Incremental Validity of Language-Based Assessments of Personality in World Trade Center Responders

Joshua R. Oltmanns, Ph.D.[1], H. Andrew Schwartz, Ph.D.[2], Camilo Ruggero, Ph.D.[3], Youngseo Son, B.S.[2], Huy Vu, M.S.[2], Gilvir Gill, B.S.[2], Jiaju Miao, M.A.[2], Monika Waszczuk, Ph.D.[4], Sean A. P. Clouston, Ph.D.[2], Evelyn J. Bromet, Ph.D.[2], Benjamin J. Luft, M.D.[2], and Roman Kotov, Ph.D.[2]

[1] Southern Methodist University, [2]Stony Brook University, [3]University of North Texas, [4]Rosalind Franklin University

This research was supported by the National Institutes of Occupational Safety and Health under Award Number U01OH011321 (PI: Roman Kotov). NIOSH had no role in the conduct of the study or preparation of the manuscript. The findings and conclusions in this article are those of the authors and do not represent the official positions of NIOSH.

Correspondence should be addressed to Joshua R. Oltmanns, 1315 Expressway Tower, Department of Psychology, Southern Methodist University, Dallas, TX, 75206. Email: jroltmanns@smu.edu.

Abstract

The predictive utility of language-based assessments (LBAs) of personality was tested across one year. Spoken language in everyday functioning interviews with 343 World Trade Center (WTC) responders was recorded and transcribed. LBAs that were previously developed using social media text were adapted to the interview transcriptions, resulting in eight LBAs: neuroticism, extraversion, agreeableness, openness, conscientiousness, depressiveness, anxiousness, and anger proneness. After one year, responders were assessed for physical and mental health problems via self-report questionnaire as well as cognitive ability via processing speed test and mental healthcare use via clinic data. LBAs predicted outcomes moderately, over functioning scores, and predicted increases in negative outcomes over the outcome variables. The neuroticism LBA emerged as the most robust individual predictor.

Keywords: personality, language, five-factor model, world trade center responders

Incremental Validity of Language-Based Assessments of Personality in World Trade Center Responders

Machine-learning based technologies have rapidly developed and proliferated in the 21st century. New research in psychiatry and psychology shows that artificial intelligence (AI) might be used to improve mental health diagnosis, prognosis, and treatment prediction (Koutsouleris et al., 2022). However, researchers and clinicians have not yet harnessed this technology. Existing assessment techniques often rely on patient-self report, which has unique biases such as over and underreporting and limited insight into psychological problems (McGrath et al., 2010). Further, clinical psychological assessment requires lengthy periods of time, sometimes consisting of several sessions. The development of AI has the potential to improve assessment through use of more objective behavioral markers such as language. This could, in turn, increase validity and scalability while simultaneously reducing resource demands required on the part of clinicians and patients alike.

Research has demonstrated that techniques can be used to assess personality with natural language. Words and parts of speech hold psychological insight into emotions, personality, situations, and cognitive styles, as well as the power to discriminate between groups of patients and controls (Pennebaker et al., 2003). In more recent years, advanced machine learning-based techniques have improved accuracy of these methods (Eichstaedt et al., 2020). For example, personality language models trained on text from on social media "status updates" (i.e., a brief social media post that tells friends what one is doing or says what is on the mind) have demonstrated test-retest reliability and convergent, discriminant, and predictive validity correlations with traditional psychological measures (Merchant et al., 2019; Park et al., 2015; Schwartz et al., 2014). These models were developed via machine learning from very large

samples of brief textual status updates (i.e., millions of posts). With the power of increased computational modeling, in addition to single words, (a) multi-word "phrases" and (b) multi-word/phrase "topics" can be extracted and used to improve language models that capture personality (Eichstaedt et al., 2020). Mining of electronic health records has also been used to provide large samples of textual language to assess mental health (Bittar et al., 2020; Boag et al., 2021). However, to date few studies have applied these techniques to spoken language, especially with patients, which provides unique challenges to obtain, but exciting opportunities for improving language models used to capture psychiatric constructs.

Several studies have used language samples from patient populations to construct and test language models of personality and related psychopathology. One study trained language models to recognize depression in a sample of 7,845 general community adults and tested in 600 older adults (Rutowski et al., 2021). Language content was obtained through conversation with an application asking questions about home life. They found an area under the curve of .82 and relatively little degradation from one sample to the other. While this study was important due to its size and collection of spoken language data, limitations included that the participants interacted with an application, rather than a human, and the models were trained as dichotomous classifiers only for depression from the Patient Health Questionnaire-9.

Another study collected open-ended interviews with 81 patients admitted to an emergency room one month after they had experienced a trauma (Schultebraucks et al., 2020). Natural language was examined using the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007) and included in multi-modal models also with speech acoustics and facial features to classify PTSD and MDD diagnoses. LIWC features such as self-assured, interrogatives, workhouse, organized, and sexual focus contributed to the classification of PTSD

and MDD. This study is important in its use of clinical interviews to assess language and its use of a clinical sample. Limitations of this study include the relatively small sample size and reliance on LIWC to assess language content.

These studies trained models in relatively small samples, which limits their accuracy and generalizability. An alternative strategy is to *adapt* models developed in large samples of written language to interview language, adjusting for words that appear in one language type but not another. In particular, Park et al. (2015) developed algorithms for assessing personality from text. They used Facebook status updates from 66,732 users to train and a separate sample of 4,800 users to test the validity. They found good convergence with self-reports of personality. Eichstaedt et al. (2020) found stronger convergent validity with self-report measures of personality from the Language Based Assessments (LBAs) compared to LIWC, as well as additional criterion validity support and test-retest reliability.

We have previously applied LBAs from these large social media derivation samples to language used in voicemail updates and oral histories from World Trade Center (WTC) responders (Oltmanns et al., 2021; Son et al., 2021). In 124 WTC responders, Son et al. (2021) showed predictive validity of LBAs for PTSD. In a different sample of 174 WTC responders, Oltmanns et al. (2021) showed convergent and discriminant validity of LBAs for depressiveness, anxiousness, anger proneness, and the FFM personality domains with self-report measures of personality and psychopathology. Oltmanns et al. (2021) also showed predictive validity for trauma-related outcomes including symptoms of depression, posttraumatic stress disorder, sleep disturbance, respiratory problems, and GERD. Limitations of Son et al. (2021) include the relatively small sample size, examining only one outcome (PTSD), and only four LBAs (neuroticism, extraversion, depressiveness and anxiousness). Limitations of Oltmanns et al.

(2021) include still relatively small sample size and daily voicemail updates rather than clinical interview language. Neither study examined objective outcome validators.

The present study advances the validation of language-based scoring of personality by adapting social media status update-derived LBAs to a sample of $N = 343$ WTC responders using language from a structured clinical interview for functional impairment. Further, the predictive validity of the LBAs is examined for outcomes measured one year later, including objective measures of cognitive ability and mental healthcare costs, which have not been previously examined in relation to the LBAs.

Method

Procedure

Participants in this study were recruited for WTC Personality and Health Study (Waszczuk et al., 2019) from the Stony Brook site of the WTC Health Program (Dasaro et al., 2017). The National Institute for Occupational Safety and Health established the WTC Health Program to monitor the physical and mental health of responders to the 9/11 WTC disaster. Participants were required to have been on the disaster site for rescue or spent significant time during the clean-up effort. Exclusionary criteria consisted of limited comprehension of English or major cognitive impairment.

The current study uses data from the second and third assessment points of the WTC Personality and Health Study. At the second assessment point (baseline for the present study), participants were interviewed for impairment in everyday functioning, in addition to completing a self-report questionnaire battery and a cognitive ability test. At the third assessment point one year later (wave 2 for the present study), outcomes were re-assessed. Language transcription, feature extraction, and analysis are described below.

The study was approved by the local Institutional Review Board and all participants provided informed consent.

Participants

Participants were 343 WTC responders. On average, responders were 56.5 years old (*SD* = 8.6 years). Participants identified as 89% male and 90% white (7.3% Black, 1.2% Asian, and 1.2% other). 7.1% identified Hispanic ethnicity. Most of the responders worked as law enforcement on 9/11 (65%), while others worked as construction workers, electricians, and paramedics. The sample was demographically similar to the overall patient population at the Stony Brook site of the WTC Health Program (Bromet et al., 2016).

Measures

Functioning Interview. At baseline, responders completed the Range of Impaired Functioning (RIFT) interview, which assesses functioning in six domains over the past month: family, social network, friends, household duties, recreation, and life satisfaction (Leon et al., 1999). Areas were probed with multiple semi-structured follow-up questions. As an example, the employment area is probed with questions regarding the number of hours worked, sick days taken, and satisfaction with work. Functioning was rated in consensus meetings of at least three interviewers on a five-point scale with labels 1 (severe impairment), 2 (moderate impairment), 3 (mild impairment), 4 (no impairment, satisfactory level), and 5 (no impairment, high level). The RIFT total score is the mean of ratings for the six domains.

LBAs. Audio was transcribed using *TranscribeMe*, a HIPAA-conformant transcription service. Speakers were identified and language of interviewer was removed to focus on the interviewee. Linguistic features were extracted using the Python package Differential Language Analysis Toolkit (DLATK) (Schwartz et al., 2017).

We adapted pretrained LBA models from social media (Park et al., 2015) to spoken language. Words were removed based on differences in the amount of people that used a word. Words with low prevalence in RIFT transcriptions (e.g., emojis) were removed from the Facebook status updates (Rieman et al., 2017), and the models were retrained before application to the RIFT transcriptions. Words that were absent in Facebook updates (e.g., dysfluencies such as "uh" and "um") were removed from transcripts before application of the retrained models. In particular, two dimensions were used to examine differences between social media and RIFT language: 1) the percentage of people that used the word at all, and 2) the frequency of usage among those who used it. Words were selected for inclusion based on having distributional estimates for both the discrete indicator and continuous variables that matched the two target distributions of usage. Words that have a different distribution were assumed to likely take on a different meaning. For example, "sick" in social media is more often used in a positive way, while it is used more often in a negative way in the WTC responders' language. Age was used as a covariate in the models.

Personality. At baseline, responders completed the Faceted Inventory of the Five-Factor Model (FI-FFM) to assess neuroticism, extraversion, agreeableness, and conscientiousness and their facets (Watson et al., 2019). The Big Five Inventory-2 was used to assess openness (Soto & John, 2017). Items for these questionnaires were rated on a Likert scale from 1 (disagree strongly) to 5 (agree strongly).

Health Measures at baseline and wave 2. The Trail Making Test Part A is a widely used neuropsychological test of simple attention and processing speed (Bowie & Harvey, 2006). Participants used a pencil to connect a series of scattered numbers in numerical order. The key

outcome is time. The TMT Part A has been associated with motor speed and intelligence (Bowie & Harvey, 2006).

Total cost of mental healthcare visits per responder at the WTC Health Program were computed from electronic health records. This includes costs of therapy and mental health medication for treatment related to mental health problems resulting from exposure to the 9/11 disaster that the responders received from the CDC-funded WTC Health Program—Long Island site. To adjust positive skewness of the variable, it was winsorized at the 97% percentile.

The PTSD Checklist for DSM-5 (PCL-5) was used to assess PTSD symptoms in the past month (Weathers et al., 2013). The PCL-5 consists of 20 items rated from 1 (not at all) to 5 (extremely). General Depression, Suicidality, and Well-Being scales from the Inventory of Depression and Anxiety Symptoms, expanded version (IDAS-II) were administered to the responders (Watson et al., 2012). Ratings were made of the past two weeks on a Likert scale from 1 (not at all) to 5 (extremely).

Lower respiratory symptoms (LRS) were assessed over the past one week with a six-item questionnaire rated from 1 (none) to 5 (6 or 7 days of the week) (Waszczuk et al., 2017). An example item is, “How often did your chest feel tight?” GERD symptoms were assessed over the past one week with the Reflux Disease Questionnaire (RDQ) (Shaw et al., 2001). Six items were rated for severity from 1 (did not have) to 6 (severe). An example RDQ item is, “A pain in the center of the upper stomach.” Pain symptoms were measured over the past seven days with four items rated from 1 (not at all to 5 (very much) from the Pain Interference scale (Askew et al., 2016). The items assessed pain interference with day-to-day activities, work around the home, participation in social activities, and household chores. During the course of two weeks after the in-person assessment, responders completed daily diaries that included sleep quality assessed

each morning with items based on the Pittsburgh Assessment Conference sleep diary (Natale et al., 2015). Responders rated sleep quality from 1 (very poor) to 5 (very good).

Analyses

Correlations were used to examine the convergent and discriminant relationships among the variables. Hierarchical multiple regression analysis was used to examine predictive validity of the LBAs. In one set of hierarchical regressions, wave 2 outcomes (e.g., depression, suicidality, PTSD) were predicted in two steps: the predictive effect of the RIFT total was examined in step 1, and the 8 LBA scores were entered into step 2. In the second set of hierarchical regressions, wave 2 outcomes were predicted again in two steps: the outcome variable at baseline (rather than RIFT total) entered step 1, and then the 8 LBAs were entered in step 2. Thus, in the first set of hierarchical regressions we examined the predictive utility of the LBAs over and above the baseline RIFT ratings by interviewers made from the same session as the LBAs. And in the second set of analyses, we examined the ability of the LBAs to predict change in the outcome variables across one year. Missing data were imputed with ipsative mean imputation if less than 20% of a scale was missing. Data are available from the last author.

Results

Descriptives for the study variables are presented in Supplemental Table 1. The median absolute value intercorrelation among the LBAs was $r = .19$, indicating that the LBAs were weakly intercorrelated. Similarly, the intercorrelations between the LBAs and their corresponding self-report scales ranged from $r = .09$ (anger proneness) to $r = .27$ (conscientiousness), with a median of $r = .21$ (Table 1). Correlations between the LBAs and the outcomes at both waves are presented in Table 2. LBAs correlated significantly with most of the outcomes at baseline and wave 2. The neuroticism LBA’s median absolute value correlation with

the outcomes across both waves was $r = .23$, with a maximum correlation of $r = .32$ (wave 2 IDAS Depression). Across the significant relationships with wave 2 mental health markers general depression, suicidality, and PTSD symptoms, the neuroticism LBA's median *r* was .30. In comparison, the median *r* between the RIFT score itself and those outcomes was .42. Across the physical health outcomes lower respiratory symptoms, GERD symptoms, and pain, the neuroticism LBA's median *r* was .26. In comparison, the median *r* between the RIFT score and those outcomes was .30. The correlation between the neuroticism LBA and mental healthcare costs was $r = .17$, while the same correlation for the RIFT was $r = .30$.

Table 3 displays results of the first set of hierarchical regression analyses, which were the LBAs predicting wave 2 outcome variables over and above baseline total RIFT scores. The results indicate that the LBAs provided significant incremental variance to the prediction of seven of the eleven outcomes over and above the RIFT scores. At the level of individual predictors, the Neuroticism LBA significantly predicted six of the outcomes (general depression, suicidality, PTSD, LRS, GERD symptoms, and pain), ranging from $\beta = .19$ to $\beta = .29$, with a median of $\beta = .23$ (all *p* values < .01) (displayed in Table 5). The Neuroticism LBA did not significantly predict mental healthcare costs in the model over and above the RIFT and other LBAs. These results indicate that personality and psychopathology LBAs scored from language used in the RIFT interview predicted seven outcomes over and above ratings made by RIFT interviewers.

Table 4 displays results of the second set of hierarchical regression analyses, which were the LBAs predicting wave 2 outcome variables over and above baseline outcome scores—thus predicting change in the outcomes. The results indicate that the LBAs provided significant incremental variance to the prediction of change in six of the eleven outcomes. At the level of

individual predictors, the Neuroticism LBA predicted general depression β = .13, p = .008, and the Neuroticism LBA and Anger Proneness LBAs predicted PTSD symptoms β = .11 and .13, respectively (p = .009 and p = .007) (Table 5). This indicates that personality and psychopathology LBAs scored from clinical language used in the RIFT interview predicted worsening of both physical and mental health outcomes across one year—in particular, that for neuroticism.

Discussion

Increasing evidence suggests that AI can be used to assess psychiatric constructs, with language markers being one promising area (Koutsouleris et al., 2022). However, much research relies solely on written social media language and only a few studies use objective behavioral markers to validate the AI. The present study extends prior research by adapting language models developed from social media to spoken clinical interview language in a larger clinical sample. We examined predictive utility for several longitudinal health outcomes including EMA assessments of sleep quality and objective measurements of cognitive ability and mental healthcare costs. Results indicate that the LBAs have incremental validity for the prediction of several outcomes across time over and above the RIFT interview score and the outcomes themselves. In particular, the Neuroticism LBA had relatively strong incremental validity for predicting the outcomes.

The results of this study are important for four reasons: 1) the convergent validity findings indicate that language use by interviewees can be scored for personality. While this has been demonstrated previously, the present findings are especially important because the algorithms used to assess personality from the interviews here were developed on written social media status updates and then applied to spoken clinical interview language. This indicates that

models developed on social media text can translate to spoken clinical language. However, results also indicate that there is potential for improvement in that algorithms trained on clinical interview language in large samples may provide even more promising results. 2) The findings indicate that language used in clinical interviews can provide predictive validity for important physical and mental health outcomes *over and above ratings made by interviewers*. That is, the language used to convey answers to clinical assessment contains predictive validity information over and above the answers coded by raters. This advances the idea that meaningful nuanced language indicators can be captured by AI and harnessed by clinicians and researchers to improve treatment and research. 3) The findings indicate that LBAs can predict increases in problem levels of relevant treatment outcomes over and above baseline levels of these outcomes, consistent with findings in our prior work using language scored for personality using voice messages (Oltmanns et al., 2021) and open-ended interviews (Son et al., 2021). In the present study, LBAs significantly predicted increases in PTSD, GERD, depressive symptoms, cognitive ability, suicidality, and pain over corresponding baseline symptoms. Importantly, these are some of the most common negative health outcomes that WTC responders confront due to exposure to the 9/11 attacks. 4) The present study added to our prior work the test of LBAs for predicting objective outcomes. Results of hierarchical regression indicate that the LBAs predicted increases in total mental healthcare dollars spent on a responder across one year. Further, the results provide supplemental evidence of the predictive ability of the LBAs for the objective Trail Making Test of cognitive ability over and above the scoring of the RIFT interview itself.

Effect sizes of the relations between AI-based LBAs and outcomes approached standards for a “moderate” magnitude (Cohen, 1992). In zero-order correlations, effect sizes approached 10% of the variance, or $r = .30$, which is comparable to predictive validity of self-report measures of

personality (Funder & Ozer, 2019; Soto, 2019). In hierarchical regressions controlling for the RIFT interview score itself, the incremental variance accounted for by the LBAs also approached moderate effect sizes (maximizing at change in R^2 = 9% for general depression, PTSD symptoms, and pain). The ability to predict important outcomes such as these at a moderate effect size, even over and above the RIFT interview itself, demonstrates the LBAs as valuable contributors to assessment. This is especially impressive considering that there are few variables, in general, that predict these outcomes at a moderate effect size. LBA prediction of moderate variance is an exciting and promising result.

Effect sizes for the LBA prediction of *change* in the outcomes was somewhat smaller (maximizing at change in R^2 = 4% for the Trail Making Test, suicidality, and GERD symptoms). This is to be expected, as change in an outcome is difficult to predict—especially during a relatively brief one-year follow-up period. Thus, the ability of the LBAs to predict these changes is notable and further demonstrates criterion validity of the LBAs for outcomes over and above the RIFT interview questions themselves and *also* over and above autoregressive effects of the outcomes predicting themselves.

All LBAs significantly predicted the outcomes. In correlations, Neuroticism emerged as the most robust LBA with the outcomes, with its maximum effect size being r = .32. The neuroticism LBA displayed criterion validity r's for wave 2 outcomes similar to the RIFT score itself. In regressions, it was clearer that Neuroticism was the most powerful LBA predictor of the outcomes. These findings indicate that language related to neuroticism emerging from the RIFT interview was important for the prediction of a variety of health outcomes. This is consistent with prior indicators of a relationship between neuroticism (assessed via self-report and through language) and lower respiratory symptoms (Oltmanns et al., 2021; Waszczuk et al., 2018). In the

prediction of change in the outcomes, the Neuroticism and Anger Proneness LBAs predicted increases in PTSD symptoms. This is consistent with the literature indicating the neuroticism and anger are key symptoms of PTSD prediction and severity (DiGangi et al., 2013). The present findings bolster prior evidence of associations between personality and psychopathology traits and health outcomes.

The present study advances empirical support for natural language prediction of important clinical health outcomes across time. However, it has several limitations to be improved in future research. First, the sample was assessed many years after the initial trauma (the 9/11 terrorist attacks). LBA predictive validity directly after trauma experience requires continued investigation (c.f., Schultebraucks et al., 2020). Language directly after a trauma may have even greater predictive effects. Second, the present study depended on language in structured clinical interviews. There are clinical situations where richer language could be obtained, for example in initial patient history interviews. However, the current LBA performance despite this limitation is promising. Finally, the present study is limited by a narrow demographic range (mostly white male WTC responders). More research is needed to establish generality of the findings across various gender, race, and ethnicity groups.

Despite early evidence of promise for AI for use in mental healthcare, the technology has yet to be implemented by mental health professionals. AI deployed to score personality from clinical interview language could significantly reduce time demands and increase predictive validity of assessments using behavioral data in a scalable way (e.g., automatic feedback provided to clinicians). The present results provide further evidence of the predictive validity of personality LBAs for trauma-related outcomes and objective mental health outcomes. This line of research

could lead to a promising prognostic tool to increase the validity and scalability of mental healthcare in the future.

References

Askew, R. L., Cook, K. F., Revicki, D. A., Cella, D., & Amtmann, D. (2016). Evidence from diverse clinical populations supported clinical validity of PROMIS pain interference and pain behavior. *Journal of Clinical Epidemiology*, *73*, 103–111. https://doi.org/10.1016/j.jclinepi.2015.08.035

Bittar, A., Velupillai, S., Downs, J., Sedgwick, R., & Dutta, R. (2020). Reviewing a Decade of Research Into Suicide and Related Behaviour Using the South London and Maudsley NHS Foundation Trust Clinical Record Interactive Search (CRIS) System. *Frontiers in Psychiatry*, *11*, 553463. https://doi.org/10.3389/fpsyt.2020.553463

Boag, W., Kovaleva, O., McCoy, T. H., Rumshisky, A., Szolovits, P., & Perlis, R. H. (2021). Hard for humans, hard for machines: Predicting readmission after psychiatric hospitalization using narrative notes. *Translational Psychiatry*, *11*(1), 32. https://doi.org/10.1038/s41398-020-01104-w

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the Trail Making Test. *Nature Protocols*, *1*(5), 2277–2281. https://doi.org/10.1038/nprot.2006.390

Bromet, E. J., Hobbs, M. J., Clouston, S. A. P., Gonzalez, A., Kotov, R., & Luft, B. J. (2016). DSM-IV post-traumatic stress disorder among World Trade Center responders 11–13 years after the disaster of 11 September 2001 (9/11). *Psychological Medicine*, *46*(4), 771–783. https://doi.org/10.1017/S0033291715002184

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Dasaro, C. R., Holden, W. L., Berman, K. D., Crane, M. A., Kaplan, J. R., Lucchini, R. G., Luft, B. J., Moline, J. M., Teitelbaum, S. L., Tirunagari, U. S., Udasin, I. G., Weiner, J. H.,

Zigrossi, P. A., & Todd, A. C. (2017). Cohort Profile: World Trade Center Health Program General Responder Cohort. *International Journal of Epidemiology*, *46*(2), e9–e9. https://doi.org/10.1093/ije/dyv099

DiGangi, J. A., Gomez, D., Mendoza, L., Jason, L. A., Keys, C. B., & Koenen, K. C. (2013). Pretrauma risk factors for posttraumatic stress disorder: A systematic review of the literature. *Clinical Psychology Review*, *33*(6), 728–744. https://doi.org/10.1016/j.cpr.2013.05.002

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C., Tobolsky, V., Smith, L. K., Buffone, A., Iwry, J., Seligman, M., & Ungar, L. H. (2020). *Closed and Open Vocabulary Approaches to Text Analysis: A Review, Quantitative Comparison, and Recommendations* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/t52c6

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), Article 2. https://doi.org/10.1177/2515245919847202

Koutsouleris, N., Hauser, T. U., Skvortsova, V., & De Choudhury, M. (2022). From promise to practice: Towards the realisation of AI-informed mental health care. *The Lancet Digital Health*, S2589750022001534. https://doi.org/10.1016/S2589-7500(22)00153-4

Leon, A. C., Solomon, D. A., Mueller, T. I., Turvey, C. L., Endicott, J., & Keller, M. B. (1999). The Range of Impaired Functioning Tool (LIFE-RIFT): A brief measure of functional impairment. *Psychological Medicine*, *29*(4), Article 4.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*(3), 450–470. https://doi.org/10.1037/a0019216

Merchant, R. M., Asch, D. A., Crutchley, P., Ungar, L. H., Guntuku, S. C., Eichstaedt, J. C., Hill, S., Padrez, K., Smith, R. J., & Schwartz, H. A. (2019). Evaluating the predictability of medical conditions from social media posts. *PLOS ONE*, *14*(6), Article 6. https://doi.org/10.1371/journal.pone.0215476

Natale, V., Léger, D., Bayon, V., Erbacci, A., Tonetti, L., Fabbri, M., & Martoni, M. (2015). The Consensus Sleep Diary: Quantitative Criteria for Primary Insomnia Diagnosis. *Psychosomatic Medicine, 77*(4), 413–418. https://doi.org/10.1097/PSY.0000000000000177

Oltmanns, J. R., Schwartz, H. A., Ruggero, C., Son, Y., Miao, J., Waszczuk, M., Clouston, S. A. P., Bromet, E. J., Luft, B. J., & Kotov, R. (2021). Artificial intelligence language predictors of two-year trauma-related outcomes. *Journal of Psychiatric Research*, *143*, 239–245. https://doi.org/10.1016/j.jpsychires.2021.09.015

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

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count (LIWC): LIWC 2007* [Computer software]. LIWC.net

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Rieman, D., Jaidka, K., Schwartz, H. A., & Ungar, L. (2017). Domain Adaptation from User-level Facebook Models to County-level Twitter Predictions. *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, 764–773.

Rutowski, T., Shriberg, E., Harati, A., Lu, Y., Oliveira, R., & Chlebek, P. (2021). Cross-Demographic Portability of Deep NLP-Based Depression Models. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 1052–1057. https://doi.org/10.1109/SLT48900.2021.9383609

Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., & Galatzer-Levy, I. R. (2020). Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. *Psychological Medicine*, 1–11. https://doi.org/10.1017/S0033291720002718

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., & Ungar, L. (2014). Towards Assessing Changes in Degree of Depression through Facebook. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 118–125. https://doi.org/10.3115/v1/W14-3214

Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., & Eichstaedt, J. (2017). DLATK: Differential Language Analysis ToolKit. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 55–60. https://doi.org/10.18653/v1/D17-2010

Shaw, M. J., Talley, N. J., Beebe, T. J., Rockwood, T., Carlsson, R., Adlis, S., Fendrick, A. M., Jones, R., Dent, J., & Bytzer, P. (2001). Initial validation of a diagnostic questionnaire for gastroesophageal reflux disease. *The American Journal of Gastroenterology*, *96*(1), 52–57. https://doi.org/10.1111/j.1572-0241.2001.03451.x

Son, Y., Clouston, S. A. P., Kotov, R., Eichstaedt, J. C., Bromet, E. J., Luft, B. J., & Schwartz, H. A. (2021). World Trade Center responders in their own words: Predicting PTSD symptom trajectories with AI-based language analyses of interviews. *Psychological Medicine*, 1–9. https://doi.org/10.1017/S0033291721002294

Soto, C. J. (2019). How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Waszczuk, M. A., Li, K., Ruggero, C. J., Clouston, S. A. P., Luft, B. J., & Kotov, R. (2018). Maladaptive Personality Traits and 10-Year Course of Psychiatric and Medical Symptoms and Functional Impairment Following Trauma. *Annals of Behavioral Medicine*, *52*(8), 697–712. https://doi.org/10.1093/abm/kax030

Waszczuk, M. A., Li, X., Bromet, E. J., Gonzalez, A., Zvolensky, M. J., Ruggero, C., Luft, B. J., & Kotov, R. (2017). Pathway from PTSD to respiratory health: Longitudinal evidence from a psychosocial intervention. *Health Psychology*, *36*(5), 429–437. https://doi.org/10.1037/hea0000472

Waszczuk, M. A., Ruggero, C., Li, K., Luft, B. J., & Kotov, R. (2019). The role of modifiable health-related behaviors in the association between PTSD and respiratory illness. *Behaviour Research and Therapy, 115*, 64–72. https://doi.org/10.1016/j.brat.2018.10.018

Watson, D., Nus, E., & Wu, K. D. (2019). Development and Validation of the Faceted Inventory of the Five-Factor Model (FI-FFM). *Assessment, 26*(1), 17–44. https://doi.org/10.1177/1073191117711022

Watson, D., O'Hara, M. W., Naragon-Gainey, K., Koffel, E., Chmielewski, M., Kotov, R., Stasik, S. M., & Ruggero, C. J. (2012). Development and Validation of New Anxiety and Bipolar Symptom Scales for an Expanded Version of the IDAS (the IDAS-II). *Assessment, 19*(4), Article 4. https://doi.org/10.1177/1073191112449857

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). *The PTSD Checklist for DSM-5 (PCL-5)—Standard [Measurement instrument]*. http://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp

Table 1

Correlations Between Language-Based Assessments and Self-Reports

LBA/Self-Report Construct	Pearson's r
Extraversion	.23
Conscientiousness	.27
Agreeableness	.21
Openness	.23
Neuroticism	.21
Anger Proneness	.09
Anxiousness	.10
Depressiveness	.20

Note. Each row presents a correlation between a language-based Assessment and its corresponding self-report score.

Table 2. *Correlations Between LBAs and Outcomes at Baseline and Wave 2.*

Outcomes	LBA							
Baseline	E	C	A	O	N	ANG	ANX	DEP
RIFT Functioning	**.18**	**.13**	**.11**	-.07	**-.15**	**-.18**	**-.19**	**-.24**
Trail Making Test Time	-.03	-.03	-.07	-.03	-.04	.00	-.02	-.04
IDAS General Depression	**-.20**	**-.22**	**-.14**	**.13**	**.28**	**.13**	**.18**	**.17**
IDAS Suicidality	-.11	-.11	**-.13**	.03	**.17**	.06	.06	.02
IDAS Well-Being	**.20**	**.25**	**.14**	-.03	**-.27**	**-.17**	**-.21**	**-.22**
PCL PTSD	**-.12**	**-.16**	**-.12**	.09	**.24**	.10	**.19**	**.12**
Lower Respiratory	.00	**-.14**	**-.19**	.05	**.25**	.10	**.12**	.08
GERD	-.03	-.07	-.02	.06	**.12**	.07	**.15**	.08
Pain	**-.16**	-.11	-.04	**.14**	**.26**	.03	**.14**	**.14**
Sleep Quality	.07	**.14**	.06	**-.12**	-.08	-.07	-.08	-.10
Mental Healthcare Cost	**-.11**	**-.16**	**-.15**	.05	**.19**	.02	.08	.06
Wave 2								
RIFT Functioning	**.15**	**.16**	.11	-.01	**-.18**	**-.15**	**-.13**	**-.16**
Trail Making Test Time	.07	.00	-.08	-.07	.00	-.04	**-.14**	**-.12**
IDAS General Depression	**-.13**	**-.18**	**-.14**	.14	**.32**	**.22**	**.23**	**.20**
IDAS Suicidality	**-.16**	**-.15**	-.05	.09	**.24**	**.13**	**.12**	.10
IDAS Well-Being	**.13**	**.17**	**.12**	-.09	**-.22**	**-.18**	**-.18**	**-.20**
PCL PTSD	**-.11**	**-.19**	**-.15**	.11	**.30**	**.20**	**.21**	**.16**
Lower Respiratory	-.05	**-.18**	**-.21**	.04	**.31**	**.12**	**.13**	.07
GERD	-.03	-.11	**-.16**	.04	**.23**	**.13**	**.14**	.09
Pain	**-.17**	**-.18**	**-.17**	**.14**	**.29**	.05	**.15**	**.19**
Sleep Quality	.08	.03	.11	**-.13**	-.10	-.09	-.05	-.10
Mental Healthcare Cost	-.08	**-.17**	**-.14**	.04	**.17**	.00	.07	.03
Median absolute value *r*	.11	.16	.13	.07	.23	.10	.14	.11

Note. Bold = significant correlation ($p < .05$). LBA = language-based assessment, RIFT = Range of Impaired Functioning interview, IDAS = Inventory of Depression and Anxiety Symptoms-II, E = extraversion, C = conscientiousness, A = agreeableness, O = openness, N = neuroticism, ANG = anger proneness, ANX = anxiousness, DEP = depressiveness, PTSD = PCL-5 Symptom Checklist, LRS = lower respiratory symptoms, GERD = gastro-esophageal reflux disease symptoms.

Table 3. *Hierarchical Regression Test of Baseline LBAs Over RIFT Interview Total Score for Wave 2 Outcomes*

	Step 1 (IV: RIFT Total)		Step 2 (IVs: 8 LBAs)		
DV: Wave 2 Outcome	R^2	*p*	R^2	ΔR^2	*p*
RIFT Functioning	.431	.000 ***	.446	.014	.471
Trail Making Test	.000	.736	.043	.043	.210
IDAS General Depression	.220	.000 ***	.308	.088	.000 ***
IDAS Suicidality	.051	.000 ***	.116	.065	.006 **
IDAS Well-Being	.136	.000 ***	.178	.042	.055
PCL PTSD	.175	.000 ***	.263	.088	.000 ***
Lower Respiratory	.067	.000 ***	.177	.110	.000 ***
GERD	.103	.000 ***	.161	.057	.011 *
Pain	.087	.000 ***	.178	.091	.000 ***
Daily Sleep Quality	.023	.015 *	.059	.036	.319
Mental Healthcare Dollars	.091	.000 ***	.144	.053	.010 *

Note. *** $p < .001$, ** $p < .01$, * $p < .05$ in step 2. LBA = language-based assessment, RIFT = Range of Impaired Functioning interview, IDAS = Inventory of Depression and Anxiety Symptoms - II, PTSD = PCL-5 Symptom Checklist, GERD = gastro-esophageal reflux disease symptoms. Each row is a separate model.

Table 4. *Hierarchical Regression Test of Baseline Language Predictors Over Baseline Outcomes for Wave 2 Outcomes*

	Step 1 (IV: Wave 1 Outcome)		Step 2 (IV: 8 LBAs)		
Wave 2 Outcome	R^2	*p*	R^2	ΔR^2	*p*
RIFT Functioning	.431	.000 ***	.446	.014	.471
Trail Making Test	.443	.000 ***	.482	.039	.024 *
IDAS General Depression	.587	.000 ***	.613	.026	.019 *
IDAS Suicidality	.372	.000 ***	.415	.042	.013 *
IDAS Well-Being	.495	.000 ***	.502	.008	.831
PCL PTSD	.686	.000 ***	.710	.024	.005 **
Lower Respiratory	.640	.000 ***	.657	.017	.096
GERD	.476	.000 ***	.513	.037	.009 **
Pain	.476	.000 ***	.509	.033	.021 *
Daily Sleep Quality	.549	.000 ***	.574	.025	.097
Mental Healthcare Dollars	.798	.000 ***	.801	.004	.614

Note. *** $p < .001$, ** $p < .01$, * $p < .05$ in step 2. LBA = language-based assessment, RIFT = Range of Impaired Functioning interview, IDAS = Inventory of Depression and Anxiety Symptoms - II, PTSD = PCL-5 Symptom Checklist, GERD = gastro-esophageal reflux disease symptoms. Each row is a separate model.

Table 5. *Significant Individual LBA Predictors of Outcomes*.

	LBA	Beta	p	Zero-order *r*
Models Including RIFT in Step 1				
General Depression	Neuroticism	.25	.000	.32
Suicidality	Neuroticism	.19	.008	.24
PTSD	Neuroticism	.23	.000	.30
Lower Respiratory	Neuroticism	.29	.000	.31
GERD	Neuroticism	.21	.003	.23
Pain	Neuroticism	.22	.001	.29
Mental Healthcare Cost	Anxiousness	.21	.029	.07
Models Including Outcome in Step 1				
Trail Making Test	Extraversion	.12	.024	.07
Trail Making Test	Neuroticism	.12	.046	.00
Trail Making Test	Anxiousness	-.21	.016	-.14
General Depression	Neuroticism	.13	.008	.32
Suicidality	Neuroticism	.13	.029	.24
PCL PTSD	Neuroticism	.11	.009	.30
PCL PTSD	Anger Proneness	.13	.007	.20
GERD	Neuroticism	.13	.016	.23
Pain	Agreeableness	-.10	.037	-.17

Note. LBA = language-based assessment, RIFT = Range of Impaired Functioning interview, PTSD = PCL-5 Symptom Checklist, GERD = gastro-esophageal reflux disease symptoms. Each DV is one model within the two sets of regressions, each model controls for all other LBAs.