous findings cannot simply be generalized to job interviews. Third, our study is one of the first to explore the criterion-related validity of a text-to-personality technique for work-related criteria. Previous studies showed that text-based personality ratings are related to relevant (self-reported) social psychological criteria (e.g., number of friends; Park et al., 2015) and academic achievement (Hickman et al., 2021). To date, however, no studies have related text-based personality ratings to the most crucial criterion in selection contexts, namely, job performance.

Personality Assessment in Personnel Selection Contexts

According to personality psychologists (e.g., Allport & Odbert, 1936), salient individual differences are encoded in our language. Through factor analysis of self-ratings on personality-descriptive words, the underlying structure of personality can be exposed. Research using this so-called “lexical hypothesis” led to the emergence of one of the most successful models in psychology, the Big Five model of personality (Goldberg, 1990), and its successor, the six-dimensional HEXACO model of personality (Ashton et al., 2004). Personality is an important and frequently measured psychological construct in personnel selection contexts, as various aspects of personality are important predictors of work outcomes (Sackett & Walmsley, 2014; Schmitt, 2014). Most notably, the personality trait conscientiousness substantially and incrementally predicts work performance over and above cognitive ability (e.g., Oh et al., 2014; Salgado, 1997; Schmidt & Hunter, 1998). Other personality traits tend to predict performance in specific jobs (Barrick & Mount, 1991), or other work-related outcomes, such as honesty-humility predicting counterproductive work behaviour (Pletzer et al., 2019).

In personnel selection contexts, personality traits are typically measured with self-report personality inventories. However, there is a continuing debate on the value of self-report personality inventories as they are susceptible to faking (Anglim et al., 2017; Morgeson et al., 2007), rely on the ability and motivation to introspect accurately (De Cuyper et al., 2017), and suffer from measurement artefacts (e.g., primacy and recency effects or consistency motivation; Paulhus & Vazire, 2007). Consequently, there has been a search for alternative techniques to measure personality, including job interviews (Levashina et al., 2014) and, more recently, Asynchronous Video Interviews (AVIs; Lukacik et al., 2020). AVIs are digital job interviews in which applicants first record their responses to interview questions with a webcam. At a later moment in time, recruiters can review the applicants’ responses.

Early meta-analytical studies showed that job interviews often contain personality-related questions and information (e.g., Cortina et al., 2000; Huffcutt et al., 2001). Subsequent studies showed that job interviews can also be designed to specifically assess personality traits (e.g., Barrick et al., 2000;

Blackman, 2002; Van Iddekinge et al., 2005) and that interviewers’ (other-reported) personality ratings are less susceptible to faking than interviewees’ (self-reported) personality ratings (Van Iddekinge et al., 2005). Interviewers’ ratings, however, vary in quality (Melchers et al., 2011). Several studies have shown that not everyone is equally good at making predictions from interviews or text excerpts (Dipboye et al., 2012; Hall et al., 2016; Van Iddekinge et al., 2006). Furthermore, interviewers’ personality ratings are influenced by irrelevant candidate characteristics (e.g., attractiveness or gender; Huffcutt, 2011) and similarity effects (Graves & Powell, 1996). Finally, conducting and scoring interviews is cognitively demanding (Christiansen et al., 2005) and time consuming (i.e., costly).

In sum, despite the importance of measuring personality traits in personnel selection contexts and the compelling evidence that personality can be assessed with job interviews, the evaluation process is highly dependent on interviewer idiosyncrasies and expensive. We argue that, for job interviews and AVIs (which are recorded by default), text-to-personality techniques have great potential to make more valid personality assessments while at the same time reducing costs.

Personality Assessment with Text-analysis Techniques

Over time, a wide range of text-analysis techniques have been developed, ranging from somewhat automated word-counting software to nearly fully automated Natural Language Processing (NLP) programs (Banks et al., 2018). The most notable and frequently applied text-analysis techniques can be categorized into two methods: the closed-vocabulary techniques and the open-vocabulary techniques (Park et al., 2015).

The closed-vocabulary technique (also referred to as content analysis or dictionary-based methods; Banks et al., 2018) is the oldest technique and uses a top-down approach to generate a predefined and limited lists of keywords. Each keyword list represents one topic (e.g., extraversion) and is based on active, theory-driven decisions by its designers. Advantages of the closed-vocabulary technique are that it can be used for short texts with less than 250 words (Muralidhar et al., 2018) and that it is relatively transparent as long as it is not combined with machine learning. The LIWC program (pronounced “Luke”; Pennebaker et al., 2015) is a popular application of the closed-vocabulary technique. It is essentially a word-counting program that computes a trait score based on how often people mention words associated with that trait relative to the total amount of words. In the development of LIWC, a large number of words has been assigned to (sometimes several) categories by human experts. Hence, even though interrater reliability can be checked, the classification of these keywords remains rather subjective. To our knowledge, the closed-vocabulary technique has only been used to measure personality dimensions indirectly (e.g., L. L. Chen et al., 2016; Mehl et al., 2006). However, no research to date used a top-down technique to develop word lists for personality traits. Moreover, this technique has not yet been applied in personnel selection contexts.

The open-vocabulary technique is a more recent technique to obtain text-based (personality) ratings and includes bag-of-words approaches (Banks et al., 2018), language-based assessment (e.g., Park et al., 2015), and NLP techniques. In its

simplest form, these techniques can effectively create a list of words for any trait. That list can then be applied to other texts to assess personality traits, in a similar fashion to the closed-vocabulary technique (i.e., by counting words). More advanced open-vocabulary methods, such as the NLP technique “BERT” (Devlin et al., 2018) no longer work with word lists, but instead aim to understand the context in which words are used. Most open-vocabulary techniques require large quantities of data for optimization, but still appear to be the preferred approach in personnel selection contexts. In fact, organizations (e.g., Hirevue, Outmatch, Talview, and many more) are rapidly developing NLP tools to assess candidates’ personality traits based on AVIs, often in combination with assessments of non-verbal and para-verbal variables through machine learning. A well-known example of a text-analysis program that is based on advanced open-vocabulary techniques is IBM Watson Personality Insights (Hickman et al., 2019; IBM Watson, 2019).

When evaluating a new technique for personnel selection purposes, construct and criterion-related validity are the foremost characteristics (e.g., Evers et al., 2009; Sackett et al., 2018). The closed-vocabulary technique has not yet been applied to job interviews, so evidence in terms of construct validity is lacking. Open-vocabulary methods show weak to moderate convergence with self-ratings and medium to strong convergence with interviewer ratings (e.g., L. L. Chen et al., 2016; Hickman et al., 2021; Naim et al., 2018). We note that all these studies investigated the construct validity of text analysis with a combination of closed- (i.e., LIWC) and open-vocabulary (i.e., Word-2-vec, *n*-grams, or Latent Dirichlet Allocation) techniques and machine-learning statistics. Construct validity evidence for the same programs seems to differ between samples and contexts. For example, IBM Watson (2019) reports an average correlation of .31 between their Personality Insights, validated on tweets (Hickman et al., 2019), and self-reported personality traits. In contrast, when Hickman et al. (2019) applied this same technique to short AVIs, they discovered low and conflicting construct validity coefficients. Importantly, we could not find any research linking these estimates to job performance.

Although open-vocabulary techniques have found their way into practice and their validity evidence is promising, these techniques lack transparency (Kar & Dwivedi, 2020) and are often seen as a black box (Cheng & Hackett, 2021). This lack of transparency is problematic, because candidates will only trust personnel selection techniques if they have adequate explainability (Liem et al., 2018). Indeed, the opacity of advanced algorithms for personnel selection is such a concern that the EU is in the process of labelling such approaches to employment and personnel selection as high risk – the highest risk category before being considered unacceptable (Liboreiro, 2021). Already, there have been instances of litigation against organizations for using “opaque” analysis technology to infer personality-like characteristics from AVIs (Harwell, 2019). Altogether, there is a societal push for more transparent techniques to evaluate interviews.

The Present Study

Having concluded that 1) existing closed-vocabulary techniques possess limited *construct validity* for personality assessment and 2) most open-vocabulary methods lack *explainability*, we considered alternative text-analysis approaches to evaluate job interviews. We hasten to add that both closed- and open-vocabulary techniques have a strong body of research behind them and it would be unrealistic to address all limitations in one attempt.

Specifically, we built on existing closed-vocabulary techniques with the aim to improve their validity by not relying on subjective ratings and using more complete word lists. Therefore, to avoid the closed-vocabulary technique’s dependency on subjective human judgment, the HTTP technique is not based on expert judgements but on lexical studies that produced lists of words that people use to describe each personality trait (Ashton et al., 2004; De Raad & Barelds, 2008; De Raad & Hoskens, 1990). Thus, the premise of our approach is that people use the very same words to describe themselves in job interviews that are found in lexical studies to reflect personality traits.

The HTTP technique differs from other closed-vocabulary approaches in two ways, in addition to the fact that it constitutes the first closed-vocabulary attempt to measure personality directly. First, to compute the HEXACO personality traits, the HTTP technique uses weighted keyword lists. The loading of each keyword on each personality trait was derived from the lexical studies, which showed clear differences in the extent to which words are representative for each personality trait. In contrast, other closed-vocabulary techniques unit-weight all keywords that load on the same characteristic. Second, the HTTP technique uses a grammar engine (Sentimentics, 2019) to adjust the keywords’ factor loadings for words that change the keywords’ meaning (e.g., not, very, a little). For example, consider that “cheerful” is an indicator of high extraversion. The grammar engine allows the HTTP technique to rate “not cheerful” as an indicator of low extraversion, whereas other closed-vocabulary techniques would consider this an indicator of high extraversion. The first purpose of our series of studies was to establish the construct validity of the HTTP technique by relating its ratings to corresponding self-rated (Study 1, 2, and 3) and observer-rated (Study 2) personality traits. We expected that the HTTP technique would show positive correlations with personality ratings from other sources.

Hypothesis 1: HTTP traits are positively related to corresponding self-reported personality traits.

Hypothesis 2: HTTP traits are positively related to corresponding observer-rated personality traits.

In the search to optimize construct validity of text-analysis techniques, some researchers have suggested to process text more nuancedly than only counting keywords. For example, Park et al. (2015) suggested future research to use a grammar engine. Hence, the second purpose of this investigation was to

discover if the weighted loadings and grammar engine provide additional validity over a more straightforward keyword count for automated personality assessment. For this comparison, we used the keyword lists from the HTTP technique to construct a simple word-counting dictionary. Subsequently, the construct validity of the HTTP scales was compared to the construct validity of the keyword-counting technique without extra features.

Hypothesis 3: HTTP traits are more strongly related than keyword-counted traits to corresponding a) self-reported and b) observer-rated personality traits.

Psychologists (Tett & Burnett, 2003) have posited that traits are expressed as a response to trait-relevant cues (e.g., interview questions). Following the trait-activation perspective, job interviews that activate specific traits generate more trait relevant responses (e.g., Lievens et al., 2006; Speer et al., 2015). As such, a job interview question that targets a specific personality trait is likely to elicit a response that is most informative for the targeted trait and less informative for other traits. These differences should also affect the construct validity of text-analysis assessments. Nonetheless, studies of text analysis in job interviews have not explored this perspective yet (Hickman et al., 2021).

Hypothesis 4: HTTP traits based on responses to trait-relevant questions are more strongly related to corresponding self-reported and observer-rated personality traits than HTTP traits based on responses to trait-irrelevant questions.

Woods et al. (2020) argue that, alongside our limited understanding of construct validity, the lack of knowledge about the criterion-related validity of new selection techniques is the most critical gap in the literature. To expand our knowledge about the criterion-related validity of text-analysis techniques, we tested the HTTP techniques' criterion-related validity by relating its traits scores to known correlates of self-reported personality: self-reported stress and self-efficacy (Study 1), self-reported job performance (Study 2), and supervisor-rated job performance (Study 3). We expected our technique to predict these

socioemotional and performance outcomes for several reasons. First, personality is a relevant predictor of these job-relevant outcomes (e.g., Sackett & Walmsley, 2014; Schmitt, 2014; Zettler et al., 2019). Second, interviews are able to reveal relevant personality-related information (e.g., Barrick et al., 2000; Blackman, 2002; Van Iddekinge et al., 2005). Third, this information is likely to be encapsulated in the words that candidates speak during an interview and the HTTP technique is designed to capture these words.

Hypothesis 5: HTTP traits will be significantly correlated with socioemotional and performance outcomes (see, Table 1 for a summary of the predictions).

Last, we compared the criterion-related validity of HTTP traits and self-reported personality traits. Compared to interviews, self-reports of personality are easy to administer and cost-effective (Heimann et al., 2020). Considering that recruiters are often looking for efficiency and accuracy, it is prudent to compare any alternative measure of personality to the current (most efficient) best practices. Therefore, we formulated the following research question.

Research question: Do HTTP traits and/or self-reported traits have incremental validity over one another for predicting socioemotional and performance outcomes?

Overview of Studies and General Method

In all three studies (see, Table 1 for an overview), participants responded to interview questions targeting specific personality traits, which were analysed both by the HTTP and keyword-counting technique. Additionally, all participants completed a version of the HEXACO Personality Inventory-Revised (Lee & Ashton, 2006). The three studies in this manuscript represent various ways of conducting job interviews, two studies used asynchronous interviews (i.e., pre-recorded questions and recorded responses; Study 1 and 3) and one used traditional face-to-face interviews (Study 2). In all cases the interviews were structured and prompting was minimal. All interviews were audio recorded.¹ In this section, we describe the HTTP technique, the keyword-counting technique, and the HEXACO personality inventory, which were used in all three studies. For

Table 1. Overview of Studies.

Number	Sample	N	Interview type	Trait targeted (by n interview questions)	Construct validity for targeted trait	Criterion-related validity for targeted trait
1	Heterogeneous snowball sample	102	Asynchronous	Emotionality (3)	HEXACO-100: Emotionality	Self-rated perceived stress (+) Self-rated general self-efficacy (-)
				Extraversion (3)	HEXACO-100: Extraversion	Self-rated perceived stress (-) Self-rated general self-efficacy (+)
2	Heterogeneous snowball sample	155	Face-to-face	Honesty-Humility (4)	HEXACO-200: Honesty-Humility Observer-rated: Honesty-Humility	Self-rated normative behaviour (+)
				Extraversion (4)	HEXACO-200: Extraversion Observer-rated: Extraversion	
				Conscientiousness (4)	HEXACO-200: Conscientiousness Observer-rated: Conscientiousness	Self-rated Task performance (+)
3	Employees engineering firm	72	Asynchronous	Conscientiousness (5)	HEXACO-60: Conscientiousness	Supervisor-rated job performance (+)

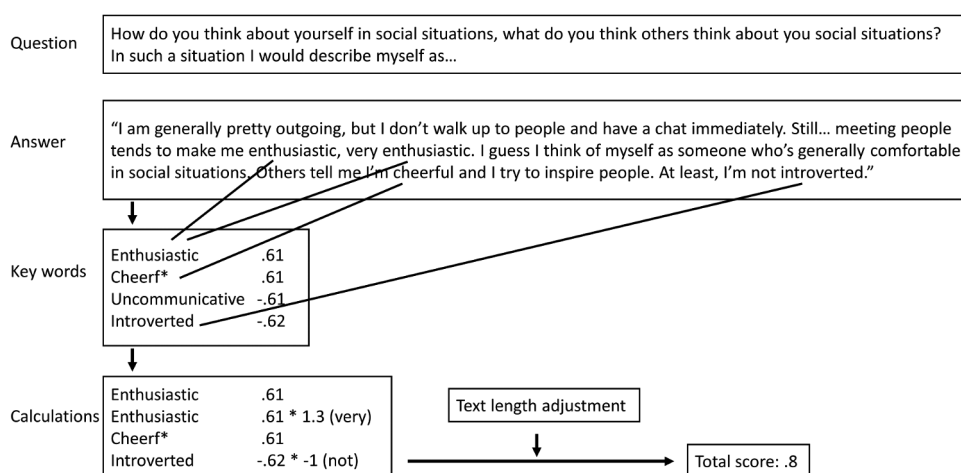


Figure 1. Hypothetical example of HTTP technique's procedure for estimating extraversion from text. First words that match the keyword list are identified from the interview transcript. Cheerf* is stemmed. Second, the relevant weights are applied to each word. Third, the weights are adjusted for adjacent modifiers (e.g., not, very). Finally, the scores are combined and adjusted for frequency relative to text length resulting in a score ranging from -1 to 1 .

a detailed description of the development of the HTTP technique and the keyword-counting technique please refer to this project's Open Science Framework (OSF) webpage.²

The HEXACO Text-to-Personality (HTTP) Technique

To develop the HTTP technique, we obtained the HEXACO library of 551 Dutch adjectives (De Vries et al., 2009), 755 Dutch nouns (De Raad & Hoskens, 1990), and 2331 Dutch nouns, adjectives, and verbs (De Raad & Barelds, 2008). The factors from the lexical data from De Raad and Hoskens (1990) and De Raad and Barelds (2008) were not perfectly aligned with the HEXACO dimensions, respectively, they found a seven- and eight-factor structure. To determine which of these factors appropriately reflected the HEXACO dimensions, we evaluated the content of the seven factors (De Raad & Hoskens, 1990) and the correlation coefficients with the HEXACO traits of the eight factors (De Raad & Barelds, 2008). Next, prior to merging the three keyword lists, we removed overlapping keywords, standardized the factor loadings, and removed keywords with weak loadings. The end result was a list of 2,652 Dutch stemmed keywords for all six HEXACO personality dimensions with 493 (extraversion) to 979 (emotionality) words per trait with loadings from -1.00 to $-.30$ and $.30$ to 1.00 .

Unless otherwise mentioned, the HTTP traits presented in this manuscript were based on the responses to interview questions that specifically activated the respective trait. See, Figure 1 for an illustration of how the HTTP technique produces these scores. First, the program searched the text for the words that also appeared in the keyword lists. Second, it applied the appropriate factor loading, obtained from the lexical research, to each word found in the text. Third, the matched words from the text and related factor loadings were corrected for syntactically related negative/positive adverbs (e.g., not, very, somewhat). Last, all trait information was combined, resulting

in a score on an interval scale ranging from -1 to 1 . A score of -1 corresponds with a negative trait presence, 0 corresponds to an equal amount of information indicating negative and positive trait presence, and 1 corresponds with positive trait presence. In this calculation, the score was adjusted for text length, because longer texts allow people to generate more hits. When a participant did not respond with any word from the word list for that trait, which was not unusual, the trait score for that response was coded as missing.

The Keyword-Counting Technique

To compare the HTTP technique to the more straightforward keyword-counting technique, we took the keyword lists that we created to construct the HTTP technique and, following the guidelines in the operator manual, transformed these lists into a LIWC (Pennebaker et al., 2015) dictionary. The keyword-counting technique was applied to the textual information in the same way as the HTTP technique; only questions high in trait activation were separately analysed and these values were averaged to arrive at an overall trait score.

The HEXACO Personality Inventory

In all three studies we measured self-rated personality with HEXACO personality inventories (60-form; Ashton & Lee, 2009, pp. 200-form; Lee & Ashton, 2006; or 100-form, 2018).³ Responses are provided on a 5-point Likert scale, ranging from $1 = strongly disagree$ to $5 = strongly agree$. In our studies we used the Dutch version of the HEXACO personality inventory (De Vries et al., 2009). Using the Dutch HEXACO personality inventory, De Vries et al. (2009) obtained alpha reliabilities ranging from $.71$ to $.79$ for the 60-item form, $.75$ to $.84$ for the 100-item form, and $.85$ to $.91$ for the 200-item form.

Methods and Results

Study 1

The purpose of Study 1 was to investigate the construct validity of the HTTP technique and test its criterion-related validity with self-reported criteria. To test its construct validity, we chose to focus on the personality dimensions extraversion and emotionality because these dimensions have relatively high observability (Funder, 1995; Funder & Dobroth, 1987) and self-other agreement (Lee & Ashton, 2006; Lee et al., 2009). To test the criterion-related validity of the HTTP technique, we selected two work-relevant individual outcomes that relate differentially to extraversion and emotionality, namely general self-efficacy and perceived stress levels. Previous research (Ebstrup et al., 2011) found extraversion to be positively related to self-efficacy and negatively related to perceived stress levels, whilst finding the opposite pattern for neuroticism (which has substantial overlap with emotionality from the HEXACO model).

Participants and Procedure

The data for Study 1 were collected in two stages. First, participants completed an online survey; second, they completed an asynchronous digital interview. Participants were recruited by a post-graduate student through his social network. Only Dutch-speaking participants were included in the current study. Participants who completed the study went into a draw for 10 cash prizes of €50. When participants agreed to participate in the study, they received an email with the survey link and instructions to record – after completing the survey – their responses to the interview questions. Both the survey and the interview questions could be completed at a location of participants' convenience. Altogether, the online survey took around 15 minutes to complete. The online survey was completed by 187, mostly female, participants ($\text{♀} = 78.1\%$, $M(\text{age}) = 23.12$ years, $SD(\text{age}) = 6.08$).

The Interview

The instructions for the interview were emailed to the participants and included a PowerPoint attachment that contained the interview questions. On the first slide, the PowerPoint file instructed the participants to orally respond to six open-ended questions in less than two minutes. We also instructed them to "Use as many adjectives as you can to describe your behaviour", and provided some examples. The interview questions were split up between extraversion and emotionality. Two questions (no. 1 and 3) were constructed to activate both traits simultaneously (i.e., mixed trait activation), two questions (no. 4 and 5) to only activate emotionality (i.e., single trait activation), and two questions (no. 2 and 6) to only activate extraversion (please see this project's OSF webpage for all interview protocols). Upon completing all six questions, the participants saved the voice recording and sent it to the test leader.

Of the participants who had completed the survey, 104 (55.6%) participants also submitted their asynchronous interview recording. This recording was then manually transcribed. One of the participants delivered the interview in writing and one participant delivered an inaudible audio file. Therefore, 102 interviews were available to process with the text-based personality rating software ($\text{♀} = 83.3\%$, $M(\text{age}) = 23.59$ years,

$SD(\text{age}) = 6.03$). The HTTP traits were based on the program's assessment of the responses to the four questions activating the trait (i.e., single and mixed trait activation combined), these values were averaged into one score per trait. On average, participants generated 783 words in response to the four questions activating emotionality ($SD = 269$) and 781 words to the four questions activating extraversion ($SD = 272$). A number of participants did not use any words from a trait's HTTP word list when responding to one or more questions. Therefore, to estimate the reliability of the HTTP traits, missing values were imputed. McDonald's ω (Dunn et al., 2014) was found to be .42 for HTTP emotionality and HTTP extraversion (i.e., for the combination of single and mixed trait activation questions). Further investigation showed that the reliability could not be substantially improved by disregarding a specific text-based personality rating. Following a similar procedure for the keyword-counting scores, McDonald's ω was .64 for keyword-counted emotionality and .70 for keyword-counted extraversion.

Additional Instruments

HEXACO-PI-R

Emotionality and extraversion were measured with items of the HEXACO-PI-R 100. Both traits showed a good McDonald's ω , respectively .80 and .87 ($n = 102$).

New General Self-Efficacy Scale

The NGSES (G. G. Chen et al., 2001) measures "individuals' perception of their ability to perform across a variety of different situations" (pp. 170, Judge et al., 1998). The NGSES was translated to Dutch by the authors (the translated scales can all be found on the OSF webpage of this project). Participants rated the extent to which the statements were applicable to them on a 5-point Likert-scale, ranging from 1 = *strongly disagree* to 5 = *strongly agree*. Previous research found a high level of alpha reliability (.85; G. G. Chen et al., 2001); for the Dutch items, the present study found an equivalent high McDonald's ω (.84, $n = 102$).

Perceived Stress Scale

The PSS (Cohen et al., 1983) measured the self-reported perceived stress of the participants with 14 items. These items were also translated to Dutch by the authors. The items were rated on a 5-point Likert scale, ranging from 1 = *never* to 5 = *very often*. Cohen et al. (1983) reported internal consistencies of .84, .85, and .86, the current study found a McDonald's ω of .88 ($n = 102$) for the Dutch items.

Results

Table 2 shows the correlations between the self-reported personality traits, the text-based personality traits, the criteria (self-efficacy and perceived stress), and the keyword-counted traits. Hypothesis 1, that HTTP traits would be positively related to corresponding self-reported traits, was supported by the moderately positive correlations between the HTTP and self-

Table 2 Study 1: Correlations and descriptive statistics

		M	SD	Age	Gender	Self-reported personality		HTTP technique		Keyword count		Self-reported criteria		Word count
						Emot.	Extr.	Emot.	Extr.	Emot.	Extr.	GSE	PS	
Self-reported personality	Emotionality	3.35	0.50	-.13	.26**	.80								
	Extraversion	3.52	0.54	-.07	-.15	-.15	.87							
HTTP technique	Emotionality	.01	.19	-.12	.13	.45**	-.33**	.42						
	Extraversion	.18	.16	-.16	.07	-.08	.36**	-.30**	.42					
Keyword count	Emotionality	-0.31	1.58	-.10	.24*	.37**	-.19	.49**	-.11	.64				
	Extraversion	1.01	1.10	.02	-.12	-.24*	.24*	-.35**	.40**	-.69**	.70			
Self-reported criteria	General Self-Efficacy	3.80	0.51	.04	-.19	-.10	.59**	-.31**	.24*	-.27**	.22*	.84		
	PerceivedStress	2.71	0.52	-.09	.17	.43**	-.58**	.42**	-.22*	.26**	-.11	-.39**	.88	
Word count	Emotionality	783	269	-.03	.24*	.23*	.08	.07	.20*	.30**	-.43**	.08	-.05	-
	Extraversion	781	272	-.07	.24*	.22*	.04	.06	.21*	.30**	-.44**	.07	.00	.96**

Note. N = 102. McDonald ω reliabilities are italicized on the diagonal. Hypothesized relations are **bolded**. Self-reported personality was measured with the HEXACO-PI-R 100-form (Lee & Ashton, 2018), general self-efficacy was measured with the GSE (G. G. Chen et al., 2001), and perceived stress was measured with the PSS (Cohen et al., 1983). Scores on self-reported personality, general self-efficacy, and perceived stress ranged from 1–5, scores from the text-based personality ratings ranged from –1 to 1, Gender was coded as male = 1 and female = 2.

* p < .05, ** p < .01

Table 3. Study 1: Correlations between self-reported personality scores and HTTP scores based on responses to interview questions varying in trait activation.

HTTP scores split on level of trait activation	Self-reported personality					
	Emotionality			Extraversion		
	r	Z-difference with high activation	n	r	Z-difference with high activation	n
High activation	.45**	-	102	.36**	-	102
Mixed Emotionality and Extraversion	.23*	1.79*	102	.20*	1.24	101
Low activation (other trait)	.04	3.16**	102	.30**	0.47	86

Note. Six interview questions were asked in total, pertaining to either emotionality, extraversion, or a mix. Therefore, two questions prompted each level of trait activation. The number of participants differs for the correlation coefficients because of missing values. Missing values represent participants who did not utter a key word in response to either question with this level of trait activation.

* p < .05, ** p < .01

reported personality traits ($r = .45$ for emotionality and $r = .36$ for extraversion). Using Fisher’s r -to- z -to- r transformation the average correlation between the text-based and self-reported personality traits was $\bar{r} = .41$, showing signs of construct validity. The HTTP traits showed higher correlations with the self-reported personality traits than the keyword-counted traits (.45 vs .37 for emotionality and .36 vs .24 for extraversion). Although the direction of the findings was in line with Hypothesis 3, which stated that keyword weights and a grammar engine would improve construct validity, the correlations between the HTTP and self-reported personality traits did not significantly differ from the correlations between the keyword-counted and self-reported personality traits (respectively $Z = 0.88, p > .05$ and $Z = 1.16, p > .05$). Consequently, Hypothesis 3 was not supported.

Hypothesis 4 stated that HTTP traits based on responses to questions with high levels of trait activation are more valid than HTTP traits based on responses to questions with low levels of trait activation. Therefore, the average HTTP scores for both traits were calculated for each set of questions (high, mixed, and low activation). Table 3 shows the correlations between self-reported personality and the HTTP traits for the different levels of trait activation. For HTTP emotionality, high trait activation scores correlated significantly stronger with self-reported emotionality ($r = .45$) than mixed trait activation ($r = .23, Z = 1.79, p(\text{difference}) < .05$) or low trait activation

scores (i.e., activating extraversion instead: $r = .04, Z = 3.16, p(\text{difference}) < .01$). For HTTP extraversion, high trait activation scores did not correlate significantly stronger with self-reported personality ($r = .36$) than mixed trait activation ($Z = 1.24, p > .05$) or low trait activation scores ($Z = 0.47, p > .05$). In short, the results showed mixed support for trait-activation effects.

The HTTP traits showed mostly similar correlations with the criteria as self-reported personality did. HTTP emotionality correlated significantly and negatively with GSE ($r = -.31$) and positively with perceived stress ($r = .42$). Inversely, HTTP extraversion correlated negatively with GSE ($r = .24$) and positively with perceived stress ($r = -.22$). These correlations are largely similar to previous research findings and to those of the self-reported personality traits in this study (with the exception of self-reported emotionality and GSE: $r = -.10, p = .31$), therefore offering support for Hypothesis 5. The research question addressing whether the HTTP traits have incremental validity over self-reported personality traits or vice versa, was investigated with several regression analyses (see, Table 4). Only HTTP emotionality, and not HTTP extraversion, showed incremental validity over self-reported personality.

Discussion

Overall, HTTP emotionality and extraversion showed construct validity by having medium correlations with their self-reported counterparts. The more straightforward keyword-counting technique showed similar construct validity to the

Table 4. Study 1: Stepwise regression analyses for Emotionality and Extraversion with General Self-Efficacy and Perceived Stress.

		Independent variables: Emotionality scores											
		Dependent variable: General self-efficacy						Dependent variable: Perceived stress					
		R^2	ΔR^2	R^2	ΔR^2	b^1	b CI ¹	R^2	ΔR^2	R^2	ΔR^2	b^1	b CI ¹
Step 1	Constant					3.904	3.00 to 4.81					1.59	0.75 to 2.44
	Age	.037				0.001	-0.02 to 0.02	.036				0.00	-0.02 to 0.01
Step 2	Gender					-0.230	-0.50 to 0.04					0.08	-0.17 to 0.33
	Self-rated	.040	.003					.189	.153**				
Step 3	Text-based			.120	.083**					.192	.156**		
	Self-rated			.126	.006	0.091	-0.13 to 0.31			.253	.061**	0.30**	0.09 to 0.50
	Text-based	.126	.086**			-0.921**	-1.51 to -0.33	.253	.064**			0.81**	0.25 to 1.36
		Independent variables: Extraversion scores											
		Dependent variable: General self-efficacy						Dependent variable: Perceived stress					
		R^2	ΔR^2	R^2	ΔR^2	b^1	b CI ¹	R^2	ΔR^2	R^2	ΔR^2	b^1	b CI ¹
Step 1	Constant					2.01	1.19 to 2.83					4.67	3.83 to 5.51
	Age	.037				0.01	-0.01 to 0.02	.036				-0.01	-0.03 to 0.00
Step 2	Gender					-0.15	-0.37 to 0.08					0.12	-0.11 to 0.33
	Self-rated	.363	.326**					.360	.324**				
Step 3	Text-based			.104	.067**					.095	.059*		
	Self-rated			.366	.262**	0.52**	-0.36 to 0.70			.362	.266**	-0.54**	-0.70 to -0.37
	Text-based	.366	.003			0.21	-0.44 to 0.82	.362	.001			-0.15	-0.77 to 0.48

Note. Each combination of personality trait (emotionality or extraversion) and criterion (general self-efficacy or perceived stress) was tested with two regression analyses to estimate the incremental validity of text-based personality ratings over self-reported personality and vice versa. Gender was coded as 1 = male and 2 = female. ¹The reported Confidence Intervals (CI) of the unstandardized regression coefficients (b) are for the complete regression model only.

* $p < .05$, ** $p < .01$

HTTP technique. For emotionality, the construct validity of the HTTP technique was higher when interview questions activated the trait, but not for extraversion. Finally, the HTTP traits showed criterion-related validity that approached that of self-reported personality traits. Altogether, this study shows that text-based personality ratings from a closed-vocabulary technique have some potential for personality assessment, especially when respondents are activated to provide information relevant to the trait.

The most salient limitation of the present study is the low reliability of the HTTP traits. In contrast, the keyword-counting technique yielded more acceptable reliability coefficients. To explain this difference in reliability, the differences between the keyword-counting technique and HTTP technique (i.e., the grammar engine and weighted word list) should be considered. The interview questions focused on different topics that varied in situational strength (e.g., general social situation, leadership situation, crisis situation). Possibly, some situations trigger the use of strongly weighted words and other situations do not, affecting the HTTP ratings to a great extent from one situation to another, reducing reliability.

In addition to the HTTP technique's low reliability, another noteworthy limitation is that the participants were not instructed to behave as job applicants. Applicant instructions would promote impression management and therefore affect the personality trait scores. Therefore, it is important to test the validity of the HTTP technique in a selection setting.

Study 2

The purpose of Study 2 was threefold. The first purpose was to test the construct and criterion-related validity of the HTTP technique based on face-to-face job interviews with applicant instructions. The second purpose was to compare the HTTP technique's ratings to those of human observers. The third purpose was to assess two other personality traits, namely honesty-humility and conscientiousness, and (again) extraversion.

Although asynchronous interviews are gradually gaining popularity, face-to-face interviews remain overwhelmingly popular. The main difference between face-to-face and asynchronous interviews is that face-to-face interviews involve back-and-forth communication with another human. A large volume of research has shown that interpersonal interaction shapes the way job applicants respond (e.g., Dipboye et al., 2012; Word et al., 1974). We decided it was important to also test the HTTP technique with face-to-face interviews because of their popularity and their notable differences with asynchronous interviews. Additionally, this study uses observer-rated personality, in addition to self-reported personality, to assess the construct validity of the HTTP technique. In practice, the application of text analysis is likely to substitute or supplement recruiter ratings. Therefore, it could be argued that the correlations of HTTP traits with observer ratings are the most vital test of the technique's construct validity.

In addition to extraversion, this study targeted two other traits that are valid predictors of work outcomes: honesty-humility and conscientiousness. Honesty-humility is predictive of counterproductive behaviours (Pletzer et al., 2019) and conscientiousness is the most predictive personality dimension of task performance across jobs (Sackett & Walmsley, 2014; Schmidt & Hunter, 1998). Consequently, to test the criterion-related validity of HTTP honesty-humility and conscientiousness, we asked participants to report normative and task performance.

Participants and Procedure

The data for Study 2 were collected in two stages. First, the participants completed an online survey at a time and place of their own choosing. Second, the participants were instructed to behave as job applicants for a fictitious organization and answered 12 interview questions. Participants were recruited and interviewed by two post-graduate students via their social network and interested organizations (82%, $n = 127$), and through the university's participant pool (18%, $n = 28$). Upon completion of the study, participants were offered €20 compensation.

Altogether, 155 Dutch-speaking people participated in this study. One hundred-forty-nine participants were born in the Netherlands and the remaining six participants were born in various other countries. Gender was distributed equally in this sample ($\text{♀} = 49\%$) and the average age was 29.81 years ($SD = 12.18$). One hundred twenty-nine participants had either a part- or fulltime job, with an average of 6.80 ($SD = 9.50$) years of work experience in their field. Eighty-six participants studied either part- or fulltime. The most prevalent educational level was post-graduate education (51.9%), followed by undergraduate education (24.1%).

The Interview

Prior to commencing the mock job interview, the participants were asked to read an introduction to a fictitious company called "Utopia & Co" and imagine that they were applying for a position (of their own choosing). Participants were informed that the interview allowed approximately two minutes per question. The interviewers were instructed to use simple prompts to help participants to respond for the full two minutes, such as "Can you give an example of that?" or "Can you tell me a bit more about [repeat question]?" Altogether, the interviewers asked the participants 12 questions about conscientiousness, extraversion, and honesty-humility (four per trait). On average, the participants spoke 609 ($SD = 168$) words while responding to the honesty-humility questions, 488 ($SD = 169$) words to the extraversion questions, and 565 ($SD = 190$) words to the conscientiousness questions.

Again, a substantial number of participants responded to one or more trait-relevant questions without saying a word from the HTTP keyword lists. Similar to Study 1, missing values were imputed for calculating the reliabilities and results showed low (extraversion $\omega = .43$; conscientiousness: $\omega = .57$) to very low reliability (honesty-humility: $\omega = .01$). Using a similar

procedure for the keyword-counting technique, the McDonald's ω was .10 for honesty-humility, .39 for extraversion, and .57 for conscientiousness.

Additional instruments

HEXACO-PI-R

We used the items from the HEXACO-200 (Lee & Ashton, 2004) to assess honesty-humility, extraversion, and conscientiousness. These scales all showed good reliabilities, respectively McDonald's $\omega = .90$, .90, and .89.

Observer-rated personality

The interviews were each rated by two independent judges to represent ratings that would usually be given by recruiters or hiring managers. Immediately after the interview, the post-graduate student who had interviewed the participant rated the participant's personality. Later, each audio recording was also rated by one of two research assistants (with a completed post-graduate education). To rate the participants' personality from the interviews, we developed Behaviourally Anchored Rating Scales (BARS; Kingstrom & Bass, 1981) based on the HEXACO definitions of honesty-humility, extraversion, and conscientiousness (available on the OSF page). BARS are frequently recommended to improve the structure of job interviews (e.g., Dipboye et al., 2012; Levashina et al., 2014). The observer ratings were converted to Z-values for each observer to reduce the effect of leniency/severity. The one-way random effects ICC for consistency were .53, .57, and .68 for respectively the ratings of honesty-humility, extraversion, and conscientiousness. These inter-rater reliabilities are similar to the ones reported in the meta-analysis by Conway et al. (1995) and somewhat lower than the ones reported by Heimann et al. (2020).

Work Performance Questionnaire

Self-reported job performance ratings were obtained from all participants who indicated having a job ($n = 129$) with the work performance questionnaire (WPQ). The WPQ is under development by the authors and, as such, no previous information on the psychometric qualities of this instrument is available. The current study reports the results for the six-item *task performance* scale and the six-item *normative performance* scale (i.e., the reverse of counterproductive behaviour). Responses were provided on a Likert-scale ranging from 1 = *never* to 7 = *very often*. The McDonald's ω for the task performance scale was .71 and .69 for the normative performance scale.

Results

Table 5 shows the correlations and descriptive statistics for honesty-humility, extraversion, and conscientiousness (self-rated, observer-rated, HTTP, and keyword-counted scores), self-rated normative and task performance, and the word count of the four interview questions for each trait. On average, self-rated personality traits correlated strongly with the

Table 5. Study 2: Correlations between text-mining, observer, and self-reported personality ratings and self-reported performance.

	M	SD	Age	Gender	Self-report				Observer-rated				HTTP technique				Keyword count				Performance				
					HH	Ex	Co	HH	Ex	Co	HH	Ex	Co	HH	Ex	Co	HH	Ex	Co	HH	Ex	Co	Norm.	Task	HH
Self-reported personality	HH	3.58	0.49	.34**	.30**	.90																			
	Ex	3.77	0.41	-.21**	-.01	-.12	.90																		
Observer-rated personality	Co	3.50	0.47	.21**	.09	.19*	-.04	.89																	
	HH	0.01	0.82	.20*	.24**	.42**	-.13	.16*	.53																
HTTP technique	Ex	0.00	0.84	-.10	.02	-.12	.64**	-.07	-.13	.57															
	Co	0.00	0.87	.21**	.17*	.15	.01	.55**	-.02	.68															
Keyword count	HH	.21	.15	-.04	.16	.17*	-.01	-.04	.20*	.11	.01														
	Ex	.13	.17	-.16*	.06	-.08	.38**	-.06	.03	.33**	.02	.16*	.43												
Self-reported performance	Co	.12	.16	.16*	.09	.07	.05	.19*	.14	.02	.26**	.02	.04	.57											
	HH	2.05	1.07	.16	.12	.25**	-.16*	-.12	.17*	-.05	.02	.44**	.01	-.02	.10	.16	.56**	.02	.24**	.39	-.04	.57			
Word Count	Ex	1.15	1.44	-.05	-.01	-.09	.17*	-.12	-.04	.10	-.04	.10	.06	.19*	-.03	.00	.59**	-.03	-.04	.11	.11	.69			
	Co	0.47	0.82	.19*	.00	.09	.04	.24**	.08	.06	.17	.17	.09	.06	.19*	.01	.07	.14	.05	.23**	.38**	.71			
Performance	Norm.	6.16	0.62	.08	.05	.30**	-.07	.16	.05	-.16	.17	.12	.09	.06	-.24**	-.03	.05	-.02	-.08	-.09	-.09	-.06			
	Task	5.26	0.66	-.10	.04	.11	.05	.19*	-.08	.03	.12	-.13	-.13	-.09	.09	.06	.08	.09	-.11	.05	-.17	.03	.37**		
Word Count	HH	608.52	168.36	-.20*	-.10	-.03	.20*	-.11	.01	.27**	-.23**	.03	.12	-.13	-.02	-.02	-.02	-.08	-.11	.05	-.17	.03	-.37**		
	Ex	488.15	169.20	-.17*	.04	-.10	.18*	.02	.05	.24**	.07	.06	.08	.09	-.31**	-.11	.05	-.11	.05	-.11	.05	-.17	.03	.37**	
Word Count	Co	565.49	189.68	-.11	.05	-.03	.18*	.00	.05	.26**	-.03	.04	.14	.08	-.29**	-.04	.01	-.09	-.09	-.09	.04	.49**	.69**		

Note. $N = 149-155$ for all personality correlates and $n = 127-129$ for performance correlates. McDonald ω reliabilities and one-way random effects ICC for consistency are italicized on the diagonal. Hypothesized correlations are **bolded**. The four text-based ratings for honesty-humility showed negative covariance. Self-reported personality was measured with the HEXACO-PI-R 200-form (Lee & Ashton, 2004). Scores on self-reported personality ranged from 1-5, scores from the text-based personality ratings ranged from -1 to 1, observer-rated personality scores were z-scored, self-reported work performance scores ranged from 1-7, Gender was coded as male = 1 and female = 2.

* $p < .05$, ** $p < .01$

Table 6. Study 2: Correlations between self-reported personality scores and HTTP scores based on responses to interview questions varying in trait activation.

HTTP scores split on level of trait activation	Self-reported personality								
	Honesty-Humility			Extraversion			Conscientiousness		
	<i>r</i>	Z-difference with high activation	<i>n</i>	<i>r</i>	Z-difference with high activation	<i>n</i>	<i>r</i>	Z-difference with high activation	<i>n</i>
High activation	.17*		155	.38**		155	.19*		152
Low activation (other traits)	-.07	2.14*	155	-.03	4.02**	153	.04	1.25	148
	Observer-rated personality								
	<i>r</i>	Z-difference with high activation	<i>n</i>	<i>r</i>	Z-difference with high activation	<i>n</i>	<i>r</i>	Z-difference with high activation	<i>n</i>
	High activation	.20*		155	.33**		155	.26**	
Low activation (other traits)	.06	1.25	155	-.04	3.53**	153	.09	1.58	148

Note. The number of participants differs for the correlation coefficients because missing values. Twelve interview questions were asked in total, four pertaining to each honesty-humility, extraversion, and conscientiousness. High activation is the average HTTP score for the four questions activating that trait, low activation are HTTP scores based on the remaining eight questions.

* $p < .05$, ** $p < .01$

corresponding observer-rated traits ($\bar{r} = .54$). When assessing the text-based personality ratings' construct validity, all HTTP traits correlated significantly with the corresponding self-reported ($\bar{r} = .25$) and observer-rated traits ($\bar{r} = .26$) and did not correlate significantly with any of the non-corresponding traits. These findings lend support to Hypotheses 1 and 2 that HTTP traits have construct validity. However, the correlations with self-reported and observer-rated personality were small for HTTP honesty-humility (respectively $r = .17$, $p < .05$ and $r = .20$, $p < .05$) and HTTP conscientiousness (respectively $r = .19$ and $.26$). Only HTTP extraversion showed moderately strong correlations (respectively $r = .38$ and $r = .33$).

When comparing the HTTP technique to simple keyword counting, only HTTP extraversion correlated significantly stronger with self- and observer-reported extraversion than keyword-counted extraversion did (respectively: $r = .38$ vs $.17$, $Z = 2.91$, $p(\text{difference}) < .01$ and $r = .33$ vs $.10$, $Z = 3.13$, $p(\text{difference}) < .01$). Contrary to expectations, the correlations of HTTP conscientiousness and honesty-humility with self- and observer-reported personality were not significantly different from those of simple keyword counting. In conclusion, we found weak support for Hypothesis 3 that HTTP traits outperform keyword-counted traits; only HTTP extraversion was significantly more construct valid than the keyword-counted trait.

To investigate if trait activation affected the construct validity of the HTTP technique, the HTTP scores for each trait were also calculated based on responses to the remaining eight questions that did not activate the trait. The correlations of the HTTP traits based on high and low trait activation with self-reported and observer-rated personality are displayed in Table 6. The construct validity of activated traits was only significantly higher in three out of six comparisons. As such, Hypothesis 4 – that HTTP traits based on trait activating questions are more construct valid than HTTP traits based on trait-irrelevant questions – was partly supported.

Finally, when assessing the HTTP technique's criterion-related validity, HTTP honesty-humility did not correlate significantly with normative work performance ($r = .12$, $p > .05$) and HTTP conscientiousness correlated negatively with task performance ($r = -.24$, $p < .01$). These findings for the HTTP traits were not only contrary to expectations, they were also in contrast to the correlations of self-reported personality traits: both self-

reported honesty-humility and conscientiousness were significantly correlated with their respective outcomes (respectively $r = .30$, $p < .01$ and $r = .19$, $p < .05$). Thus, the HTTP traits did not show the expected criterion-related validity and Hypothesis 5 was not supported, as a consequence we refrained from testing the research question.

Discussion

The average construct validity of the traits produced by the HTTP technique were somewhat lower in Study 2 than in Study 1. The correlations of the HTTP traits with both self-reported and observer-rated personality traits were small to medium. This indicates that, in simulated selection contexts, closed-vocabulary text analysis has low to medium construct validity. Interestingly, the HTTP technique showed similar agreement with observer-rated personality and self-reported personality, despite the fact that text analysis and observer ratings are largely based on the same information (i.e., the verbal utterances during the interview). Perhaps the observers' judgment was affected by signals that the HTTP technique is not privy to, such as non-verbal and paraverbal signals (Dipboye et al., 2012). Alternatively, observers may pay attention to different verbal cues than the HTTP technique does and thus arrive at somewhat different ratings.

In terms of trait activation, the results, combined with similar partial support in Study 1, indicate that word-counting approaches (based on lexical research) are likely to be more construct valid when they are applied to responses to trait-relevant questions. Part of this increased construct validity might be that trait-relevant questions seemed to substantially increase the frequency with which participants spoke words that were included in the word list. That is, closed-vocabulary approaches cannot estimate trait levels without participants using the words in the word list.

Overall, the results were most pronounced for extraversion. First, HTTP extraversion showed the highest convergence with self-rated (similar to Study 1) and observer-rated extraversion. Second, the HTTP technique was only superior to simple word counting in the case of extraversion. Third, trait activation mostly improved the construct validity of extraversion (in contrast to Study 1). The more

favourable results for extraversion (compared to honesty-humility and conscientiousness) could be a consequence of the fact that extraversion is the most visible personality trait (Watson et al., 2000). This clearer visibility is supported by the fact that observer-rated extraversion showed a stronger correlation with its self-reported counterpart (again compared to honesty-humility and conscientiousness). Hence, extraversion seems to be the trait that can be somewhat better automatically assessed from interviews, especially when conditions for text analysis are optimized.

In addition to the limited evidence of construct validity, no evidence was found for the criterion-related validity of text-based and observer-rated personality ratings. HTTP conscientiousness even correlated negatively with self-reported task performance. Perhaps this effect occurred because participants occupied all sorts of different jobs, making it difficult to compare their performance levels. However, a major limitation of this study's design is the common method and source of self-reported personality and the criteria. That is, it may be questionable whether self-reported performance adequately captures actual performance. Study 3 addresses this limitation by using supervisor-rated performance as the criterion.

Study 3

In the first two studies, to evaluate the criterion-related validity of the HTTP technique, the correlation of HTTP traits with self-reported criteria were compared to those of self-reported personality with the same criteria. This approach has the limitation that self-reported predictors and criteria share variance merely because these ratings originate from the same source (i.e., the participant) and were obtained through the same method (i.e., a Likert-scale response format; e.g., Paulhus, 1991; Van Vaerenbergh & Thomas, 2012). The relation of self-reported criteria with self-reported predictors is therefore likely to be stronger than the relationship with predictor ratings from other sources or methods (i.e., text-to-personality techniques). To address the limitation of the previous two studies, Study 3 aimed to investigate the criterion-related validity of HTTP traits with a non-self-reported criterion.

When considering criteria, job performance can be thought of as "the holy grail" of organizational psychology. Indeed, a primary concern of the Society for Industrial and

Organizational Psychology (SIOP) principles for test validation is that performance on an assessment method is related to job performance (Sackett et al., 2018). The most researched (and therefore perhaps the most valued) aspect of job performance is role proficiency on an individual and team level (Carpini et al., 2017). To address the same-source problem with a relevant criterion, Study 3 was designed to obtain text-based and self-reported personality ratings, and supervisor-rated performance of employees from one organization.

Participants and Procedure

All participants worked for a large engineering company in The Netherlands. An email was sent to 461 employees, inviting them to complete the HEXACO-60 (Ashton & Lee, 2009) and a mock job interview. In total, 125 employees completed this inventory (27.1% response rate). Subsequently, 103 participants completed the asynchronous (mock) job interview. Of the 103 participants who completed the interview, 10 participants (9.7%) did not provide enough input to compute a conscientiousness score (i.e., they provided us with zero hits on the library for all interview questions) and afterwards, 22 audio files were lost due to a data management error. The final sample consisted of 71 participants who had completed all measures ($\delta = 88.7\%$, $M(\text{age}) = 38.11$ years, $SD(\text{age}) = 11.35$). Supervisor ratings of performance were obtained from the organization's database for all participants.

The Interview

Participants completed the interview in a dedicated quiet office. Upon arrival, the participants were asked to sit in front of a laptop and follow the instructions on the screen. Next, they responded to five interview questions within a maximum of two minutes. As Studies 1 and 2 showed the importance of asking trait-relevant questions, the five questions were all aimed to measure conscientiousness. Similar to Study 1, participants were encouraged to use self-descriptive words when answering the questions and some examples were provided. Most participants finished the interview in under 10 minutes. On average, participants spoke 628 words while responding to all five questions ($SD = 255$). It is worth noting that some participants were so uncomfortable answering the interview questions that they uttered one sentence or two sentences and expressed discomfort. This apparent reluctance of the participants resulted in a lack of hits for a large number of participants for at least one question response ($n = 42$). Expectation maximization imputation was completed for the purpose of the

Table 7. Study 3: Correlations between self-rated and text-analysis conscientiousness and job performance.

		Conscientiousness							
		<i>M</i>	<i>SD</i>	Age	Gender	Self-reported	HTTP	Keyword	Performance
Conscientiousness	Self-reported	3.68	0.48	.24*	.11	.77			
	HTTP	0.19	0.14	-.19	.18	.22	.52		
	Keyword count	1.71	1.95	-.02	.04	.15	.25*	.70	
Supervisor-rated performance		3.27	0.32	-.04	-.01	.35**	.24*	.06	.69
Word count		629	254	-.10	.04	-.24*	.14	-.33**	-.01

Note. $N = 72$. McDonald ω reliabilities are *italicized* on the diagonal. Hypothesized correlations are **bolded**. Scores on self-reported personality and performance from 1–5, scores for HTTP conscientiousness ranged from –1 to 1, Gender was coded as male = 1 and female = 2.

* $p < .05$, ** $p < .01$

McDonald's ω reliability analysis for HTTP conscientiousness ($\omega = .52$) and simple keyword-counted conscientiousness ($\omega = .70$).

Additional Instruments

HEXACO-PI-R

The participants completed the HEXACO-60. The ω reliabilities were mostly acceptable, ranging from .65 to .78 and for conscientiousness specifically .77.

Job performance

Individual job performance ratings were collected from the organization's database. In this particular company, job performance was evaluated on a yearly basis with a structured form that uses the same items for all employees. Employees were rated on four core values (integrity, excellence, teamwork, and health, safety & environmental) and on three performance factors (knowledge of work, quality of work, and quantity of work). These ratings appear indicative for role proficiency on individual and team level. Ratings were provided on a 5-point Likert scale, ranging from 1 = *Does not yet meet* to 5 = *Far exceeds*. For the analyses in the present study the mean proficiency rating of all items was used ($\omega = .69$).

Results and Discussion

Table 7 shows the correlations and descriptive statistics for the key variables. HTTP conscientiousness was marginally significantly correlated with self-reported conscientiousness ($r = .22$, $p = .06$), although this correlation was only marginally significant, HTTP conscientiousness correlated substantially lower with the other self-reported personality dimensions ($r = -.14$ [agreeableness] to $.08$ [emotionality]). HTTP conscientiousness also correlated significantly with supervisor-rated performance ($r = .24$, $p = .04$). However, self-reported conscientiousness showed a stronger relation with supervisor-rated performance ($r = .35$, $p < .01$) and stepwise regression confirmed that self-reported conscientiousness had incremental validity over HTTP conscientiousness ($\Delta R^2 = .10$, $p < .01$), but not vice versa ($\Delta R^2 = .03$, $p > .05$). These results align with the findings from the prior two studies – now with a criterion that was not self-reported – that self-reported personality has superior criterion-related validity over HTTP traits. In contrast to the previous studies, keyword-counted conscientiousness did not significantly relate to self-reported conscientiousness ($r = .15$) and performance ($r = .06$). However, in both cases, the correlations of HTTP conscientiousness were not significantly stronger (respectively $Z = 0.49$, $p > .05$ and $Z = 1.24$, $p > .05$).

In summary, we found marginal support for the construct and criterion-related validity of HTTP conscientiousness. One potential explanation for the marginally significant results could be the small sample size. The features of the HTTP technique seemed to add little over simple keyword counting. Lastly, self-reported conscientiousness predicted more variance in job performance than HTTP conscientiousness.

General Discussion

Text analysis, and then mainly the open-vocabulary technique, is rapidly gaining popularity for personnel selection and is contributing to the automation of this field. The appetite for these efficiency improvements seems largely driven by the advent of AVIs (Lukacik et al., 2020). However, academic knowledge is still grossly lacking about the validity of these text-analysis techniques for personnel selection.

The present study describes the development of the HTTP technique, a closed-vocabulary technique based on prior lexical research, with the purpose to assess personality traits from job interviews. The HTTP technique differs from popular open-vocabulary techniques in two important ways: it is more transparent and it is theoretically driven. The validity of the HTTP technique was tested in three different samples. This research led to four key findings. First, the HTTP technique showed small to medium-sized correlations with self-reported and observer-rated personality. Second, we found mixed, but generally favourable, evidence for the HTTP technique's criterion-related validity. However, self-reported personality was found to be more strongly predictive of the criteria. Third, the HTTP technique produced a more construct valid personality score when the interview questions activated that specific personality trait. Fourth, the HTTP technique's additional features (i.e., using factor loadings for each keyword and adjusting the keywords' loading for adjacent quantifiers) did not seem to improve its construct and criterion-related validity substantially in comparisons to a straightforward keyword-counting technique. Altogether, developing a personality text-analysis technique based on previous lexical research may have merit, but would require much more research to achieve the high standards necessary for personnel selection purposes.

Theoretical Implications

Our research contributes to the personnel selection literature in a number of ways. First, we contribute to the growing literature using text-analysis techniques to automate the assessment of job interviews (e.g., Hickman et al., 2021; Naim et al., 2018). So far, these previous studies have primarily focused on open-ended vocabulary techniques. Although initial validity evidence of such open-ended techniques is promising (e.g., L. L. Chen et al., 2016; Hickman et al., 2021; Naim et al., 2018), these techniques often lack transparency. A lack of transparency and explainability is a serious concern within the personnel selection context, as laws and regulations (e.g., Council of the European Union, 2016) grant candidates the explicit right to demand transparency on how their information is processed and require assessments to be justifiable. In an attempt to solve this dilemma, we developed a transparent, theory-driven closed-vocabulary technique (i.e., the HTTP technique) to automatically score personality traits based on job interviews and provided some modest initial evidence for its construct and criterion-related validity. We hope that the present study inspires future research and algorithm developers to use similar transparent and theory-driven approaches.

Second, we contribute to the literature on trait activation theory (Tett & Burnett, 2003; Tett et al., 2021). Previous studies already demonstrated that interview questions that activate specific traits generate more trait relevant responses (e.g., Lievens et al., 2006; Speer et al., 2015). The present study shows that asking trait-relevant questions may also affect the construct validity of text-analysis techniques. Specifically, the construct validity of HTTP emotionality (Study 1) and HTTP extraversion (Study 2) was higher when these ratings were based on responses to high trait-activating questions rather than mixed or low trait-activating questions. These findings indicate that asking trait-relevant questions make candidates verbalize more thoughts pertaining to a specific trait, and thereby use more relevant keywords, leading to more accurate estimates of the level of that trait.

Third, the results of this study brought up some considerations for the use of text-analysis in personnel selection research in general. When considering the convergence with self-ratings of personality (average $r = .41, .25, \text{ and } .22$), the construct validity of the HTTP technique is not as high as that of observer ratings (average $r = .54$). Moreover, we found lower convergence than some studies that predicted personality from job interviews with text analysis in conjunction with non-verbal and para-verbal analyses (Hickman et al., 2021; Naim et al., 2018), but not much lower than the convergence advertised by the recently discontinued IBM Watson personality insights (IBM Watson, 2019). Which beckons the same question as Woods et al. (2020) asked in their review of digital selection practices: "What are appropriate and effective benchmarks for construct validity . . . ?" (pp. 70). As we see it, efficiency improvement is the primary goal of automatic text analysis and, therefore, these solutions likely have the aim to replace or assist recruiters when making selection decisions. Therefore, we believe that the relations with observer ratings are the most appropriate construct validity target. In our second study, we found an average correlation of .26 between the HTTP technique and observer ratings. These correlations are similar to those found in similar endeavours (e.g., Hickman et al., 2021) and appear insufficient for personnel selection purposes.

Finally, in contrast to typical psychometric test evaluations, text-analysis research rarely investigates internal consistency, but instead reports accuracy/construct validity (e.g., Naim et al., 2018) and/or test-retest validity (e.g., Hickman et al., 2021). We observed relatively low reliabilities for the HTTP technique. Of course, internal consistency reliabilities can be increased by collecting more and longer interview responses. Indeed, the developers of IBM Watson recommend using at least 3500 words for a reliable measure (IBM Watson, 2017), which means that applicants would have to speak well over an hour to provide this much text (or provide 5–7 pages of writing), whereas an AVI typically lasts much shorter. Therefore, collecting longer responses might reduce the ecological validity of text-analysis experiments. In line with the approach of most text-analysis research, however, we question whether internal consistency is the appropriate metric for establishing the reliability of text-based personality ratings. Other personality measures using a similar number of items, like the brief HEXACO inventory or the HEXACO Situational Judgment Test, show comparable low internal consistency reliabilities, but still high criterion-related validities (Oostrom et al.,

2019; De Vries, 2013). Consequently, future research might like to investigate test-retest reliabilities as an alternative test of the reliabilities of text-based personality ratings.

Practical Implications

Obviously, the psychometric properties of the HTTP technique do not yet meet the psychometric standards needed for this technique to be used in actual high-stakes selection contexts. Nevertheless, the present study offers several important insights for organizations who wish to improve their assessment methods. First, organizations may want to consider using job interviews or AVIs to supplement self-report personality inventories. However, considering the present study's findings, we urge organizations to not use text-analysis techniques as a substitute of personality inventories but rather as an *additional* tool in selection and assessment batteries.

Second, when using interviews to assess personality traits, it is important to ask questions activating the traits that the organization wishes to measure. Although previous studies already showed that asking trait-activating questions enhances the construct validity of job interviews that rely on human judgement (e.g., Lievens et al., 2006; Speer et al., 2015), and Hickman et al. (2021) eluded to the effect of trait-activating questions in AVIs, the present study is the first to show that this also applies to job interviews that are automatically scored by text-analysis techniques. Using trait-activating questions seems to improve construct validity of a text-analysis tool. Although this advice may seem obvious at a surface level, in practice, text-analysis tools provide ratings regardless of the question asked, therefore, a recruiter is not directly incentivized to ask targeted questions. Generally speaking, we recommend to follow the best practices for job interviews regardless of text-analyses' ability to analyse any piece of text.

Third, we wish to re-iterate the call by Woods et al. (2020) and emphasize that there are no clear requirements for text-analysis tools' reliability and construct/criterion-related validity. Considering the level of scrutiny that is being levied at algorithmic assessment in personnel selection contexts (Liboreiro, 2021), we believe that the W/O psychology community needs to propose clear guidelines for when text-analysis solutions are effective. For these guidelines, we can largely tap into over a century of research that critically evaluated psychometric measurements, but also need to consider which kinds of reliability and validity coefficients are most appropriate and which levels are needed; what targets are most appropriate for construct validity; and finally, how overfitting (e.g., Ilievski et al., 2016) should be reduced and generalizability be improved.

Limitations

The present investigation knows several limitations. First, the interviews were not entirely representative of actual job interviews. The interviews were from either a low-stakes or simulated selection situation, but none were from a true personnel selection context. Second, for the asynchronous interviews in Study 1 and 3, prompt keywords (both positive and negative) were provided to encourage participants to use sufficient keywords. Although prompting is quite common in interviews,

prompts are generally supposed to be neutral. These prompts were provided because pilot interviews showed that some participants struggled to provide enough information (i.e., keywords) to assess their personality with text-analysis techniques. This conundrum emphasizes the need for a large keyword list when assessing personality from text, so that prompting can be restricted to a minimum while participants still speak enough keywords. Third, the results from this research may not be entirely representative of a text-to-personality assessment in practice in one more way. Our studies used high quality manually transcribed texts because the transcripts from automatic transcription software packages (in Dutch) were insufficiently accurate. Practically speaking, companies will want to use automated transcription because it is too expensive and cumbersome to manually transcribe interviews, leading to lower quality data and probably worse validity. Finally, the keywords' factor loadings had to be adjusted to be able to merge the word lists. It would have been much cleaner to use the results of one lexical study, but that would have significantly reduced the number of keywords in the HTTP technique's dictionary. More lexical studies with several word types (nouns, verbs, and adjectives) using a six-factor solution should be conducted to establish robust estimates of the loadings.

Future Research

The opportunities for future research are numerous, but we would like to highlight possible venues that could be pursued in future studies. First, when we consider our proposed approach to text analysis specifically, there are possible improvements that future research can test to address the previously mentioned limitations. For instance, to further optimize and extend the keyword list, we believe that new lexical research could contribute by including more words and increasing the amount of hits an applicant can generate. Moreover, ideally, such future lexical research would apply a work frame-of-reference (e.g., Schmit et al., 1995) to maximize its applicability to personnel assessment. Such a frame-of-reference has been shown to improve criterion-related validity (Shaffer & Postlethwaite, 2012).

Second, more research is needed on the criterion-related validity of the HTTP technique for predicting other-rated and objective work outcomes. As Woods et al. (2020) noted, when assessing the criterion-related validity of new techniques, it is important to examine what this technique contributes in terms of validity as part of a complete selection procedure, which often consists of a series of tests, questionnaires, and interviews. Furthermore, as text-analysis techniques rely on the use of particular keywords, differences in language proficiency and word use might affect their validity. Hence, future studies should also look at differential prediction based on ethnicity and socioeconomic class.

Third, most tools that automatically analyse interviews are multimodal, meaning that they assess verbal, prosodic, and non-verbal behaviour simultaneously. Some other studies have already combined text analysis with other features. Although these studies are appealing because they illustrate how existing solutions might work, we believe that

much more work is required on the separate modalities in a job interview setting, because each modality is likely to have different challenges that are specifically relevant for personnel selection. For example, we showed that trait-activation improves the validity of text analysis, but we would expect trait-activation to have a less strong effect on non-verbal behaviour. In contrast, adverse impact might be a greater concern for the prosodic and non-verbal modality. We therefore recommend future research to study each modality in detail instead directly focussing on the sum of all modalities.

Last, once the validity of text-to-personality techniques has been optimized further, it is not clear how well these assessments will be received by applicants. Interviews are typically a human affair and automated scoring of responses might not be well-received. Already, Langer et al. (2019) found that participants generally prefer videoconference interviews with a human interviewer over highly automated interview platforms with an avatar. It remains unknown how receptive applicants are to text-to-personality assessments by computers. Future research could investigate if applicants react differently to personality assessments based on surveys, interviewers, and text-analysis programs.

Conclusion

We view this study as an illustration of the potential value of new technology for personnel selection. The results show that closed-vocabulary text-analysis techniques can be used to assess traces of personality in interview transcripts and that it can predict socioemotional and performance outcomes. Additionally, the quality of the text-to-personality ratings improves when interview questions activate the personality trait under investigation. Still, significant improvements are needed before automatically computed text-to-personality ratings can be used to replace, or supplement, self-report inventories or interviewer ratings.

Notes

1. Initially, we intended to use speech-to-text software to automatically generate texts from the voice recordings and then feed these texts to the text-to-personality technique. However, the accuracy of several speech-to-text software packages in the Dutch language proved disappointing. Regardless of the audio quality, all speech-to-text programs made an unacceptable number of mistakes. It was thus decided to transcribe all spoken text manually.
2. This project's Open Science Framework (OSF) webpage can be found on <https://osf.io/w76px/>
3. For a detailed list of research on the HEXACO model and inventories, see hexaco.org.

Author contributions

Study conception, design, material preparation, and data collection: D. Holtrop, J. K. Oostrom, W. R. J. van Breda, and R. E. de Vries. Analyses were performed by D. Holtrop and W. R. J. van Breda. The first draft of the manuscript was written by D. Holtrop and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of the Vrije Universiteit Amsterdam (file number: VCWE-2017-046).

Open Science

The materials for this study are available on its Open Science Framework webpage <https://osf.io/w76px/>
DOI10.17605/OSF.IO/W76PX

Prior versions

A very early and partial version of this work was presented at annual Work and Organizational Psychology (WAOP) conference, Amsterdam, The Netherlands (25 November, 2016).

Disclosure statement

Ward R. J. van Breda is the owner of NeedForward, a commercial entity that developed and owns the grammar engine used for the text-analyses in this manuscript. The other authors have no affiliation with this entity.

Funding

Stichting Abbas Fonds, a Dutch organization that aims to stimulate test development and research into psychometric tests in The Netherlands and Flanders, financially supported Study 2.

ORCID

Djurje Holtrop  <http://orcid.org/0000-0003-3824-3385>
Janneke K. Oostrom  <http://orcid.org/0000-0002-0963-5016>
Ward R. J. van Breda  <http://orcid.org/0000-0002-7358-4444>
Antonis Koutsoumpis  <http://orcid.org/0000-0001-9242-4959>
Reinout E. de Vries  <http://orcid.org/0000-0002-4252-5839>

References

- Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, 47(1), i–171. <https://doi.org/10.1037/h0093360>
- Anglim, J., Morse, G., De Vries, R. E., MacCann, C., & Marty, A. (2017). Comparing job applicants to non-applicants using an item-level bifactor model on the HEXACO personality inventory. *European Journal of Personality*, 31(6), 669–684. <https://doi.org/10.1002/per.2120>
- Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91(4), 340–345. <https://doi.org/10.1080/00223890902935878>
- Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., Boies, K., & De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, 86(2), 356–366. <https://doi.org/10.1037/0022-3514.86.2.356>
- Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology*, 33(4), 445–459. <https://doi.org/10.1007/s10869-017-9528-3>
- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1), 1–26. <https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>
- Barrick, M. R., Patton, G. K., & Haugland, S. N. (2000). Accuracy of interviewer judgments of job applicant personality traits. *Personnel Psychology*, 53(4), 925–951. <https://doi.org/10.1111/j.1744-6570.2000.tb02424.x>
- Blackman, M. C. (2002). Personality judgment and the utility of the unstructured employment interview. *Basic and Applied Social Psychology*, 24(3), 241–250. https://doi.org/10.1207/S15324834BASP2403_6
- Carpini, J. A., Parker, S. K., & Griffin, M. A. (2017). A look back and a leap forward: A review and synthesis of the individual work performance literature. *Academy of Management Annals*, 11(2), 825–885. <https://doi.org/10.5465/annals.2015.0151>
- Chen, L., Feng, G., Leong, C. W., Lehman, B., Martin-Raugh, M., Kell, H., Lee, C. M., & Yoon, S.-Y. (2016). Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm. *18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan.
- Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods*, 4(1), 62–83. <https://doi.org/10.1177/109442810141004>
- Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review*, 31(1). <https://doi.org/10.1016/j.hrmr.2019.100698>
- Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance*, 18(2), 123–149. https://doi.org/10.1207/s15327043hup1802_2
- Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, 24(4), 385–396. <https://doi.org/10.2307/2136404>
- Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, 80(5), 565–579. <https://doi.org/10.1037/0021-9010.80.5.565>
- Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology*, 53(2), 325–351. <https://doi.org/10.1111/j.1744-6570.2000.tb00204.x>
- Council of the European Union. (2016). *The European General Data Protection Regulation* <https://gdpr-info.eu/>
- De Cuyper, K., De Houwer, J., Vansteelandt, K., Perugini, M., Pieters, G., Claes, L., & Hermans, D. (2017). Using indirect measurement tasks to assess the self-concept of personality: A systematic review and meta-analyses. *European Journal of Personality*, 31(1), 8–41. <https://doi.org/10.1002/per.2092>
- De Raad, B., & Barelds, D. P. (2008). A new taxonomy of Dutch personality traits based on a comprehensive and unrestricted list of descriptors. *Journal of Personality and Social Psychology*, 94(2), 347–364. <https://doi.org/10.1037/0022-3514.94.2.347>
- De Raad, B., & Hoskens, M. (1990). Personality-descriptive nouns. *European Journal of Personality*, 4(2), 131–146. <https://doi.org/10.1002/per.2410040206>
- De Vries, R. E. (2013). The 24-item brief HEXACO inventory (BHI). *Journal of Research in Personality*, 47(6), 871–880. <https://doi.org/10.1016/j.jrp.2013.09.003>
- De Vries, R. E., Ashton, M. C., & Lee, K. (2009). De zes belangrijkste persoonlijkheidsdimensies en de HEXACO Persoonlijkheidsvragenlijst. *Gedrag & Organisatie*, 22(3), 232–274. <https://doi.org/10.5117/2009.022.003.004>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. <https://doi.org/10.48550/arXiv.1810.04805v2>
- Dipboye, R. L., Macan, T., & Shahani-Denning, C. (2012). *The selection interview from the interviewer and applicant perspectives: Can't have one without the other*. Oxford University Press.
- Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399–412. <https://doi.org/10.1111/bjop.12046>
- Ebstrup, J. F., Eplöv, L. F., Pisinger, C., & Jørgensen, T. (2011). Association between the Five Factor personality traits and perceived stress: Is the effect mediated by general self-efficacy? *Anxiety, Stress, and Coping*, 24(4), 407–419. <https://doi.org/10.1080/10615806.2010.540012>
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2009). COTAN beoordelings-systeem voor de kwaliteit van tests (COTAN evaluation system for the quality of tests). NIP.

- Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102(4), 652–670. <https://doi.org/10.1037/0033-295X.102.4.652>
- Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, 52(2), 409–418. <https://doi.org/10.1037/0022-3514.52.2.409>
- Goldberg, L. R. (1990). An alternative “description of personality”: The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216–1229. <https://doi.org/10.1037/0022-3514.59.6.1216>
- Graves, L. M., & Powell, G. N. (1996). Sex similarity, quality of the employment interview and recruiters’ evaluation of actual applicants. *Journal of Occupational and Organizational Psychology*, 69(3), 243–261. <https://doi.org/10.1111/j.2044-8325.1996.tb00613.x>
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30. <https://doi.org/10.1037/1040-3590.12.1.19>
- Hall, J. A., Goh, J. X., Mast, M. S., & Hagedorn, C. (2016). Individual differences in accurately judging personality from text. *Journal of Personality*, 84(4), 433–445. <https://doi.org/10.1111/jopy.12170>
- Harwell, D. (2019). *Rights group files federal complaint against HireVue, a hiring company that uses artificial intelligence*. Seattle times. <https://www.seattletimes.com/business/rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-and-deceptive-practices/>
- Heimann, A. L., Ingold, P. V., Debus, M. E., & Kleinmann, M. (2020). Who will go the extra mile? Selecting organizational citizens with a personality-based structured job interview. *Journal of Business and Psychology*, 36, 1–23. <https://doi.org/10.1007/s10869-020-09715-2>
- Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated Video Interview Personality Assessments: Reliability, Validity, and Generalizability Investigations. *Journal of Applied Psychology*, Preprint. <https://doi.org/10.1037/apl0000695>
- Hickman, L., Tay, L., & Woo, S. E. (2019). Validity evidence for off-the-shelf language-based personality assessment using video interviews: Convergent and discriminant relationships with self and observer ratings. *Personnel Assessment and Decisions*, 5(3), 12–20. <https://doi.org/10.25035/pad.2019.03.003>
- Huffcutt, A. I. (2011). An empirical review of the employment interview construct literature. *International Journal of Selection and Assessment*, 19(1), 62–81. <https://doi.org/10.1111/j.1468-2389.2010.00535.x>
- Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, 86(5), 897–913. <https://doi.org/10.1037/0021-9010.86.5.897>
- IBM Watson. (2017). *Watson Personality Insights Exercise*. <https://tinyurl.com/y5tmsn7j>
- IBM Watson. (2019). <https://cloud.ibm.com/docs/services/personality-insights?topic=personality-insights-models>
- Ilievski, F., Postma, M., & Vossen, P. (2016). *Semantic overfitting: What ‘world’ do we consider when evaluating disambiguation of text?* *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan.
- Judge, T. A., Erez, A., & Bono, J. E. (1998). The power of being positive: The relation between positive self-concept and job performance. *Human Performance*, 11(2–3), 167–187. <https://doi.org/10.1080/08959285.1998.9668030>
- Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research—Moving away from the “What” towards the “Why”. *International Journal of Information Management*, 54. <https://doi.org/10.1016/j.ijinfomgt.2020.102205>
- Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. *Personnel Psychology*, 34(2), 263–289. <https://doi.org/10.1111/j.1744-6570.1981.tb00942.x>
- Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, 98(6), 1060–1072. <https://doi.org/10.1037/a0034156>
- Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, 27(3), 217–234. <https://doi.org/10.1111/ijsa.12246>
- Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, 39(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8
- Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO Personality Inventory: Two new facet scales and an observer report form. *Psychological Assessment*, 18(2), 182–191. <https://doi.org/10.1037/1040-3590.18.2.182>
- Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, 25(5), 543–556. <https://doi.org/10.1177/1073191116659134>
- Lee, K., Ashton, M. C., Pozzebon, J. A., Visser, B. A., Bourdage, J. S., & Ogunfowora, B. (2009). Similarity and assumed similarity in personality reports of well-acquainted persons. *Journal of Personality and Social Psychology*, 96(2), 460–472. <https://doi.org/10.1037/a0014059>
- Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, 67(1), 241–293. <https://doi.org/10.1111/peps.12052>
- Liboreiro, J. (2021). Euronews. <https://www.euronews.com/2021/04/21/the-higher-the-risk-the-stricter-the-rule-brussels-new-draft-rules-on-artificial-intelligence>
- Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven (Eds.), *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 197–253). Springer.
- Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology*, 91(2), 247–258. <https://doi.org/10.1037/0021-9010.91.2.247>
- Lukacik, E.-R., Bourdage, J. S., & Roulin, N. (2020). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, 32(1), 100789. <https://doi.org/10.1016/j.hrmmr.2020.100789>
- Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30, 457–500. <https://doi.org/10.1613/jair.2349>
- Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5), 862–877. <https://doi.org/10.1037/0022-3514.90.5.862>
- Melchers, K. G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers’ rating quality. *Personnel Psychology*, 64(1), 53–87. <https://doi.org/10.1111/j.1744-6570.2010.01202.x>
- Mirowska, A. (2020). AI Evaluation in Selection: Effects on Application and Pursuit Intentions. *Journal of Personnel Psychology*, 19(3), 142–149. <https://doi.org/10.1027/1866-5888/a000258>
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, 60(3), 683–729. <https://doi.org/10.1111/j.1744-6570.2007.00089.x>
- Muralidhar, S., Nguyen, L., & Gatica-Perez, D. (2018). *Words worth: Verbal content and hirability impressions in YouTube video resumes*. 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Brussels, Belgium.
- Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2018). Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, 9(2), 191–204. <https://doi.org/10.1109/TAFFC.2016.2614299>

- Oh, I.-S., Le, H., Whitman, D. S., Kim, K., Yoo, T.-Y., Hwang, J.-O., & Kim, C.-S. (2014). The incremental validity of honesty-humility over cognitive ability and the big five personality traits. *Human Performance*, 27(3), 206–224. <https://doi.org/10.1080/08959285.2014.913594>
- Oostrom, J. K., De Vries, R. E., & De Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance*, 32(1), 1–29. <https://doi.org/10.1080/08959285.2018.1539856>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* <https://doi.org/10.1109/3468.844354>
- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6), 934–952. <https://doi.org/10.1037/pspp0000020>
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Academic Press.
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). The Guilford Press.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.
- Pletzer, J. L., Bentvelzen, M., Oostrom, J. K., & De Vries, R. E. (2019). A meta-analysis of the relations between personality and workplace deviance: Big Five versus HEXACO. *Journal of Vocational Behavior*, 112, 369–383. <https://doi.org/10.1016/j.jvb.2019.04.004>
- Sackett, P. R., Tippins, N. T., Arthur, W., Jr, Delany, T., Dunleavy, E. M., Hayes, T. L., Hough, L., Oswald, F., Putka, D. J., Ryan, A. M., & Schmitt, N. (2018). *Principles for the Validation and Use of Personnel Selection Procedures*. American Psychological Association. <https://www.apa.org/ed/accreditation/about/policies/personnel-selection-procedures.pdf>
- Sackett, P. R., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace? *Perspectives on Psychological Science*, 9(5), 538–551. <https://doi.org/10.1177/1745691614543972>
- Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology*, 82(1), 30–43. <https://doi.org/10.1037/0021-9010.82.1.30>
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262–274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology*, 80(5), 607. <https://doi.org/10.1037/0021-9010.80.5.607>
- Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. *Annu. Rev. Organ. Psychol. Organ. Behav.*, 1(1), 45–65. <https://doi.org/10.1146/annurev-orgpsych-031413-091255>
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., & Seligman, M. E. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, 8(9), e73791. <https://doi.org/10.1371/journal.pone.0073791>
- Sentimentics. (2019). Retrieved 3-12-2019 from <http://sentimentics.com/>
- Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, 65(3), 445–494. <https://doi.org/10.1111/j.1744-6570.2012.01250.x>
- Speer, A. B., Christiansen, N., & Honts, C. (2015). Assessment of personality through behavioral observations in work simulations. *Personnel Assessment and Decisions*, 1(1), 43–56. <https://doi.org/10.25035/pad.2015.006>
- Tausczik, Y. R., & Pennebaker, J. W. (2011). *Predicting the perceived quality of online mathematics contributions from users' reputations*. CHI 2011: Conference on Human Factors in Computing Systems, Vancouver, BC, Canada.
- Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, 88(3), 500–517. <https://doi.org/10.1037/0021-9010.88.3.500>
- Tett, R. P., Toich, M. J., & Ozkum, S. B. (2021). Trait Activation Theory: A Review of the Literature and Applications to Five Lines of Personality Dynamics Research. *Annual Review of Organizational Psychology and Organizational Behavior*, 8(1), 199–233. <https://doi.org/10.1146/annurev-orgpsych-012420-062228>
- Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing personality with a structured employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology*, 90(3), 536–552. <https://doi.org/10.1037/0021-9010.90.3.536>
- Van Iddekinge, C. H., Sager, C. E., Burnfield, J. L., & Heffner, T. S. (2006). The variability of criterion-related validity estimates among interviewers and interview panels. *International Journal of Selection and Assessment*, 14(3), 193–205. <https://doi.org/10.1111/j.1468-2389.2006.00352.x>
- Van Vaerenbergh, Y., & Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, 25(2), 195–217. <https://doi.org/10.1093/ijpor/eds021>
- Watson, D., Hubbard, B., & Wiese, D. (2000). Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, 78(3), 546–558. <https://doi.org/10.1037/0022-3514.78.3.546>
- Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, 29(1), 64–77. <https://doi.org/10.1080/1359432X.2019.1681401>
- Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, 10(2), 109–120. [https://doi.org/10.1016/0022-1031\(74\)90059-6](https://doi.org/10.1016/0022-1031(74)90059-6)
- Zettler, I., Thielmann, I., Hilbig, B. E., & Moshagen, M. (2019). The nomological net of the HEXACO model of personality: A large-scale meta-analytic investigation. *Perspectives on Psychological Science*, 15(3), 723–760. <https://doi.org/10.1177/1745691619895036>