8-25-2021

On Estimating Personality Traits of US Supreme Court Justices

Ryan C. Black
Michigan State University

Ryan J. Owens
University of Wisconsin

Justin Wedeking
University of Kentucky, justin.wedeking@uky.edu

Patrick C. Wohlfarth
University of Maryland

On Estimating Personality Traits of US Supreme Court Justices

Digital Object Identifier (DOI)
https://doi.org/10.1086/714888

Notes/Citation Information
Published in *Journal of Law and Courts*, v. 9, no. 2.

On Estimating Personality Traits of US Supreme Court Justices

RYAN C. BLACK, Michigan State University, USA
RYAN J. OWENS, University of Wisconsin, USA
JUSTIN WEDEKING, University of Kentucky, USA
PATRICK C. WOHLFARTH, University of Maryland, College Park, USA

ABSTRACT

Psychological scholarship on personality is uniting with political science to redefine existing theories. This is clearly the case with research on judicial behavior and the US Supreme Court. But if this new approach is to survive and thrive, it must employ measures equal to the task. We show that Supreme Court Individual Personality Estimates, which seek to estimate justices' personalities by examining their concurring opinions, suffer from a number of important methodological deficits that critically limit their usefulness. We briefly discuss what kinds of improved personality measures scholars should use instead and offer an improved set of estimates for one trait with an application that demonstrates our cautionary tale.

Pick up any biography on a Supreme Court justice. You will find a substantial amount of attention directed to his or her personality. Read news articles about high court nominees, and you are sure to come across something about their personalities. Investigate the justices' private archival papers, and instantly you will develop a sense for their different styles and personalities. Retrospective analyses about the Court's previous terms—as well as prospective analyses about its future terms—seem always to address justices' personalities. People want to know about the justices' personalities.

Scholarly interest in the empirical connection between personality and judging is growing steadily as well (e.g., Braman and Nelson 2007; Braman 2009; Collins 2011; Moyer 2012; Owens and Wedeking 2012; Hall 2018; Black et al. 2020). Whereas Epstein and Knight's (1998) pacesetting book on strategic judging did not mention the term "personality" once, Epstein, Landes, and Posner's (2013) reassessment of judicial behavior included

Contact the corresponding author, Patrick C. Wohlfarth, at patrickw@umd.edu.

Electronically published August 25, 2021.

Journal of Law and Courts, volume 9, number 2, Fall 2021.
 https://doi.org/10.1086/714888

at least 16 references to it. Personality scholarship could wind up filling key voids in our current understanding of judicial behavior. But for personality research to take hold, and for scholars to maximize its potential, we require accurate measures of justices' personalities.

Two recent books empirically examine the effects of justices' traits on judicial behavior: Hall (2018) and Black et al. (2020). As this article dialogue hopefully demonstrates, we believe both have the potential to push the scholarly agenda. Hall investigates how all Big Five traits influence five judicial actions: agenda setting, opinion assignment, intra-Court bargaining, voting, and writing separate opinions. Our work focuses on a single trait, conscientiousness, and examines how it influences nine judicial actions: agenda setting, legal persuasion in oral argument and legal writing, the decision to side with the solicitor general, majority opinion assignments, opinion bargaining, the content of the Court's opinions, the treatment of precedent, whether justices follow public opinion, and when justices recuse. Taken together, these works show that personality influences nearly all aspects of judicial behavior.

We have two goals with this paper. First, we aim to raise concerns about the personality measures originally employed by Hall (2018) and subsequently published in Hall et al. (2021): the Supreme Court Individual Personality Estimates (SCIPEs). To do so, we deploy a more inclusive set of ideology measures to demonstrate there is ultimately limited evidence of SCIPE's empirical validity. We then seek to uncover the source of this validity deficit by examining SCIPE's reliance on a single source of text: concurring opinions written by the justices. Beyond the inherent circularity of this approach, our results find that concurring opinions may not actually reflect justices' personalities.

Second, we then seek to provide constructive insights into what kinds of estimates could be appropriate to examine justices' personalities. After explaining what those estimates might look like, we apply the conscientiousness trait—the personality concept on which we have focused extensively (Black et al. 2020)—to the relationship between lower court conflict and justices' agenda-setting votes. The results reveal the significant substantive limitations and inferential risks of using SCIPEs to study personality on the Supreme Court and the need for more careful measures for this growing scholarly market.

REASSESSING SCIPE VALIDITY

To validate their five trait estimates, Hall et al. (2021) subject those estimates to two examinations related to the concept of ideology. First, they analyze the correlation between their trait scores and the ideological direction of each justice's vote in each case (Spaeth et al. 2010).[1] Second, they analyze the correlation between their trait scores and a justice's Clerk-Based Ideology (CBI) score, which come from the political campaign contributions made by their former law clerks (Bonica et al. 2017).

1. See Hall et al. (2021, n. 8), where they state that their binomial logit model (with 34 observations) is "tantamount to estimating a logit model where the *unit of analysis is the individual vote*," which would have well over 60,000 observations (emphasis added).

We suspect few scholars of the Court would identify these two particular measures as offering the most direct or intuitive assessment of justice ideology, so we examine four additional approaches to measuring judicial preferences. First, we look at the simple proportion of conservative votes cast by each justice over the course of his or her career. We count the number of conservative votes and divide that into the total number of votes cast by the justice that had a determinable ideological direction. This is the approach used in basically every existing study that introduces a new approach to measuring preferences, ranging from Segal-Cover scores to Martin-Quinn scores (Martin and Quinn 2002) to even the validation of CBI scores themselves.

Second, we examine each justice's Segal-Cover score, which is derived from a content analysis of the preconfirmation hearing newspaper editorials written about each nominee (see also Segal et al. 1995). These scores have the added bonus of being the only measure of ideology not derived from any aspect of how a justice behaves once on the Court (cf. voting on cases or hiring law clerks). That strength is arguably also their greatest weakness, however, since a good number of justices seem to show evidence of ideological change over the course of their careers (e.g., Epstein, Martin, Quinn, and Segal 2007).

Our final two additions are more sophisticated approaches that summarize judicial votes through the use of item response theory models. In particular, we estimate a two-parameter item response model to produce a single, career-level estimate of each justice's ideal point. That is, as our third measure, we estimate a justice's static Martin-Quinn score (Martin and Quinn 2002).[2] Our final measure is a career-averaged ideology score as estimated by the approach suggested by Bailey (2007, 2013). The key difference between these latter two approaches is that the Martin and Quinn–inspired version uses only the actual votes cast by justices on cases, whereas Bailey's formulation creatively leverages instances where a justice takes a position on a case in which she was not a voting member of the Court (e.g., Justice Thomas writing that he believes the Court ruled wrongly in a previously decided case, decades before he joined the Court).

All four of our additional measures as well as the CBI score are near continuous in nature, so we use ordinary least squares regression to examine the relationship between alternative ideology measures and SCIPE scores. And, because the individual votes justices cast in each case are dichotomous (1 if liberal, 0 if conservative), we use logistic regression for those data.[3]

2. Hall et al. (2021, n. 7) reject using this approach. They assert that summarizing the Martin-Quinn scores to a single value per justice would result in "a great degree of imprecision." This remark is paradoxical, because the CBI scores on which they rely follow precisely such an approach. Each year, a justice hires multiple law clerks, some of whom will have Campaign Finance scores, as identified by Bonica et al. (2017). Justices serve for multiple years. As calculated by Bonica et al. (2017), a justice's CBI score is determined by averaging all of those clerks who have Campaign Finance scores. Thus, it pools together all clerks, and all of those terms of data are being reduced to a single value. We fail to see what actual unique hazard a career-averaged Martin-Quinn score poses that a career-averaged CBI score does not.

3. The data come from both the "modern" and "legacy" versions of the Supreme Court Database, which give us a total of over 72,000 votes from the Court's 1937 to 2019 terms. As no SCIPE estimates

In addition to the individual trait estimates, Hall et al. (2021) also include controls for a justice's sex and birth year. We estimate our models both with and without these controls, both to assess sensitivity to their exclusion and provide practical guidance to would-be users of SCIPEs. Analyzing the data using both approaches is important since, although Hall et al. include these controls in their validation models (and three of the four are statistically significant), neither they nor Hall (2018) includes them in any of the substantive applications/chapters.

Figures 1 and 2 show our results. In figure 1, each of the individual panels corresponds to one of the Big Five traits. The *Y*-axis within each panel identifies which measure of ideology we examine. Each *X*-axis reports the coefficients from our OLS regressions. In order to facilitate apples-to-apples comparisons among the different dependent variables, we standardized all of them prior to estimating each model. SCIPEs are already standardized, so a coefficient of 0.75 means that a one standard deviation increase in a given trait translates into a 0.75 standard deviation increase in the ideology measure. Finally, within the plotting space, we identify estimates for models that do (circles) and do not (squares) include controls for justice sex or birth year. The horizontal whiskers denote the confidence intervals around the coefficient estimates. The thicker lines show the 80% values, the vertical ticks show the 90% values, and the thinner line extends to the 95% values (all two-tailed tests).

Consider the conscientiousness trait. Scholarship shows that it consistently correlates with ideological conservatism both among the mass public and among political elites like state legislators (Dietrich et al. 2012) and members of Congress (Ramey, Klinger, and Hollibaugh 2017).[4] We therefore should expect to see a positive and statistically significant correlation between SCIPE-assessed conscientiousness and conservative ideology.

The SCIPE-assessed conscientiousness measure provides no such correlation, however. In figure 1, we examine three measures that directly tap into ideology, as evidenced in the votes cast by a justice. Only one—the Bailey score—provides even weak evidence of an association. What is more, that weak association only appears when simultaneously controlling for a justice's sex and age. In other words, the measures by themselves are insufficient. Consistent with Hall et al. (2021), we find decently strong evidence of an association with the CBI measure, but this too requires the inclusion of the demographic control variables.

yet exist for Justices Gorsuch and Kavanaugh, we exclude the 184 and 109 votes cast by them, respectively. Gorsuch has written 14 concurrences to date and Kavanaugh has written 10. It is unclear at what point justices will have written enough opinions to be able to estimate their personality.

4. Given the centrality of the conscientiousness trait in our earlier work, we also examined five additional criterion validity dependent variables for SCIPEs. Out of the five, SCIPEs were significantly correlated with only one at the $p < .20$ level (Black et al. 2020, 56). Note that the Hall (2018) SCIPE measures are identical to those reported in Hall et al. (2021), so that existing result is applicable to the scores as republished in Hall et al. (2021). (Our measures were significant at the $p < .10$ level or better for four of the five variables.)

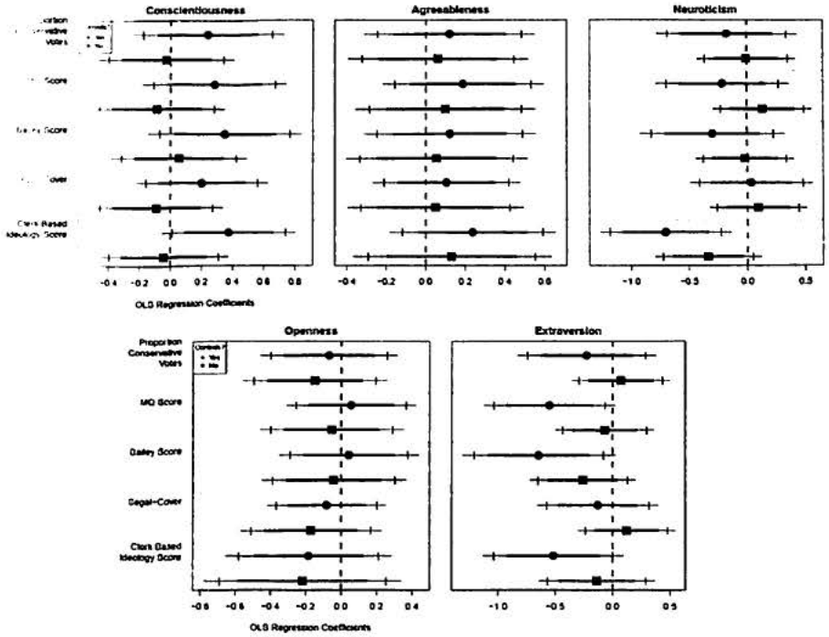

Figure 1. Summary of ideology models with SCIPE trait estimates. The horizontal whiskers report 95% (thin lines), 90% (vertical ticks), and 80% (thick lines) confidence intervals (all two-tailed). See text for additional description.

Next, consider agreeableness. Scholarship on the correlation between ideology and agreeableness is mixed, but Dietrich et al. (2012, 203), who conducted a comprehensive review of roughly two dozen studies, note that "several studies have found [a] modest link . . . between agreeableness and liberal identification." These authors found that more agreeable state legislators were more liberal. Ramey et al.'s (2017) study of Congress found similar results. As figure 1 indicates, however, no such relationship exists between SCIPEs and any of the five measures of judicial ideology. None of the coefficients is signed in the correct direction, and, in any event, all have very wide confidence intervals.

Dietrich et al. (2012) classify neuroticism as having a "modest" negative relationship with ideological conservatism (or, if you prefer, a positive correlation with liberalism). Across five dependent variables and two different model specifications, however, we find only a single result that is statistically significant and in the expected direction. As reported by Hall et al. (2021), the CBI score paired with demographic controls yields a statistically significant coefficient. As our plot shows, however, this effect is weakened both in magnitude and significance level if one omits the demographic control variables. And, of course, none of the four other measures of ideology exhibits any connection with neuroticism.

Next, the trait of openness has, by Dietrich et al.'s (2012, 203) accounting, "consistently . . . been found to be a strong predictor of ideological liberalism." As such, we should

expect to find a negative relationship between openness and conservatism, since all five of our measures are scaled with positive values being conservative. This is not what we see in figure 1, however. Indeed, not a single one of the 10 models produced a result that even approaches statistical significance.

Next, we consider extraversion. Predictions here are probably the most challenging of any of the traits. On one hand, Dietrich et al.'s (2012, 203) comprehensive canvassing of the literature lead them to conclude that this trait "routinely produces null results when included as a predictor of ideology." On the other hand, it is unclear where Hall et al. (2021) put their chips, as they first note that one mass public study found a negative relationship between liberalism and extraversion. They then go on to observe that the analysis of members of Congress by Ramey et al. (2017) found a positive association.

Starting with the five aggregate measures, we find some moderately consistent evidence that extraversion is positively associated with liberalism, with results in the range of statistical significance for a justice's Martin-Quinn score, Bailey score, and CBI score. As with the earlier significant results, however, all of these results are sensitive to the inclusion/exclusion of the demographic control variables. And, given that the Hall et al. measures are derived from the same software as the Ramey et al. study, it is possible this is an artifact of the particular method as opposed to something substantively informative.

We next turn our attention to figure 2, which follows the same approach to predict the ideological direction of justices' votes. It shows coefficient estimates for the SCIPE variables from two logistic regression models. Circles represent models that include controls, and squares denote models that exclude them. We use point color to mark whether a variable is (black) or is not (gray) statistically significant.[5] All of the black points are significant at the .05 level and the two gray points have two-tailed p-values greater than .20.

For conscientiousness, the vote-level results in figure 2 underscore the SCIPE's sensitivity to demographic controls. We observe a strong and statistically significant effect in the expected direction only when the models include demographic controls. That is, increased conscientiousness decreases the probability of a liberal vote only when the models include demographic controls. If the models exclude demographic controls—and contain only the SCIPEs—the coefficient size shrinks nearly to zero and the standard errors correspondingly increase, rendering them statistically indistinguishable from zero.

For agreeableness, the vote-level models in figure 2 also show sensitivity to the inclusion of demographic controls, though here the effect is actually reversed as compared to conscientiousness. A model that includes controls fails to show any significant correlation,

5. Given the panel nature of the vote data, an additional consideration is the sensitivity of the results to using different approaches to calculating standard errors. Hall et al. (2021) report, in the parlance of Zorn (2006), the equivalent of "naive" standard errors, which is to say that they assume conditional independence. The inferences regarding significance are unchanged if one utilizes so-called robust standard errors, or standard errors clustered on the approximately 9,000 cases in the data. If, however, one uses standard errors clustered on justice, then most of the SCIPE coefficients fall out of significance, owing to the much smaller number of observations (only 34).

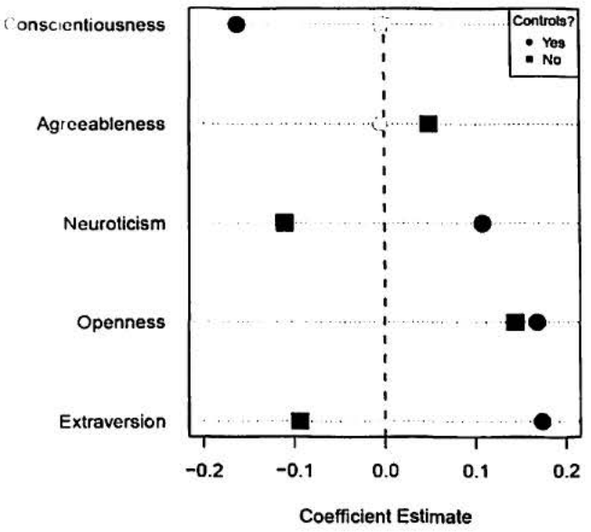

Figure 2. Summary of justice vote models with SCIPE trait estimates. The dependent variable is coded 1 if a justice voted liberally and 0 if he or she voted conservatively in a case. For both models, N = 72,657. Gray points are statistically insignificant ($p > .20$), and black points are significant at the .05 level. Significance calculated using naive standard errors (see n. 5 for additional discussion).

but one that excludes them does. In fact, the coefficient size increases by a factor of 10, changes signs, and becomes statistically significant in the expected direction when excluding the demographic controls.

For neuroticism, in the context of the vote-level models, we again confront a confusing pair of results. The inclusion of demographic controls determines whether the effect of the SCIPE neuroticism estimate is positive and statistically significant (controls included) or negative and statistically significant (controls excluded).

For openness, the results are more encouraging if one considers the individual vote models, where the effect is positive and largely unchanged regardless of how one handles the demographic controls or the calculation of standard errors. Yet, as Hall et al. (2021) themselves note, this is due in no small part to the fact that these models contain more than 70,000 observations.

As to the vote-level analyses on extraversion, we again observe that the sign of the effect depends critically on whether one includes the demographic controls. When they are included, extraversion is positively correlated with casting a liberal vote; when the controls are excluded, however, the effect switches signs, decreases in absolute magnitude, and yet still retains its statistical significance.

Taken together, these findings show that the SCIPEs do not appear to provide valid estimates of Supreme Court justices' personality traits. This is consistent with our previous work, which also shows that the SCIPE trait estimates fail to perform in the way one would

expect across a half dozen nonideological outcome variables.[6] We turn next to offering some data-driven insights into why this might be the case.

THE DANGER OF USING CONCURRING OPINIONS TO ESTIMATE JUSTICES' PERSONALITIES

As a preliminary matter, it is worth emphasizing that we fully agree with the general approach utilized by Hall et al. (2021). That is, a text-based approach is likely to be the optimal (though by no means the only) way of assessing judicial personalities from afar.[7] As Pennebaker et al. (2015, 1) tell us: "The ways people use words in their *daily lives* can provide rich information about their beliefs, fears, thinking patterns, social relationships, and personalities" (emphasis added). Herein lies the challenge. Scholars are generally not privy to the conversations, text messages, and emails political elites like Supreme Court justices generate in their daily lives, and so we need to look elsewhere for that information. Where we part ways from the approach of Hall (2018) and Hall et al. (2021) is in where one ought to go looking to find such information.

Concurrences Primarily Reflect the Court's Majority Opinion

As the sole source of input for their trait estimates, Hall et al. examine the concurring opinions justices have written while serving as a Supreme Court justice. We are skeptical (Black et al. 2020). Perhaps the biggest problem is that Hall et al. (2021) simply ask too much of a single type of document to be able to provide valid insights into personality. An approach that puts all of its eggs in a single basket is likely to be insufficient.

To appreciate the magnitude of the gap between the ideal and Hall et al.'s approach, it is useful to consider the primary sources that the researchers who develop text-to-trait models employ to identify the textual correlates of personality traits. That is, when scholars are trying to discern what fundamental patterns exist, from where do their "daily words" come?

Typically, scholars will employ texts of a more "natural" cast to estimate personality. Personality Recognizer—the approach utilized by Hall et al. (2021)—repurposed a corpus of essays written by undergraduate psychology students who responded to a prompt that solicited their "thoughts, feelings and sensations are at this moment. . . . Your goal in this assignment is to reveal in your writing the way your mind works naturally" (Pennebaker and King 1999, 1301). Concurring opinions, by contrast, seem to us anything but an open-ended opportunity to express one's self. As Corley (2010, 96) notes, concurring opinions often communicate a justice's "understanding *of the majority opinion*" (emphasis added). Similarly, Ray (1990, 783) observes that "by writing separately, a concurring author always offers an internal commentary on the court's judgment." In other words, concurring opinions do not provide an open-ended opportunity for expression. Majority opinions often substantially frame and constrain the content of concurring opinions in a given case.[8]

In defending concurrences, Hall (2018, 38) quotes approvingly from a longtime circuit court judge: "[SCIPE's] focus on concurrences reflects Judge Frank Coffin's description of the 'feeling of unjudicial glee as one shucks off the normal restraint of writing for a panel and proceeds to thrust and parry with gay abandon.'" The same colorful quote also appears in Hall et al. (2021). However, when read with the added context of the paragraph from which the quote originates, one cannot help but doubt whether Coffin was actually talking about concurrences, as Hall and Hall et al. suggest. Here is that quotation along with its source paragraph, which is absent from both Hall and Hall et al.'s presentation of it:

> A concurrence is like a fencing foil; it elegantly makes its usually bloodless points. A dissent, on the other hand, is more like a broadsword. It takes more resolution and commitment to wield it and there is the expectation of drawing at least a little blood. In any event, there is a feeling of unjudicial glee as one shucks off the normal restraint of writing for a panel and proceeds to thrust and parry with gay abandon. For this very reason, we judges are well advised to resist the temptation unless we find a compelling interest and no more effective alternative. Sometimes, however, a dissent is the precise instrument that should be used. [Note: Coffin then goes on to enumerate the five conditions in which he believes a dissent is appropriate.] (Coffin 1994, 227)

Perhaps most dispositive, however, is the fact that the entire paragraph itself—"thrust and parry" quote included—appears in a section that is labeled with the heading "Dissenting Opinions" and not in the distinctive "Concurring Opinions" section on the previous page. Interestingly, what thoughts Coffin does offer about concurring opinions are all consistent with our majority-response view of concurring opinions; that is, they center on how the concurring opinion will expand or limit the majority opinion (Coffin 1994, 226–27).

Even if we accept the Coffin quote as deployed by Hall and Hall et al., it still identifies what we believe to be a critical and inherent limitation of looking only at concurrences (or

6. In particular, we examined how each of the five SCIPE trait estimates performed across six additional outcomes variables, for a total of 30 comparisons. Even with an inclusive definition of consistency, we find that only 10 of the 30 comparisons are consistent with the literature. The trait-by-trait breakdown is conscientiousness, 1/6; agreeableness, 3/6; neuroticism, 3/6; openness, 1/6; and extraversion, 2/6. The appendix (available online) contains additional details and discussion of these results (see also Black et al. 2020, 69–77).

7. A text-based measure is not the only alternative for present-day researchers. One could attempt to get experts to evaluate or rate justices on each personality dimension, as has been attempted numerous times in the psychological literature. However, after doing a thorough meta-analysis of this body of work, Connelly and Ones (2010) found that correlations between self and "other's" personality ratings only added value if the raters were high on interpersonal intimacy (e.g., family members). Thus, this option is highly unlikely to yield better results for Supreme Court justices, almost all of whose families are either very private or would be temporally limited to only recent justices.

8. The approach that built the models we used in our work on personality-leveraged Twitter feeds—perhaps the ultimate open-ended response format—from over 1,500 individuals, though obviously with a character limit for each individual tweet (Black et al. 2020, 35).

dissents for that matter): a justice or judge is not writing in a manner that elicits that jurist's instant "thoughts, feelings, and sensations," but rather is fundamentally constrained by the content, topics, and, to a large degree, the specific words utilized by the majority opinion (i.e., the would-be opponent in Coffin's fencing match).

At any rate, this question is testable. If we are correct, the personality content of concurring opinions should appear, on average, like the majority opinions. Moreover, it should also appear more like the majority opinion than other concurring opinions written by that same justice. In our earlier work, we tested this argument by analyzing opinion-level personality content in the Court's majority opinion and any of the concurring opinions that accompanied it. We then also examined the personality content in pairs of concurring opinions written by the same author. What we found when we compared the two was that the correlation in personality content of majority-concur pairings was at least three times larger than the analogous correlation between pairings of concurring opinions written by the same author (Black et al. 2020). Or, stated a bit differently, opinions written by two different justices in the same case had more alike in terms of their apparent personality than opinions written by the same justice across two different cases.

Here we build upon this previous analysis in two important ways. First, when selecting a different concurring opinion, we originally only constrained the pool of possibilities to be other concurrences written by the same justice. This meant that we could be pairing an economic activity concurrence with one, say, about the First Amendment. Given that the majority and concurring opinion comparison come from the exact same case, there is a guaranteed agreement in terms of issue area. This, in turn, could be inflating the correlations we found between a case's majority and concurring opinions. To address this issue, we now limit the pool of possible concurrences to be those in the same issue area (while still requiring that it was the same justice authoring it).

Second, the results we report in our book come from a single set of random pairings. That is, we took each of the more than 2,700 concurring opinions in our data and randomly paired them with one other concurrence by the same author and computed the resulting correlations for each of the Big Five traits. Justices in our data, however, often authored quite a few concurring opinions (the median is 65). One plausible concern is that by hanging our hat on a single set of results, perhaps we just pulled a particularly "unlucky" draw of concurrences with which to make those calculations. To address this concern, we then repeated this entire process a total of 10,000 times. That is, we paired each of the 2,700 (or so) concurrences with a random one written by the same justice and within the same issue area and calculated the trait correlations. We then did it a second time, and then a third, and so on.

Figure 3 displays the results from this analysis. Each of the five panels represents one personality trait. Within each panel, we show the level of correlation between the majority and concurring opinions in a case with the dashed line. The violin plots in each panel portray the distribution of correlations we recovered from the 10,000 sampling iterations that we performed. The gray area shows the distribution like a kernel density plot. The

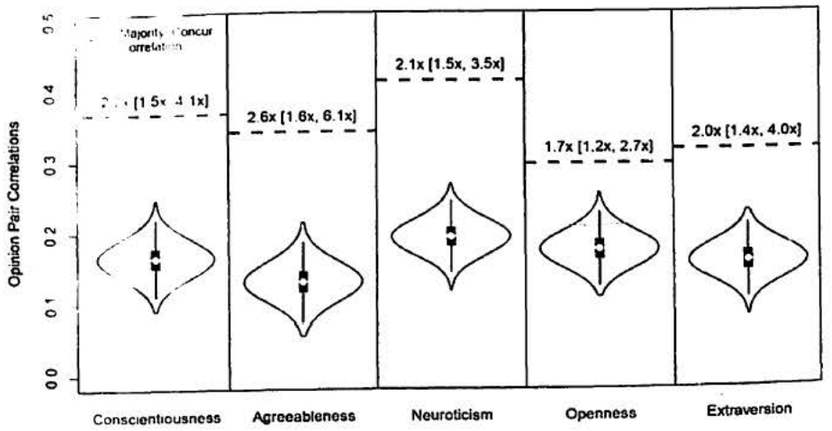

Figure 3. Comparison of opinion-level personality content. The dashed lines report the correlation between the content of a majority opinion and a concurring opinion in the same case. The violin plots show the correlation between the concurring opinion and another, randomly chosen concurring opinion written by the same justice in the same issue area. We repeated this random assignment process a total of 10,000 times. The circles, black rectangles, and vertical whiskers correspond to the median, interquartile range, and minimum/maximum, respectively. Text annotations above the dashed line report how much bigger the majority-concur correlations are compared to the concur-concur correlations.

vertical whisker spans the full range of the 10,000 iterations. The black rectangle identifies the 25th and 75th percentiles. The white circle shows the median. The text annotations above the dashed line show the factor by which the majority-concur correlations exceeds the concur-concur correlations, using the median and, in brackets, the minimum and maximum values for the 10,000 samples.

Median values for the concur-concur correlations range from a low of .14 for agreeableness to a high of .19 for neuroticism, which, to be fair, are elevated relative to the values we found in our initial analysis (they ranged from .08 to .13). That being said, our ultimate conclusion remains the same: the same case pairings continue to overwhelmingly dominate the same justice pairings across the board. Four out of five traits observe correlations that are at least twice as strong as compared to the median result from our 10,000 samples. Importantly, in not one of the 10,000 samples we analyzed did the concur-concur correlations match, let alone surpass the majority-concur correlation in strength. Thus, by even this rather conservative test, we find the fundamental theory that underlies SCIPEs to be wanting.[9]

9. The same-case majority-concur correlation provides a substantively motivated baseline against which we can compare our results. Importantly, our bar is actually (much) lower than what one would expect in previous studies on the retesting of personality traits. Individuals in a Supreme Court justice's

The Impact of a Concurrence's Legal Purpose

The above suggests a general lack of consistency within the corpus of a justice's concurring opinions—even those that were written on the same general topic. Again, this is not especially surprising given what published research on the topic more than a decade ago has already demonstrated: when justices write concurring opinions, they do so for a wide range of substantive legal purposes (Corley 2010). Although this point isn't addressed in Hall (2018), Hall et al. (2021) do acknowledge this—as well as other features—as sources of "potential" bias but then, without providing any assessment of the impact of such a bias, assert that their approach is still valid.

In our earlier work, we already examined the impact of these biases by drawing out a number of comparisons among different types of concurring opinions. Our results suggested that they are not just hypothetical but are both substantial and systematic. Justices vary, for example, in the rate at which they write a regular versus a special concurrence. Not only that, but the trait estimates one would arrive at vary significantly if you examine one type of concurrence versus another (Black et al. 2020). Relatedly, one of the main assertions made by Hall et al. (2021) in defense of concurrences is that their authors lack an incentive to accommodate the requests of other colleagues. Here, too, our previously published work demonstrated that justices vary quite a bit in terms of whether their concurrences are joined by other colleagues; Justice Douglas, for example, wrote for only himself in more than 90% of his nearly 130 concurrences. At the other end of spectrum, Justice Brennan concurred nearly 170 times in his career and the majority of those were joined by one or more of his brethren. And, once again, we find considerable discrepancies in the resulting trait estimates one would produce by using one set of concurrences versus another (Black et al. 2020).

So, it turns out we already know quite a bit about systematic bias in the singular documents that create SCIPEs. Here we build on these existing efforts by probing, in more detail, the importance of the legal purpose of concurring opinions. To do so, we turn to Corley (2010), who identified a total of six different categories of concurrences: doctrinal, emphatic, expansive, limiting, reluctant, and unnecessary.[10] Doctrinal concurrences are what judicial scholars often refer to as "special concurrences." Here, the justice joins the majority's result but not the rationale it uses to reach that result. And so the justice writes a concurring opinion to explain how her rationale differs from the majority's. Emphatic concurrences seek to clarify a particular aspect of the majority opinion. Expansive concurrences and limiting concurrences seek to enlarge or restrict the scope of the majority

age category (i.e., middle to late adulthood) typically generate test-retest correlations that range between 0.50 and 0.80 (Costa and McCrae 1994, cited in Fraley and Roberts 2005, 60). Even this lower threshold is still more than double the magnitude of even the single highest correlation we found in our 10,000 samples. Fraley and Roberts (2005, 61) find evidence of a temporal effect for test-retesting results whereby longer time intervals reduce the correlation, but even a gap of 30 years still yields a correlation of more than 0.50.

10. Our description of these types borrows heavily from Corley (2010, 15–19).

opinion, respectively. Reluctant concurrences indicate that a justice may join the majority opinion, but with reservations. Finally, unnecessary concurrences exist when a justice concurs specially but does not write an opinion clarifying why she disagrees with the majority's rationale.

Corley performs a painstaking content analysis of nearly 300 concurrences written during the Court's 1986–89 terms, coding each concurring opinion as belonging to one of the six previously described categories.[11] Her work further documents considerable variation in each justice's tendency to write or join different types of concurring opinions (Corley 2010, 32). In terms of doctrinal concurrences, for example, just over 45% of Justice Brennan's concurring activity took place in such opinions as compared to only 26% for Justice Kennedy; when it came to expansive concurrences, however, Kennedy showed a stronger preference for those (22% of his activity) as compared to Brennan (only 10% of his activity).

We next seek to ascertain whether this varying mixture of concurrences translates into the same sort of differences we previously uncovered with respect to overall personality trait estimation. Again, the intuition is fairly straightforward: if these different opinions are producing equivalent information about the writer's personality, then there should be a reasonable degree of consistency among the estimates we obtain. To that end, we utilize Corley's coding of the concurring opinions to assess how sensitive estimates of justice personalities are depending on the type of concurring opinion used to generate those inputs.

To do so, we started by gathering the 269 written concurrences coded by Corley. Unnecessary concurrences, by definition, do not include a written opinion, and so we necessarily had to exclude those. We then aggregated all of a justice's opinions by each concurring opinion type, which yielded a total of 43 merged opinions.[12] There is considerable variation in the resulting length of each of these files. Justice Scalia's corpus of doctrinal concurrences weighs in at more than 45,000 words in length. Scalia's body of restrictive concurrences, by contrast, contains just over 400 words. We follow Hall et al. (2021, n. 3) and exclude four justice-concurring combinations that contain fewer than 500 words.[13] In addition to these justice-type pairings, we also generated a single file for each justice that contains all of her concurrences. This is equivalent to the approach utilized by Hall et al., and so we include it for comparison purposes.

Opinion files in hand, we then simply process this corpus of files using Personality Recognizer as utilized by Hall et al. Because we have six different types of inputs (one for each

11. Corley chose these terms because they align with a subsequent qualitative analysis she undertook using the papers of Justices Blackmun and Marshall.

12. This value is lower than the product of the number of justices and concurring opinion types (i.e., $10 \times 5 = 50$) because some justices never wrote a particular type of concurrence. For example, neither Justice Powell nor Justice O'Connor ever wrote a reluctant concurrence.

13. The four justice/type pairings are Kennedy-Restrictive, Scalia-Restrictive, Powell-Expansive, and Rehnquist-Restrictive. We do retain one pairing that just barely misses the 500-word mark: Scalia-Emphatic, which has 497 words.

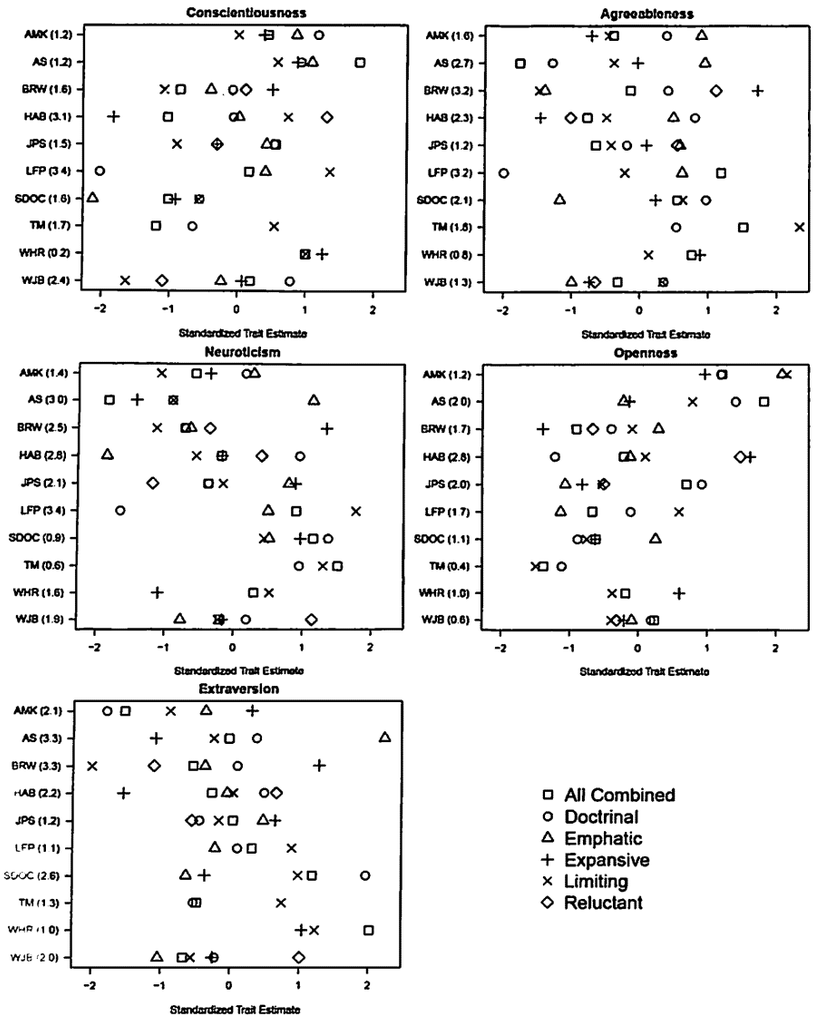

Figure 4. Personality traits as estimated by different concurring opinion types

concurrence type plus one containing all concurrences), we have a total of six different sets of estimates for each of the Big Five traits for the 10 justices in the Corley data. Figure 4 plots all of these estimates. Each of the five panels corresponds to an individual trait. The *X*-axis shows the standardized trait score. Within the plot, we use different symbols to identify what specific inputs were used to generate that estimate.[14] The *Y*-axis identifies

14. One might notice that some symbols appear to be "missing" from the plots; this is due to a justice not having written a sufficient number of concurrences in that type for Personality Recognizer to be able to generate a valid estimate.

each justice's initials. The parenthetical values located after those initials report the absolute value of the difference between the smallest and largest trait estimate for that justice. Consider Justice Blackmun (HAB) and the conscientiousness trait. Using the six different types of inputs, we estimated six different conscientious scores for him, which are, in ascending order: −1.8 (expansive), −1.0 (all combined), −0.1 (doctrinal), 0.0 (emphatic), 0.7 (limiting), and 1.3 (reluctant). It is worth reiterating that these scores are all standardized, and so a score of −1.8 means Blackmun is estimated to be 1.8 standard deviations below the mean and a score of 1.3, by contrast puts him 1.3 standard deviations above the mean for the trait. The absolute difference between these two extremes is 3.1 units, which is quite large.

Even a casual eyeballing of these results shows that most justices seem to bounce around in relation to their colleagues. Moreover, this movement based on input does not follow a consistent or predictable pattern. For example, in terms of his conscientiousness, Blackmun's emphatic concurrences (the triangle) evince considerably more conscientiousness than do his reluctant concurrences (the diamond). The opposite is true, however, for Justice Brennan (WJB), whose reluctant concurrences show higher conscientiousness than his emphatic ones. And, as the annotations by their initials reveal, Brennan, much like Blackmun, also observes a lot of variation in his concurrence-assessed conscientiousness (2.4 units between the minimum and maximum).

To provide more of a macro-level assessment, however, we turn to figure 5, which visualizes absolute-value differences, like those of Blackmun's and Brennan's we just

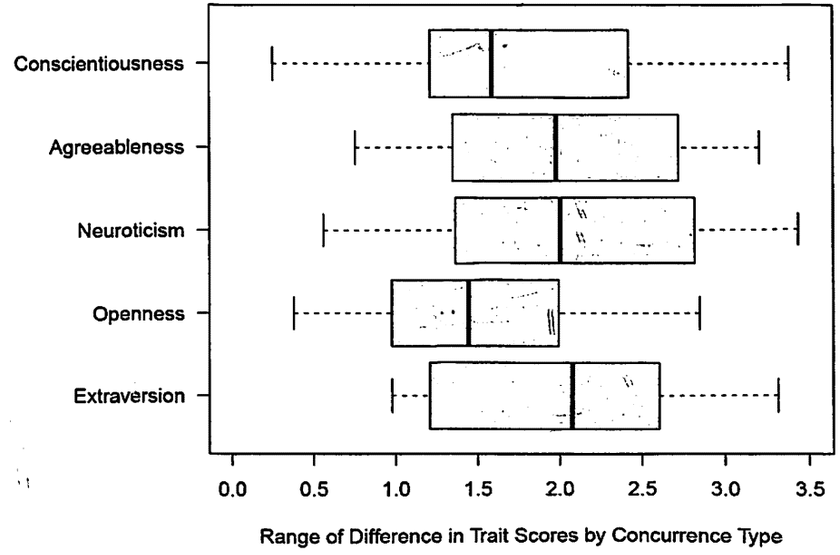

Figure 5. Box plots of the absolute value differences in each justice's concurrence-type trait estimate. Each plot portrays 10 data points, which are identified in the parentheticals of figure 4.

described. In that sense, it provides us with an opportunity to examine whether the "potential for such bias" that Hall et al. (2021) acknowledge is simply potential as opposed to realized. Each of the Big Five traits appears along the *Y*-axis. The values portrayed in each of the box plots come from the parentheticals in figure 4. That is, for the conscientiousness box plot, we include the 3.1 value for Blackmun and the 2.4 value for Brennan plus the other eight justices in the data.

Within each box plot, the thicker black line denotes the median value and the shaded gray area identifies the interquartile range of these values. Of particular note, the median amount of difference between a justice's smallest and largest estimates for a trait is never less than 1.4 standard deviations. Another consistent feature of these results is that there are large gaps for all 10 of the justices in the data. To wit, all justices have at least one trait where their minimum/maximum gap is at least one standard deviation in magnitude. Eight of the 10 justices have one or more gaps that are at least two standard deviations, and fully four of those eight also have one or more gaps that are three or more standard deviations. This indicates a high degree of trait sensitivity driven solely by the documents included in the estimation process, which itself suggests that these different documents are, consistent with the literature on concurrences, in fact very distinctive.

WHAT KIND OF PERSONALITY MEASURES WOULD BE APPROPRIATE OR IDEAL?

We have raised serious questions about the personality measures from Hall (2018) and Hall et al. (2021). But the next question is, what kind of text-based measures would be appropriate to employ? As a baseline, it would be ideal to generate personality measures for all of the justices under consideration. We recognize, however, there may be some limitations that make it challenging to produce measures on all justices. Hall et al. (2021) are up-front about not being able to generate measures for Chief Justice Vinson and Justice Minton because they did not produce enough concurrences. While being forced to omit a few of the justices is not the end of the world—after all, this happens to most everyone studying judicial behavior at some point in time (ourselves included)—it is, however, peculiar when Hall et al. later argue against using alternative sources of justices' language because "these alternative sources pose serious availability problems" (353). In reality, however, this is not much of a challenge. In our earlier efforts, we gathered enough alternative texts to estimate personality traits for all the justices during their timespan, including Vinson and Minton. And, in preparing this article we were easily able to update our estimates to include newly appointed justices. Thus, this serious problem appears to be not so serious after all.

We also wish to highlight several important features of the ideal measurement strategy. First, the personality estimates should be exogenous to the justices' opinions written while on the Supreme Court. Scholars should want to avoid a circular measurement procedure. We would not want to use justice behavior to explain justice behavior, which is circular and inappropriate.[15] As the authors of the leading study on the estimation of Supreme Court justices' ideal points remind us, "The circularity concern [of using votes-based ideal points to explain votes] is quite important as a purely technical matter. Strictly speaking, the scores should not be used in this context" (Martin and Quinn 2005, 2). Similarly, Ho and Quinn (2010, 847–48) state:

> [If] ideal point estimates are derived from the same votes being modeled in the regression, such models are circular. Estimates of other effects become uninterpretable from a causal perspective. Votes are used to explain votes . . . [a problem that] continues to plague empirical studies of judicial voting.

Much like Segal and Cover's (1989) decision three decades ago to craft an exogenous indicator of judicial ideology, researchers ought to employ texts that are exogenous to the behavior scholars are most likely to want to use the resulting personality measures to explain. "One cannot demonstrate that attitudes affect votes when attitudes are operationalized from those same votes" (558). Thus, scholars seeking to estimate the personalities of political actors should take pains to ensure the texts they employ do not come from the same behaviors they intend scholars to examine later.

Second, the text inputs should predominantly reflect the sole work of the justices. Of course, due to the justices' involvement in government and politics, the things they say and write over their lifetimes are often for particular audiences. This creates the problem that some document types are likely to reveal more about personality traits than other types (Hall 2018, 36). This tension can be successfully addressed (*a*) by collecting texts that come from justices' prenomination speeches and writings (or even lower court opinions when other document types are in short supply) and (*b*) by collecting as much textual data as possible under conditions where the justices are not strictly beholden to the same organizational, group, or societal interest.[16]

To satisfy those two prongs suggests there is considerable virtue in collecting texts from various sources as a way to account for the fact that personality traits will be more or less visible under different situations. The key, however, to making this strategy successful is that there needs to be an appropriate standardization process in place where documents with more constraints are weighed accordingly. For example, the standardization process should be able to account for the fact that less personality can be extracted from a lower court opinion compared to a speech, presumably because there are more constraints on a lower court opinion. Our existing approach does so by comparing the personality content of actual texts with ones generated completely at random (Black et al. 2020, 38–41).

15. Note, though, that one may alleviate this problem with sophisticated structural models that simultaneously estimate justices' personality traits and the dependent variable of interest (Martin and Quinn 2005).

16. There is also the issue that some justices produced more speeches and writings than others, though this is not a significant issue for the vast majority of justices.

Third, the estimation process should be able to add quickly any newly seated justices. By relying on concurring opinions as the input method, Hall et al. (2021) must wait a significantly long time before adding new personality estimates. This point becomes sharper when considering how quickly changes happen to the Court's membership; as of this writing, we have observed three new justices (Gorsuch, Kavanaugh, and Barrett) taking the bench in a span of four years. Having to wait a little while for scores seems reasonable, since a justice's first term or two will not give scholars much data with which to work. But Hall et al. provide scant practical guidance about when a justice will have written enough concurrences to fully reveal his or her personality to scholars. An ideal approach is one that is able to produce an estimate shortly after a nominee is announced—an approach that is consistent with the standard Segal and Cover (1989) adopted.

Fourth, scholars ought to employ the most recent developments in text-to-trait technology and continue to refine their measures as far as technological advances will allow. For example, the recent approach designed by IBM—Watson Personality Insight (WPI)—provides advantages for the task of text-driven personality recognition. WPI infers personality traits from text based on an open-vocabulary approach (rather than a closed-vocabulary approach) using a word-embedding technique called GloVe (Global Vectors for Word Representation) to obtain a vector representation of the input texts, which then feeds into a model that uses an algorithm that was trained on thousands of individuals who provided both text and answers to personality surveys (IBM 2017). This approach is an improvement for multiple reasons. Perhaps most importantly, text-based personality measurement with an open-vocabulary approach outperforms a closed approach (Schwartz et al. 2013). This method lets the underlying data decide which individual words, multiword phrases, and overall topics best predict an individual's personality traits. It creates trait estimates with greater accuracy and efficiency than older approaches (Arnoux et al. 2017).

Unfortunately, as is always the case with a reliance on any cutting-edge technology, today's hot fad eventually becomes yesterday's old news. For example, Black et al. (2020) employed WPI, but that program will soon be sunsetted. So while the processes used to generate personality estimates by Black et al. are valid (as are the conscientiousness scores themselves), they will eventually need to be replaced when scholars seek to examine more justices.[17] Fortunately, research at the intersection of natural language processing, machine learning, and personality psychology is taking place at a highly vigorous rate. In one recent example, Obschonka et al. (2020) discuss how to estimate personality traits from an available dataset established by Schwartz et al. (2013) that contains over 70,000 Facebook users who also completed a personality survey (and made their posts publicly

17. This highlights a trade-off between the benefit of accessibility to an open-source process with inferior estimates versus a black-box methodology that produces superior estimates.

available).[18] This provides a viable roadmap for entrepreneurial scholars to build the next generation of personality estimates (see also Kern et al. 2014; Park et al. 2015).

Fifth, an ideal estimate of personality traits also would contain some measure of uncertainty or standard error of the estimate. The judicial equivalent would be moving from Segal-Cover scores, which only contain ideological point estimates, to Martin-Quinn scores, which have standard errors. Having standard errors would be valuable not only from a measurement standpoint but also from a substantive point of view. It could contribute to a better understanding of how much variation there is in personality traits across situations and time.

Finally, but most importantly, any set of measures should be subjected to a thorough validation before proceeding. One ought not expect that future research will be uniformly successful at tapping into all of the Big Five traits. It could be, for example, that agreeableness is particularly difficult to pin down. Because we do not have self-report measures and we cannot use reports by others, that means the primary method of validation is to compare how well measures predict or correlate with various behaviors with which the traits are known to correlate. Importantly, reliance on just one or two behavioral indicators drawn from a single concept like ideology is not optimal, especially when existing work on the topic is far from clear about what relationships one should expect. We have previously illustrated this point when validating the measure of conscientiousness by examining its relationship to other indicators beyond ideology (Black et al. 2020).[19] Relying on only a couple of indicators is a risky proposition, as we demonstrate above with our reassessment of ideology.

AN APPLICATION AND A TEST

So far, we have identified a number of issues with SCIPEs, both in terms of their construction as well as their resulting empirical validity. We have also sketched out some thoughts about how an alternative approach could address some of these concerns. In this section, we illustrate the pitfalls of applying the SCIPEs to studies of judicial decision making, and to compare them to measures that are better, we apply them to an analysis of the Supreme Court's agenda-setting process. We examine whether conscientiousness makes justices more likely to vote to grant review when there is conflict among the circuit courts. Conscientiousness is particularly important to the practice of judging.[20] As part of this inquiry,

18. While one might object to treating text from Facebook posts as similar to speeches and writings, it is important to keep in mind that the WPI approach used Twitter text and Personality Recognizer used the written ramblings of undergraduate students. Facebook posts cover a wide range of contexts and are shown to be good predictors of all sorts of social and political characteristics.

19. See n. 6, where we summarize SCIPE's poor performance in this regard.

20. The American Bar Association's (ABA) Standing Committee on the Federal Judiciary calls for Supreme Court justices that "possess exceptional professional qualifications" such as "industry and diligence . . . intellectual capacity, judgment, writing and analytical abilities, knowledge of the law" and other related characteristics (see https://www.americanbar.org/content/dam/aba/uncategorized/GAO/Backgrounder.authcheckdam.pdf). The ABA's Canons of Judicial Ethics state that judges must be conscientious (see https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/pic_migrated/1924_canons.authcheckdam.pdf).

we compare the results of empirical models that include the Hall et al. (2021) SCIPE scores versus our own (Black et al. 2020).

For substantive background, there are two kinds of legal conflict that are relevant at the Court's agenda-setting stage and also to this analysis: strong conflict and weak conflict. Strong conflict exists when there is a square conflict between or among lower courts and the conflict does not appear to be clearing itself up. In other words, if the same legal question has come up in multiple circuits and those circuits have reached opposing answers, there is strong legal conflict involving that issue. In such a case, we would expect all justices to be sensitive to the conflict and vote to resolve it (e.g., Black and Owens 2009).

Weak conflict is another matter. Here, there is tension among lower court decisions, but the nature of the conflict is qualitatively different from strong conflict. The need to address it is perhaps not as pressing—at least for most justices. The two circuit court decisions purportedly at loggerheads might not address an identical legal question. This means a conflict might be characterized as "shallow," or indirect. Or, it could be that the circuit courts appear to be working out the conflict on their own, by virtue of one answer to a legal question gaining favor over another and moving the circuits toward uniformity.

Conscientious justices will treat petitions with weak legal conflict more seriously than do less conscientious justices (Black et al. 2020). Conscientious people take their professional obligations seriously. Conscientious workers "are predisposed to be organized, exacting, disciplined, diligent, dependable, methodical, and purposeful . . . [they] thoroughly and correctly perform work tasks [and] take initiative in solving problems" (Witt et al. 2002, 164). One of the justices' central duties is to ensure uniformity in the law. It is a task that the Court—and only the Court—can accomplish. In cases with strong conflict, all justices will be inclined toward granting review. But it is in cases with weak legal conflict that the conscientious justice's heightened level of duty and problem-solving initiative should stand out. In sum, we expect that high-conscientious justices are more likely than low-conscientious justices to vote to grant review to cases that present such legal conflict.

Following Black and Owens (2009) and Black et al. (2020), we test this hypothesis with a random sample of 360 paid, non-death-penalty petitions appealed from the federal court of appeals that made the Court's discuss list during the 1986–93 terms.[21] From these 360 petitions we recovered a total of 3,024 individual justice votes. The data on the justices' votes originate from the digital images of Justice Blackmun's docket sheets, which we retrieved from Epstein, Segal, and Spaeth (2007).

21. We sample petitions from the Court's discuss list because these are petitions that have a nonzero probability of being granted, since at least one justice deemed it worthy of some discussion. We examine only petitions from federal courts of appeals because current data allows ideologically estimable comparisons only between Supreme Court justices and lower federal court judges. We exclude capital petitions because they were treated differently than their noncapital counterparts during the time period of our study. The Court automatically added capital cases to the discuss list. Once there, Justices Brennan and Marshall always vored to grant the petition, vacate the death penalty, and remand the case (Woodward and Armstrong 1979; Lazarus 2005).

Dependent Variable. The dependent variable, Grant, measures whether each justice cast a vote to grant (1) or deny (0) review to each certiorari petition in the sample.

Conscientiousness. We focus on each justice's conscientiousness; larger values correspond to greater conscientiousness. We explore the empirical results using the SCIPE scores from Hall et al. (2021) and our own indicators (Black et al. 2020). Our personality measures are derived using the text of a justice's published articles, public speeches, and lower court opinions penned prior to his or her confirmation. We translated these texts into personality trait scores using WPI while introducing a novel standardization process to account for differences in personality content across document type.[22] For this article, we have updated these measures to incorporate Justices Gorsuch, Kavanaugh, and Barrett. We also control for justices' scores on the four other personality trait dimensions: openness, extraversion, agreeableness, and neuroticism.

Legal Conflict. We create four dummy variables to measure the presence and extent of lower court conflict within the case. We derive these measures from reading the cert pool memo in each case.[23] Strong Conflict represents instances when the pool memo writer acknowledges a clear and deep split among the lower courts. Weak Conflict is present when the law clerk, while assessing the presence of alleged conflict, suggests that immediate review may not be necessary. Alleged Conflict occurs if the petitioner in the case alleges a conflict among lower courts but the pool memo writer denies the existence of this conflict. No Conflict is the baseline category in our models, and it represents instances where the petitioner did not allege any legal conflict among the lower courts.

With these four binary variables in hand, we then interact them with conscientiousness. We also control for multiple factors that previous research has shown to influence the policy and legal motivations behind Supreme Court agenda setting (see, e.g., Black and Owens 2009). (See the appendix for full coding details.)

22. We describe this process in extensive detail in Black et al. (2020, 38–41), but here's the succinct version: we first created randomly generated documents from each type of document from which we could compare the raw personality estimates. We then used these random documents to generate document-level trait scores and calculated the document-type mean and standard deviation for each of the Big Five traits. This provided a baseline estimate of how much personality is present when the documents were "written" randomly by "someone" with absolutely no meaningful personality (i.e., our computer; sorry, computer). We then used these values to standardize the raw scores for our actual corpus of preconfirmation texts. The resulting standardized scores identify observed personality above and beyond documents of the same type generated in the absence of meaningful personality.

23. It should be noted that this approach is similar to the one utilized by Caldeira and Wright (1988), who used law students to assess the presence of actual conflict in cert petitions. Our approach, however, has two added advantages. First, the cert pool memos are the actual materials used by the justices in the cert pool. And Black and Owens (2009) conducted an intercoder reliability study, indicating that all measures were reliable using common metrics. Second, our approach avoids confirmation bias. Caldeira and Wright (1988) had law students assess conflict after the Court had decided to grant (or deny) review to a petition.

METHODS AND RESULTS

We employ logistic regression models with robust standard errors, estimating two models each for our versus the SCIPE indicators: (1) a baseline traits-only model that specifies the interaction terms between conscientiousness and each legal conflict variable, along with the four other personality traits; and (2) a full model specification that includes all control predictors. (The appendix reports the table of regression results.) Figure 6 reports the average marginal effects (with 95% confidence intervals) of strong and weak legal conflict (compared to the baseline of no conflict) across the range of conscientiousness.

First consider the analysis using our own personality indicators. Both models support our theoretical expectations. The impact of legal conflict varies significantly based on conscientiousness, and in a way that is most evident among petitions with weak legal conflict. That is, weak legal conflict exhibits its greatest impact on the most conscientious justices in the sample and no meaningful impact on the least conscientious justices.

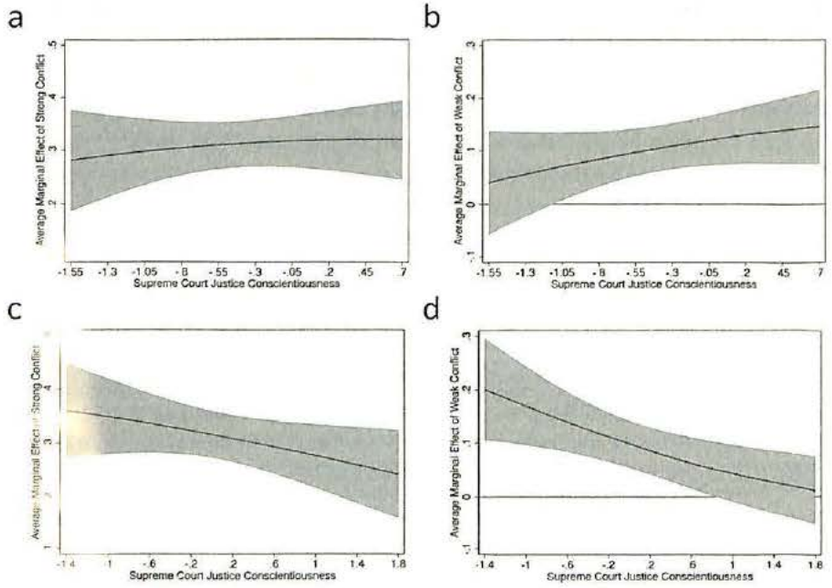

Figure 6. The conditional impact of lower court legal conflict and conscientiousness on Supreme Court justices' votes to grant certiorari. Average marginal effects of Strong Conflict (*a*) and Weak Conflict (*b*), with 95% confidence intervals, across the range of Conscientiousness using our updated personality indicators (i.e., results from model 2). Panels c and *d*, Same average marginal effects but using the Hall et al. (2013) SCIPE scores (i.e., results from model 4).

Figure 6*a* shows, as expected, that the impact of Strong Conflict is always statistically significant and positive, and its magnitude increases ever so slightly across the range of our conscientiousness measure. That is, when confronted by a petition that conveys a strong degree of lower court conflict (compared to one with no conflict), all justices are generally much more likely to seek to grant cert. Increasingly conscientious justices are only somewhat more likely to do so than less conscientious justices.

Figure 6*b* is, perhaps, the most important figure for our application. It highlights how justices treat petitions with weak legal conflict. As the figure makes clear, weak conflict fails to move justices with lower conscientiousness (using our conscientiousness indicator). In other words, the least conscientious justices do not meaningfully distinguish between weak conflict and no conflict. However, as a justice's conscientiousness increases, he or she becomes significantly more likely to grant review when confronted by weak conflict. Indeed, when conscientiousness is at the 90th percentile, a justice is 0.14 more likely to vote to grant review to a petition with weak conflict (0.27) as compared to one where no conflict (0.14) exists—roughly twice the likelihood of review. This provides considerable support, in a theoretically sensible manner, for the role of conscientiousness in addressing conflict at the agenda stage (at least if using our personality indicators).

Our results diverge significantly from those we retrieve when using the SCIPEs and Hall's (2018) original analysis, however. Hall hypothesizes that "more-conscientious justices are less likely to cast a grant vote because pursuing policy objectives violates their judicial duty" (55). His results purport to show this. But there are two major problems with this claim. First, the hypothesis assumes that granting review somehow automatically triggers the pursuit of policy objectives in violation of the judicial duty. It does nothing of the sort. It simply puts the case on the Court's docket (assuming three others also vote to grant) for the Court to decide. What happens next is within the control of the justices. At this point, justices can pursue legal or policy objectives (or both simultaneously). Second, the logical implication of this statement is that conscientious justices—those who are strongly motivated by duty and obligation—would not grant review to any cases. The idea that the justices who are most responsible, most dutiful, and most rule-abiding would simply decline to hear any cases is not sensible. Perhaps most obviously, it is well established that one of the primary duties of the Supreme Court is to reduce lower court conflict. Suggesting a conscientious justice would not grant cert because it violates their judicial duty creates a major internal conflict for the justice because it would mean they could not fulfill one of the most important duties of the Supreme Court: unifying the law for the lower courts.

Figure 6*c* shows the average marginal effect of Strong Conflict (with 95% confidence intervals), compared to the baseline of no conflict, across the range of conscientiousness using the SCIPE indicators. The figure shows that the impact of Strong Conflict is statistically significant and positive, but its magnitude decreases across the range of conscientiousness. That is, when confronted with a petition that conveys a strong degree of lower court conflict (compared to one with no conflict), all justices are generally likely seek to

grant certiorari, but the SCIPE measures suggest that more conscientious justices may be somewhat less likely to do so.

Figure 6*d* shows the opposite interactive effect from figure 6*b*. In figure 6*d*, the SCIPE measures suggest that weak conflict shapes the impact of conscientiousness in a counter-intuitive manner. That is, weak conflict matters less when viewed by a more conscientious justice, and such conflict ultimately exhibits no impact at all among the most conscientious justices. By contrast, when SCIPE conscientiousness is low (at the 10th percentile), a justice is 0.19 more likely to vote to grant review to a petition with weak conflict (0.424) as compared to one where no conflict (0.237) exists. Taken together, the results using the SCIPEs are incompatible with the premise that duty compels the conscientious justice to seek to fulfill the Supreme Court's foremost agenda-setting task: to resolve lower court conflict.

DISCUSSION

Recently, Epstein and Knight (2013, 13) sounded an alarm for judicial politics scholars, arguing: "Only by updating our theories and empirics to develop a more complete vision of judging will we continue to remain players in a field that is now more vibrant than ever." We could not agree more. As far as we are concerned, approaches that fail to treat judges and justices as "regular" individuals—including their personalities—are of limited use to fully understanding judicial behavior. But, for these types of studies to thrive, scholars must continue to be attentive to their measurements. As we have demonstrated here, however, the Hall et al. (2021) SCIPE measures that form the foundation for Hall (2018) are severely limited.

But what should one do instead so as to stay "in the game," as Epstein and Knight ask us to do? On the basis of the evidence presently available, we believe our existing approach, presented in Black et al. (2020), provides a valid indicator of the conscientiousness trait. These measures exist for 41 justices, including the three most recent appointees: Gorsuch, Kavanaugh, and Barrett. As to the other Big Five traits, we are frankly less convinced that any measure is ready yet for prime time—ours included. To establish that, one would need to gather a variety of validation indicators specifically tailored for the trait of interest.

Combining different disciplines often produces influential and informative new theories that alter the direction of those respective disciplines. That promises to be the case with the union of personality scholarship and judicial politics. We must be careful, however, to employ measures equal to the task. The SCIPE measures have critical flaws that limit their usefulness. But appropriate measures, following useful parameters, can and will move the discipline forward.

REFERENCES

Arnoux, P.-H., A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha. 2017. "25 Tweets to Know You: A New Model to Predict Personality with Social Media." *AAAI Conference on Web and Social Media*, 472–75.

Bailey, M. A. 2007. "Comparable Preference Estimates across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51 (3): 433–48.

———. 2013. "Is Today's Court the Most Conservative in Sixty Years? Challenges and Opportunities in Measuring Judicial Preferences." *Journal of Politics* 75 (3): 821–34.

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

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

Bonica, A., A. S. Chilton, J. Goldin, K. Rozema, and S. Maya. 2017. "Measuring Judicial Ideology Using Law Clerk Hiring." *American Law and Economics Review* 19 (1): 129–61.

Braman, E. 2009. *Law, Politics, and Perception: How Policy Preferences Influence Legal Reasoning*. Charlottesville: University of Virginia Press.

Braman, E., and T. E. Nelson. 2007. "Mechanism of Motivated Reasoning? Analogical Perception in Discrimination Disputes." *American Journal of Political Science* 51 (4): 940–56.

Caldeira, G. A., and J. R. Wright. 1988. "Organized Interests and Agenda Setting in the U.S. Supreme Court." *American Political Science Review* 82 (4): 1109–27.

Coffin, F. M. 1994. *On Appeal: Courts, Lawyering, and Judging*. New York: Norton.

Collins, P. M. 2011. "Cognitive Dissonance on the U.S. Supreme Court." *Political Research Quarterly* 64 (2): 362–76.

Connelly, B. S., and D. S. Ones. 2010. "An Other Perspective on Personality: Meta-analytic Integration of Observers' Accuracy and Predictive Validity." *Psychological Bulletin* 136 (6): 1092.

Corley, P. C. 2010. *Concurring Opinion Writing on the U.S. Supreme Court*. Albany, NY: SUNY Press.

Costa, P. T., Jr., and R. R. McCrae. 1994. "Set Like Plaster? Evidence for the Stability of Adult Personality." In *Can Personality Change?* 21–40. Washington, DC: American Psychological Association.

Dietrich, B. J., S. Lasley, J. J. Mondak, M. L. Remmel, and J. Turner. 2012. "Personality and Legislative Politics: The Big Five Trait Dimensions among U.S. State Legislators." *Political Psychology* 33 (2): 195–210.

Epstein, L., and J. Knight. 1998. *The Choices Justices Make*. Washington, DC: CQ Press.

———. 2013. "Reconsidering Judicial Preferences." *Annual Review of Political Science* 16:11–31.

Epstein, L., W. M. Landes, and T. H. R. A. Posner. 2013. *The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice*. Cambridge, MA: Harvard University Press.

Epstein, L., A. D. Martin, K. M. Quinn, and J. A. Segal. 2007. "Ideological Drift among Supreme Court Justices: Who, When, and How Important?" *Northwestern University Law Review* 101 (4): 1483–542.

Epstein, L., J. A. Segal, and H. J. Spaeth. 2007. "Digital Archive of the Papers of Harry A. Blackmun." http://epstein.law.northwestern.edu/research/BlackmunArchive/.

Fraley, R. C., and B. W. Roberts. 2005. "Patterns of Continuity: A Dynamic Model for Conceptualizing the Stability of Individual Differences in Psychological Constructs across the Life Course." *Psychological Review* 112 (1): 60.

Hall, M. E. 2018. *What Justices Want: Goals and Personality on the US Supreme Court*. New York: Cambridge University Press.

Hall, M. E., G. E. Hollibaugh Jr., J. D. Klinger, and A. J. Ramey. 2021. "Attributes beyond Attitudes: Measuring Personality Traits on the US Supreme Court." *Journal of Law and Courts* 9 (2): 345–70.

Ho, D. E., and K. Quinn. 2010. "How Not to Lie with Judicial Votes: Misconceptions, Measurement, and Models." *California Law Review* 98:813–76.

IBM. 2017. "The Science behind the Service." IBM Watson Personality Insight Documentation.

Kern, M. L., J. C. Eichstaedt, H. A. Schwartz, L. Dziurzynski, L. H. Ungar, D. J. Stillwell, M. Kosinski, S. M. Ramones, and M. E. Seligman. 2014. "The Online Social Self: An Open Vocabulary Approach to Personality." *Assessment* 21 (2): 158–69.

Lazarus, E. 2005. *Closed Chambers: the Rise, Fall, and Future of the Modern Supreme Court.* Reissue ed. New York: Penguin.

Martin, A. D., and K. M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10 (2): 134–53.

———. 2005. "Can Ideal Point Estimates Be Used as Explanatory Variables?" Unpublished manuscript, Washington University of St. Louis. http://adm.wustl.edu/supct/resnote.pdf.

Moyer, L. 2012. "The Role of Case Complexity in Judicial Decision Making." *Law and Policy* 34 (3): 291–312.

Obschonka, M., N. Lee, A. Rodríguez-Pose, J. C. Eichstaedt, and T. Ebert. 2020. "Big Data Methods, Social Media, and the Psychology of Entrepreneurial Regions: Capturing Cross-County Personality Traits and Their Impact on Entrepreneurship in the USA." *Small Business Economics* 55 (3): 567–88.

Owens, R. J., and J. Wedeking. 2012. "Predicting Drift on Politically Insulated Institutions: A Study of Ideological Drift on the U.S. Supreme Court." *Journal of Politics* 74:487–500.

Park, G., H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman. 2015. "Automatic Personality Assessment through Social Media Language." *Journal of Personality and Social Psychology* 108 (6): 934–52.

Pennebaker, J. W., R. L. Boyd, K. Jordan, and K. Blackburn. 2015. "The Development and Psychometric Properties of LIWC2015." Technical report. http://www.liwc.net.

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

Ramey, A. J., J. D. Klinger, and G. E. Hollibaugh Jr. 2017. *More than a Feeling: Personality, Polarization, and the Transformation of the U.S. Congress.* Chicago: University of Chicago Press.

Ray, L. K. 1990. "The Justices Write Separately: Uses of the Concurrence by the Rehnquist Court." *University of California Davis Law Review* 23:777–831.

Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, and L. H. Ungar. 2013. "Personality, Gender, and Age in the Language of Social Media: The Open Vocabulary Approach." *PLOS One* 8 (9): 1–15.

Segal, J. A., and A. D. Cover. 1989. "Ideological Values and the Votes of Supreme Court Justices." *American Political Science Review* 83 (2): 557–65.

Segal, J. A., L. Epstein, C. M. Cameron, and H. J. Spaeth. 1995. "Ideological Values and the Votes of Supreme Court Justices Revisited." *Journal of Politics* 57 (3): 812–23.

Spaeth, H. J., L. Epstein, T. Ruger, K. Whittington, J. A. Segal, and A. D. Martin. 2010. The Supreme Court Database. St. Louis, MO: Washington University in Saint Louis. http://scdb.wustl.edu/index.php.

Witt, L., L. A. Burke, M. R. Barrick, and M. K. Mount. 2002. "The Interactive Effects of Conscientiousness and Agreeableness on Job Performance." *Journal of Applied Psychology* 87 (1): 164–69.

Woodward, B., and S. Armstrong. 1979. *The Brethren: Inside the Supreme Court.* New York: Simon & Schuster.

Zorn, C. 2006. "Comparing Gee and Robust Standard Errors for Conditionally Dependent Data." *Political Research Quarterly* 59 (3): 329–41.