Considerations in Personality Measurement

REPLICABILITY, TRANSPARENCY, AND PREDICTIVE VALIDITY

MATTHEW E. K. HALL, University of Notre Dame, USA
GARY E. HOLLIBAUGH JR., University of Pittsburgh, USA
JONATHAN D. KLINGLER, University of Mississippi, USA
ADAM J. RAMEY, New York University Abu Dhabi, UAE

ABSTRACT

In another article in this issue, Black et al. discuss their preferred approach to estimating Supreme Court justices' Big Five personality traits from written text and provide several critiques of the approach of Hall et al. In this rejoinder, we show that Black et al.'s critiques are substantially without merit, their preferred approach suffers from many of the same drawbacks that they project onto our approach, their specific method of implementing their preferred approach runs afoul of many contemporary social scientific norms, our use of concurrences to estimate personality traits is far more justifiable than they suggest (especially in contrast to their use of lower court opinions), and their substantive critiques reflect a potential misunderstanding of the nature of conscientiousness. Nonetheless, we also acknowledge their broader point regarding the state-of-the-art textual analysis methodology vis-à-vis the estimation of personality traits, and we provide some constructive suggestions for the path forward.

We thank the editors for the chance to respond to Black et al.'s (2021) critique of our work. We are thrilled to see Black et al. follow our lead by studying the intersection of personality psychology and judicial behavior, and we are flattered to think that our work merits such thorough, close, and detailed attention. Black et al. aim to highlight "the magnitude of the gap between the ideal and [our] approach" (378). In short, we agree that our scores, like virtually any measure, are imperfect (in some, though not all, of the ways that Black et al. point out).

This article was reviewed by the editor, Tom S. Clark. Contact the corresponding author, Matthew E. K. Hall, at matt.hall@nd.edu.

Electronically published August 31, 2021.

Journal of Law and Courts, volume 9, number 2, Fall 2021.
© 2021 Law and Courts Organized Section of the American Political Science Association. All rights reserved. Published by The University of Chicago Press for the Law and Courts Organized Section of the American Political Science Association. https://doi.org/10.1086/715464

https://doi.org/10.1086/715464 Published online by Cambridge University Press

However, we also want to note several points that we believe are relevant to this exchange. First, though we believe the bar they have set might be a bit high, Black et al.'s estimates are also far from ideal. In fact, their estimates violate some of the most basic norms of contemporary social science; most notably, their estimates are impossible to replicate. Second, some of Black et al.'s most biting critiques also apply to their own estimates. Most strikingly, they criticize us for ignoring the distinctions between different types of concurring opinions, yet they ignore the distinctions between lower court opinions. Third, their complaints about the predictive validity of our scores are based on a basic misunderstanding of the conscientiousness personality trait. Unfortunately, the scope of this symposium does not allow for original analyses with their estimates (and many of the analyses we would like to run are impossible because their estimates are not replicable). However, we are confident that, if they were subjected to a comparable level of scrutiny, a gap of similar or greater magnitude would likely arise between "the ideal" and the Black et al. (2020) approach.

Ultimately, the limits of the Hall et al. (2021) estimates, as well as Black et al.'s, reflect the challenges inherent to this nascent field. We encourage other scholars to develop even better estimates of judicial personality. Indeed, we hope the day is not long off when all of the estimates discussed in this symposium are obsolete. We hope the talented colleagues in our discipline (including the Black et al. team) will devote their talents toward that new, creative work, rather than criticisms of initial, imperfect research—lest the perfect become the enemy of the good.

HOW SHOULD ONE ESTIMATE JUDICIAL PERSONALITY?

Both Hall et al. (2021) and Black et al. (2021) seek to estimate judicial personality traits from written text, with Black et al. (2020) utilizing a proprietary method from IBM (Watson Personality Insights [WPI]) and Hall et al. (2021) using an open-source method via a Java applet called the Personality Recognizer (Mairesse et al. 2007; see Ramey, Klingler, and Hollibaugh 2017). Both teams find personality matters, with Black et al. focusing on conscientiousness and Hall (2018) examining all Big Five traits.

However, despite these similarities, the two pieces in this issue exhibit many important differences, especially in the underlying methodology. Black et al.'s proprietary approach is based on a word-embedding technique (Global Vectors for Word Representation [GloVe]) that transforms the words in the input text into high-dimensional vectors that have been trained on enormous corpora of web-based data (Pennington, Socher, and Manning 2014). These representations are then used by IBM to train supervised models to estimate the fit between vector representations of textual corpora and other outcomes of interest (here, 1,500–2,000 respondents' Twitter feeds are transformed into their vector representations and used to estimate self-reported personality traits). Black et al. then use the pretrained models to out-of-sample estimate the personality traits of Supreme Court justices using preconfirmation opinions, speeches, and published articles or other writings.

https://doi.org/10.1086/715464 Published online by Cambridge University Press

Our approach is a dictionary-based method with a long history in personality psychology, using Pennebaker, Francis, and Booth's (2001) Linguistic Inquiry and Word Count (LIWC) and Coltheart's (1981) MRC Psycholinguistic Database (MRCPD).[1] LIWC, which has previously been used to estimate personality from text (Pennebaker and King 1999; Mairesse et al. 2007; Mairesse and Walker 2008), characterizes words by structure, tone, and topic according to a preexisting dictionary; MRCPD takes a similar approach, focusing on different linguistic features. In our case, the underlying personality trait models are trained on Pennebaker and King's (1999) essays corpus, which contains over 1.9 million words across several thousand stream-of-thought essays written by undergraduate students who also took tests to assess their Big Five personality traits. Once the model is trained on the LIWC- and MRCPD-based representations of Pennebaker and King's (1999) corpus, our approach is quite similar to Black et al.'s: we use the pretrained models to out-of-sample estimate the justices' personality traits, using concurring opinions as the input text.

Fundamentally, we have no quarrel with Black et al.'s broader approach. The use of GloVe is not a problem in any sense; indeed, three of us have used GloVe in some of our other work (Ramey, Klingler, and Hollibaugh 2019b) and hope to use it and more recent methods in future work. Indeed, GloVe is a highly promising approach that, if applied properly, can provide greater flexibility than the LIWC- and MRC-based approach of our previous work. Our criticisms are instead based on three overarching themes: (1) Black et al.'s estimates are neither replicable nor verifiable, which stems from their decision to use proprietary software instead of an open-source implementation of GloVe; (2) their criticism of our use of concurrences is overwrought, and their text is subject to the same criticisms; and (3) many of their substantive critiques are based on a misunderstanding about the nature of conscientiousness. However, we share their enthusiasm with the broader goal of estimating judicial personality and end this rejoinder with some suggestions for the path forward.

BLACK ET AL.'S PERSONALITY SCORES ARE NEITHER REPLICABLE NOR VERIFIABLE

Foremost, Black et al.'s scores are not replicable and unverifiable. The proprietary nature of IBM's model creates significant hurdles for the Black et al. application and, as such, we cannot in good conscience recommend that social scientists adopt this approach. The underlying WPI model is a proprietary black box, and it is unclear which machine-learning algorithm is used to create the specific scores once the GloVe representations for the judicial texts are provided; the IBM website simply says a "machine-learning algorithm,"

1. We used the 2001 LIWC dictionary to be consistent with the work of Mairesse et al. (2007). However, in practice, we would recommend newer versions due to the addition of several linguistic categories—such as "netspeak"—that provide additional sources of variation and would likely be of substantive interest to applied scholars.

https://doi.org/10.1086/715464 Published online by Cambridge University Press

without further clarification.[2] Additionally, the specific tweets used to train and test the model are unknown to the public and are likely to remain that way. The underlying model has also been altered since the publication of Black et al. (2020). Indeed, as Black et al. note in the readme.txt file accompanying their replication data, "subsequent to our initial estimation IBM revised the process by which they calculated the neuroticism trait and so those results are quite a bit different because of this change (i.e., the correlation is 0.59)." Not only will IBM's Personality Insights tool be discontinued in December 2021, but the tool has disallowed new instances since December 2020. In short, for many reasons, the reliance on this software has led us to a situation in which no one, not even Black et al., can replicate Black et al.'s original estimates from their texts. This saga exemplifies the pitfalls of relying on proprietary cloud-based software outside of the control of the academic researcher: the researcher is at the mercy of an unaccountable entity that can change critical aspects of their research workflow.

These problems illustrate the double standard Black et al. have applied to SCIPEs. For example, in their book, Black et al. (2020) lamented that the Personality Recognizer was no longer supported by the developer. Black et al. construed this fact to indicate that the models therein were somehow flawed, a claim akin to suggesting that regression models estimated by now-obsolete statistical programs are unreliable; while the software might be outdated, there is no reason to think the underlying models are problematic. While the models used to estimate our personality scores are no longer state of the art, it is undeniable that we have been transparent about the specific models used and the source data used to train said models, and we have provided instructions on how to set up the Personality Recognizer applet and call it through R.[3] Indeed, anyone can estimate models using our approach, whereas as of December 1, 2020, no one can estimate models using Black et al.'s approach, and Black et al. will be unable to estimate their own models after December 31, 2021.

We also differ with Black et al. regarding appropriate data used to train our respective models. As Black et al. note, IBM's Personality Insights tool is trained on Twitter feeds,[4] whereas the Mairesse et al. (2007) models we used were based on Pennebaker and King's (1999) corpus of essays written by Texas undergraduates.[5] One is free to quibble over whether the increased linguistic complexity afforded by the stream-of-thought essays

2. See https://cloud.ibm.com/docs/personality-insights?topic=personality-insights-science.

3. See the Dataverse replication website for Ramey, Klingler, and Hollibaugh (2019a) at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JDH6KR. Importantly, the Personality Recognizer is not strictly necessary for our purposes; one could simply extract the underlying models from the website set up by the original developer at http://farm2.user.srcf.net/research/personality/recognizer.html.

4. See https://cloud.ibm.com/docs/personality-insights?topic=personality-insights-science.

5. While the larger GloVe word vectors are trained using billions of words from Wikipedia, Twitter, and other sources (Pennington et al. 2014), the model implemented by IBM is much more limited, since only the 1,500–2,000 respondents who took IBM's personality tests are usable for their specific personality trait models.

https://doi.org/10.1086/715464 Published online by Cambridge University Press

outweighs the arguably increased representativeness of the Twitter sample. It is impossible to know whether this is true; while we have acknowledged the limitations of the data used to train our models, we have no information on the representativeness of IBM's Twitter sample, the number of words used in said sample, or anything to that effect. While this is partially understandable due to the proprietary nature of the model, that this information is not available to anyone (as far as we can tell) undercuts any claim that their training data are more appropriate to the task at hand. Like many of their other claims, it may be true, but the proprietary nature of their model makes it impossible to know with any degree of certainty.

CONCURRENCES VERSUS LOWER COURT OPINIONS

Black et al. also critique our use of concurring opinions to estimate personality traits, favoring instead their use of various preappointment texts. First, they raise a single-source critique, arguing that we are simply asking "too much of a single type of document" (2021, 378). Justices may adopt a professional persona when writing opinions that differs from their personality in other contexts (see Tett et al. 2013). We are not overly concerned with this critique because if the professional, judicial context prompts justices to activate certain traits, the same dynamic should occur in all of the professional activities in which we are interested. Nonetheless, we agree with the general thrust of the single-source critique in that fewer textual sources provide less information and potentially introduce bias, and Black et al. undoubtedly uses a larger set of texts to calculate their estimates.

The single-source critique might be more persuasive if Black et al. did not also raise a multisource critique, lamenting "a general lack of consistency within the corpus of a justice's concurring opinions" (2021, 382). Justices write different types of concurrences, and combining these different texts might introduce bias. In other words, Black et al. critique us for using only one type of text and too many different types of text. This critique is also valid, as we acknowledge in our prior work.[6]

However, the juxtaposition of these contradictory arguments highlights the fact that literally any corpus of text could be criticized on both of these grounds. Were we to focus on only one type of concurrence, one could further subdivide that category and complain about a lack of consistency in the texts. Were we to broaden our corpus to include other types of text, we could ameliorate the single-source problem, but we would simultaneously exacerbate the multisource problem.

Second, since this critique applies to any textual corpus, it obviously applies to the text Black et al. used to create their estimates. Their corpus is broader than ours, but it can still be characterized as a single type of text: preconfirmation, publicly available, professional writings. To their credit, Black et al. (2020) take steps to mitigate the (more serious) multisource problems in their text. Specifically, they assess the extent to which personality traits pervade each document type by comparing the justices' lower court opinions,

6. Black et al. erroneously claim that Hall (2018) does not address this concern. See Hall (2018, 38–39) for a brief discussion of the multisource critique.

https://doi.org/10.1086/715464 Published online by Cambridge University Press

speeches, and "other writings" (38–41) against randomly generated documents of each type (based on randomly selected sentences from each corpus). Results are standardized within each of the three separate corpora and combined across document types to generate a single estimate for each trait for each justice. This is a creative approach, and we commend them for taking seriously the issue of inconsistency across document types.

Unfortunately, their implementation is incomplete, as they make no attempt to correct for the "general lack of consistency" within each of these categories. Just as there are different types of concurrences, there are different types of lower court opinions, yet Black et al. (2020) lump them all together. They also combine all speeches (regardless of context, audience, etc.) and all "other writings" (coauthored and single authored, law reviews and other articles, and unspecified other writings) while making no effort to account for these differences within each category. We are not suggesting these problems are fatal to Black et al.'s (2020) estimates, just as we do not believe their critiques are fatal to ours. However, given their extensive critique of the needle in our eye, it is hard not to notice the plank in their own, and we believe it is misleading to ignore it.

Of course, these problems might be less serious in Black et al.'s estimates than in ours. We would love to conduct the same exhaustive analysis on their estimates to evaluate this possibility. Sadly, since their estimates are inherently not replicable, the best we can say is no one will ever know.

DISAGREEMENT ABOUT CONSCIENTIOUSNESS

Finally, Black et al. criticize the predictive validity of our estimates. They claim that "conscientious justices will treat petitions with weak legal conflict more seriously than do less conscientious justices" (2021, 390), based on three assumptions: "Conscientious people take their professional obligations seriously"; "One of the justices' central duties is to ensure uniformity in the law"; and "In cases with strong conflict, all justices will be inclined toward granting review" (390). Thus, they expect and find (using their estimates) that weak (but not strong) conflict has more influence on more conscientious justices. Using our estimates, both weak and strong conflict have more influence on less conscientious justices, but several aspects of Black et al.'s argument strike us as unrealistic.

First, Black et al.'s logic assumes that, when justices vote against cert in weak conflict cases, they do so because they do not take their professional obligations seriously. Based on Black et al.'s (2021) figure 6*b*, the least conscientious justices are so willing to flout their obligations that the effect of weak conflict on their cert votes is not significant, but we would not expect any justice (even less conscientious justices) to ignore their professional duty. Second, Black et al.'s theory and findings indicate a sharp juxtaposition between weak and strong conflict. In their narrative, the least conscientious justices completely discount weak conflict but take strong conflict very seriously; indeed, their response to strong conflict is no different than the most conscientious justices. Black et al., however, do not explain why justices would make such a sharp distinction between weak and strong conflict.

https://doi.org/10.1086/715464 Published online by Cambridge University Press

Most importantly, Black et al. focus on the dutifulness aspect of conscientiousness but disregard industriousness (DeYoung, Quilty, and Peterson 2007). They note "conscientious workers 'are predisposed to be organized, exacting, disciplined, diligent, dependable, methodical, and purposeful'" (Black et al. 2021, 390), yet they ignore how these traits relate to circuit conflict. Consider, however, that the justices vote on roughly 800 cert petitions each term (Johnson 2016), and the presence of strong or weak conflict is a subjective determination made by law clerks (Black and Owens 2009, 1069). Thus, Black et al. are actually testing the effect of clerks signaling that conflict exists given the justices' heavy workloads. Viewed in this light, less conscientious justices should follow simple signals from their clerks (even signals of weak conflict) in order to reduce their workload. In contrast, more conscientious justices should place less weight on these signals because they invest the time to independently examine the petitions to determine if a genuine conflict exists, not because they disregard their professional duties (and doing so may help them avoid future effort required if the case were argued before the Court; Ramey et al. 2017, 48). Indeed, we are struck by how closely Black et al.'s findings using our estimates align with our theoretical expectations.

Our point is that the translation of personality theory into expectations regarding judicial behavior is complicated and controversial. Black et al. could have chosen any one of numerous empirical tests to critique our estimates. Yet, for the one example intended to demonstrate that our estimates are flawed, they chose an application that is, at best, highly debatable.[7]

Black et al.'s critique also ignores the dozens of empirical tests Hall (2018) conducted demonstrating the validity of our estimates. We do not mean to imply that Hall's (2018) logic or findings are all correct (Hall himself would like to emphasize his humility on this front). Black et al. seem to imply that a single instance in which SCIPEs lack predictive validity would show "critical flaws" (2021, 394) in the measures. However, we note that Black et al. (2020) developed estimates for all five personality traits, yet their book examines only conscientiousness. We have no quarrel with their substantive focus, but their decision leaves us wondering: if we were allowed to test the effects of the other four traits on various forms of judicial behavior in this response, might we find a single instance in which our estimates outperform theirs? We suspect we would, but we would not conclude that their estimates have critical flaws based on such a finding.

A PATH FORWARD?

It is very possible that Black et al.'s (2020) estimates are superior to SCIPEs at recovering justices' "true" personality traits. While GloVe is less theoretically motivated than the

7. In the interest of brevity, we do not focus on some of Black et al.'s less credible arguments. For example, their claim that "the logical implication of [more conscientious justices being less likely to cast a grant vote] is that conscientious justices . . . would not grant review to any cases" (2021, 393) is a textbook example of reductio ad absurdum.

https://doi.org/10.1086/715464 Published online by Cambridge University Press

LIWC and MRCPD dictionaries, it provides greater flexibility during the fitting process. The training data underlying the Personality Insights model is potentially more representative than the training data underlying our model. However, Black et al. (2020, 2021) have painted themselves into a corner with their use of IBM's Personality Insights, as its proprietary nature makes it impossible for anyone to examine the Twitter data and/or specific algorithms used to train and/or fit their models, or even to replicate their scores. Conversely, our approach is much more open and replicable. Given the impossibility of replicating and validating their scores, we cannot in good conscience recommend their use. We would encourage the estimation of personality traits using similar—even identical—approaches, so long as such scores are verifiable and replicable. If such scores end up being superior to our own, we will gladly adopt them and encourage others to do the same.

More generally, the process of recovering elite personality traits from text should be transparent, verifiable, and replicable. Open-source software should be used whenever possible, the textual sources used to train any models should be publicly posted, and the models themselves should be made public as early as practicable. Additionally, the textual sources used to train the models should be diverse and representative in terms of contributors and sources, and they should be at least as linguistically complex as the text to be used for out-of-sample fit, which would mitigate some of the issues with using unrepresentative or potentially linguistically problematic sources.

On that note, the textual sources used for out-of-sample fit should be as comparable as possible—in terms of linguistic and structural features—to those used to train and test the underlying model. Ramey et al. (2017, 2019a) and Hall et al. (2021) show that the distributions of LIWC features within their respective corpora correlate highly with the distributions of such features in the undergraduate essays used to train the models embedded in the Personality Recognizer. This step is all the more important when the documents used for out-of-sample fit differ in terms of structure and process from those used to train the statistical models of personality.

Most importantly, the technology used to predict personality from text is still at an early stage, and the out-of-sample prediction needed to estimate such traits for elites means measurement error is going to be more prevalent than it is for, say, measures of elite ideology. Thus, the potential for newer measures to outperform and replace older ones is high, and we welcome such developments. Future measures will be subject to the problems faced by both sets of authors in this symposium. Accordingly, we recommend that all scholars treat their own measures with humility and competing measures with the open minds that the research enterprise deserves.

REFERENCES

Black, Ryan C., and Ryan J. Owens. 2009. "Agenda Setting in the Supreme Court: The Collision of Policy and Jurisprudence." *Journal of Politics* 71 (3): 1062–75.

https://doi.org/10.1086/715464 Published online by Cambridge University Press

Black, Ryan C., Ryan J. Owens, Justin Wedeking, and Patrick C. Wohlfarth. 2020. *The Conscientious Justice: How Supreme Court Justices' Personalities Influence the Law, the High Court, and the Constitution*. New York: Cambridge University Press.

———. 2021. "On Estimating Personality Traits of US Supreme Court Justices." *Journal of Law and Courts* 9 (2): 371–96.

Coltheart, Max. 1981. "The MRC Psycholinguistic Database." *Quarterly Journal of Experimental Psychology* A 33 (4): 497–505.

DeYoung, Colin G., Lena C. Quilty, and Jordan B. Peterson. 2007. "Between Facets and Domains: Ten Aspects of the Big Five." *Journal of Personality and Social Psychology* 93 (5): 880–96.

Hall, Matthew E. K. 2018. *What Justices Want: Goals and Personality on the U.S. Supreme Court*. New York: Cambridge University Press.

Hall, Matthew E. K., Gary E. Hollibaugh Jr., Jonathan D. Klingler, and Adam J. Ramey. 2021. "Attributes beyond Attitudes: Personality Traits on the US Supreme Court." *Journal of Law and Courts* 9 (2): 345–70.

Johnson, Timothy R. 2016. The Supreme Court Decision Making Process. In *Oxford Research Encyclopedia of Politics*, ed. Lee Epstein and Stefanie Lindquist. New York: Oxford University Press.

Mairesse, François, and Marilyn A. Walker. 2008. Trainable Generation of Big-Five Personality Styles through Data-Driven Parameter Estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 165–73.

Mairesse, François, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text." *Journal of Artificial Intelligence Research* 30 (1): 457–500.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pennebaker, James W., and Laura A. King. 1999. "Linguistic Styles: Language Use as an Individual Difference." *Journal of Personality and Social Psychology* 77 (6): 1296–312.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 1532–43.

Ramey, Adam J., Jonathan D. Klingler, and Gary E. Hollibaugh Jr. 2017. *More than a Feeling: Personality, Polarization, and the Transformation of the U.S. Congress*. Chicago: University of Chicago Press.

———. 2019a. "Measuring Elite Personality Using Speech." *Political Science Research and Methods* 7 (1): 163–84.

———. 2019b. "More than Words: Extracting Latent Psychological Traits and Preferences from Written Text." Unpublished manuscript.

Tett, Robert P., Daniel V. Simonet, Benjamin Walser, and Cameron Brown. 2013. "Trait Activation Theory: Applications, Developments, and Implications for Person-Workplace Fit." In *Handbook of Personality at Work*, ed. Neil D. Christiansen and Robert P. Tett, 77–100. New York: Routledge.

https://doi.org/10.1086/715464 Published online by Cambridge University Press