5-12-2021

A Test-Retest Reliability Generalization Meta-Analysis of Judgments Via the Policy-Capturing Technique

Ze Zhu
University of Nebraska at Omaha, zezhu@unomaha.edu

Alan J. Tommassetti
CPS HR Consulting

Reeshad S. Dalal
George Mason University

Shannon W. Schrader
George Mason University

Kevin Loo
George Mason University

See next page for additional authors

Recommended Citation

Zhu, Z., Tomassetti, A.J., Dalal, R.S., Schrader, S.W., Loo, K., Sabat, I.E., Alaybek, B., Zhou, Y., Jones, C., & Fyffe, S. (2021). A test-retest reliability generalization meta-analysis of judgments via the policy-capturing technique. Orgnaizational Research Methods, 25(3), 541-574. https://doi.org/10.1177/10944281211011529

Authors
Ze Zhu, Alan J. Tommassetti, Reeshad S. Dalal, Shannon W. Schrader, Kevin Loo, Isaac E. Sabat, Balca Alaybek, You Zhou, Chelsea Jones, and Shea Fyffe

A Test-Retest Reliability Generalization Meta-Analysis of Judgments Via the Policy-Capturing Technique

Ze Zhu[1], Alan J. Tomassetti[2], Reeshad S. Dalal[1], Shannon W. Schrader[1], Kevin Loo[1], Isaac E. Sabat[3], Balca Alaybek[1], You Zhou[1], Chelsea Jones[1], and Shea Fyffe[1]

Abstract

Policy capturing is a widely used technique, but the temporal stability of policy-capturing judgments has long been a cause for concern. This article emphasizes the importance of reporting reliability, and in particular test-retest reliability, estimates in policy-capturing studies. We found that only 164 of 955 policy-capturing studies (i.e., 17.17%) reported a test-retest reliability estimate. We then conducted a reliability generalization meta-analysis on policy-capturing studies that did report test-retest reliability estimates—and we obtained an average reliability estimate of .78. We additionally examined 16 potential methodological and substantive antecedents to test-retest reliability (equivalent to moderators in validity generalization studies). We found that test-retest reliability was robust to variation in 14 of the 16 factors examined but that reliability was higher in paper-and-pencil studies than in web-based studies and was higher for behavioral intention judgments than for other (e.g., attitudinal and perceptual) judgments. We provide an agenda for future research. Finally, we provide several best-practice recommendations for researchers (and journal reviewers) with regard to (a) reporting test-retest reliability, (b) designing policy-capturing studies for appropriate reportage, and (c) properly interpreting test-retest reliability in policy- capturing studies.

Keywords

policy capturing, test-retest reliability, reliability generalization, judgment analysis, meta-analysis

---

[1]Department of Psychology, George Mason University, Fairfax, VA, USA
[2]CPS HR Consulting, Sacramento, CA, USA
[3]Department of Psychology, Texas A&M University, College Station, TX, USA

Corresponding Author:

Ze Zhu, Department of Psychology, George Mason University, 4400 University Drive, MSN 3F5, Fairfax, VA 22030, USA. Email: zzhu5@gmu.edu

Policy capturing (also known as judgment analysis) is a long-standing survey technique that examines how people weigh different pieces of information when making judgments (Zedeck, 1977).[1] Policy capturing has been used in a wide array of contexts, including job choice (Rynes et al., 1983), union voting (Leigh, 1986), receptivity to various forms of advice (Dalal & Bonaccio, 2010), promotion decisions (Viswesvaran et al., 1994), and counselor judgments of client acute suicide lethality (Brown, 1972). However, although the policy-capturing technique has been used widely for over four decades, until now, no reliability generalization study has been conducted. In much the same way as researchers have questioned the reliability of novel survey techniques—with some of these techniques ultimately proving more reliable (e.g., situational judgment tests; Catano et al., 2012) than others (e.g., most Rorschach-based tests; Garb, 1999)—in the current article, we examine the average test-retest reliability of the policy-capturing survey technique across studies as well as methodological and substantive factors that may influence reliability estimates. We focus specifically on test-retest reliability because the error source most relevant to policy capturing is temporal (in)stability over the course of the (potentially very long) policy-capturing measure (Aiman-Smith et al., 2002; Cooksey, 1996; Karren & Barringer, 2002). Because reliability is necessary (although not sufficient) for validity (Cronbach, 1988; Sireci & Sukin, 2013), the validity of conclusions from policy-capturing studies cannot be accepted uncritically without demonstrating that policy-capturing judgments are stable over time.

One explanation for the lack of a policy-capturing reliability generalization study may be the low rate at which estimates of policy-capturing reliability have historically been reported. In a 2002 policy-capturing review and tutorial published in *Organizational Research Methods*, Karren and Barringer found that only a very small number of policy-capturing studies had reported any index of reliability. Therefore, Karren and Barringer—as well as a companion piece in the same journal by Aiman-Smith et al. (2002)—advocated that policy-capturing studies routinely report estimates of, in particular, test-retest reliability. Given the passage of time since the publication of these best- practice tutorials, sufficient test-retest reliability estimates now exist to facilitate a reliability generalization meta-analysis. In addition to investigating the average test-retest reliability estimate reported in policy-capturing studies, the present meta-analysis examines the degree to which these reliability estimates generalize (vs. vary) across methodological choices made by the authors of policy-capturing studies and across substantive factors studied in policy-capturing studies.

Estimating Reliability in Policy-Capturing Studies

In policy-capturing studies, decision-makers make judgments in response to a series of scenarios or profiles across which researchers have manipulated the levels of several cues (e.g., present vs. absent or high vs. low cue levels). Decision-makers' judgment "policies" are then "captured" by regressing their judgment(s) in response to the scenarios onto the combinations of cue levels present in those scenarios. This permits an assessment of the extent to which decision-makers used the cues in reaching their judgments. For example, Tomassetti et al. (2016) examined how a decision-maker's willingness to accept a job is influenced by levels (high vs. low) of six job or organization characteristics (e.g., pay level and schedule flexibility). Figure 1 provides two example scenarios out of the 64 focal scenarios used in the Tomassetti et al. study.

Because decision-makers must make a relatively long series of judgments in policy-capturing studies, the stability of their judgment policies is a concern. Test-retest reliability is designed to assess temporal stability and is therefore the preferred method for assessing reliability in policy- capturing studies (Aiman-Smith et al., 2002; Karren & Barringer, 2002). The current research therefore focuses on test-retest reliability. However, because some policy-capturing studies have reported a reliability estimate other than (or in addition to) test-retest reliability, we discuss the appropriateness of various reliability measures, including test-retest reliability, in Table 1.

To use the test-retest method in policy-capturing studies, a few of (or all) the policy-capturing scenarios are repeated such that they are presented to the participants twice. Participants' judgments across the two iterations of these scenarios are correlated to generate a test-retest reliability correlation coefficient for each decision-maker (within-person or idiographic reliability) and/or for each repeated scenario (between-person or nomothetic reliability). These coefficients are generally averaged across all the decision-makers or scenarios before being reported. It should be noted that although duplicate scenarios are obviously required for the estimation of test-retest reliability, they are unnecessary in the focal policy-capturing analyses: namely, estimating decision-makers' judgment policies. One iteration of the duplicate scenarios (generally, the second iteration) is therefore removed from the data set after assessing test-retest reliability but prior to the focal analyses.

The Current Study

The present reliability generalization study focuses on policy-capturing studies that reported test- retest reliability, as recommended by the Karren and Barringer (2002) policy-capturing tutorial. At least at the time the tutorial was published, however, reportage of test-retest reliability was rare. Our first research question therefore pertains to the overall extent of reportage of test-retest reliability estimates.

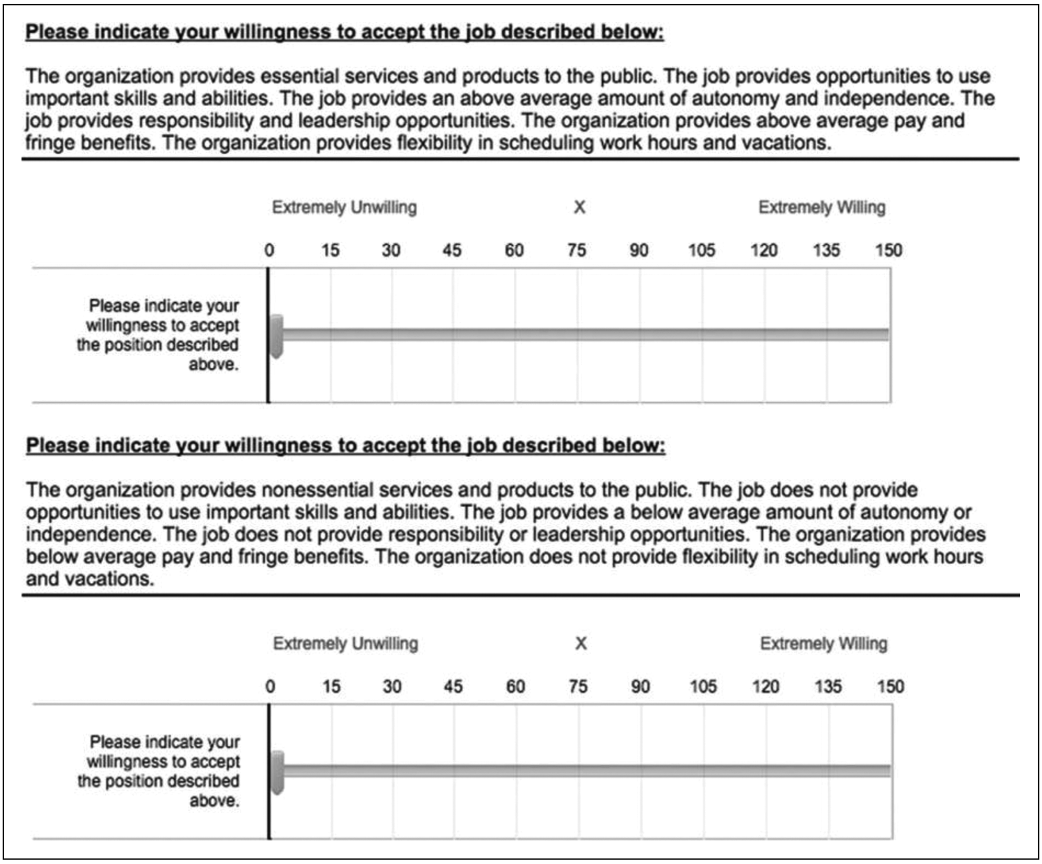

Figure 1. Two example policy-capturing scenarios (from Tomassetti et al., 2016).
Note: The first (or top) policy-capturing scenario contains all the cues with positive wording, and the second (or bottom) scenario contains all the cues with negative wording.

Research Question 1: What percentage of policy-capturing studies report a test-retest reliability estimate?

Next, we turn to the focal effect size in this reliability generalization meta-analysis: the test-retest reliability correlation coefficient. Specifically, what is the average test-retest reliability in policy- capturing studies that do report reliability information? This question is important because if policy-capturing studies do not exhibit high test-retest reliability, they cannot accurately capture decision-makers' judgment policies.

Table 1. Different Types of Reliability Estimates and Their Appropriateness for Policy-Capturing Studies.

Reliability	Appropriate for Policy-Capturing Technique?	Reason(s)
Cronbach's alpha	No, unless multiple judgments (i.e., responses) are required per scenario and these judgments must subsequently be aggregated into a composite judgment	• With a policy-capturing study, a researcher would not expect all the judgments made by a specific decision-maker *across* policy-capturing scenarios to have high communalities (communality ¼ the proportion of variance in one judgment accounted for by all the other judgments), and therefore Cronbach's alpha would be inappropriate. • However, in the relatively infrequent case in which a policy-capturing study involves multiple judgments (vs. one) *within* each policy-capturing scenario, and if these judgments are furthermore assumed to represent indicators of the same underlying construct, Cronbach's alpha could be used to assess the appropriateness of aggregating these judgments into a single composite judgment per scenario (Cortina, 1993).
Kuder-Richardson coefficient of equivalence	No	• A Kuder-Richardson coefficient of equivalence is inappropriate because policy-capturing response scales are generally not dichotomous (Aiman-Smith et al. 2002)—in other words, because the responses elicited from decision-makers are generally judgments rather than choices.
Interrater reliability	No, except possibly within clusters of judgment policies	• Policy-capturing researchers do not necessarily expect two decision-makers to make the same judgments, even given identical scenarios. This is evident in the frequent use of cluster analysis in policy-capturing studies (e.g., Dalal & Bonaccio, 2010), wherein researchers expect multiple clusters of judgment policies. Therefore, indices of interrater—as opposed to intrarater or test-retest—reliability (LeBreton & Senter, 2008) are inappropriate, except possibly within a cluster of judgment policies.
Test-retest reliability	Yes	• The error source most relevant in policy-capturing measures is that of temporal (in)stability over the course of the (potentially very long) policy-capturing measure (Aiman-Smith et al., 2002, Cooksey, 1996; Karren & Barringer, 2002). Stated differently, the extent to which the decision-maker uses a temporally stable judgment policy across scenarios in the study is the primary reliability-based concern. Therefore, the use of a test-retest Pearson product-moment correlation is appropriate. • In cases where not just reliability but also absolute agreement is of interest, a test-retest equivalent to r_{wg} (LeBreton & Senter, 2008; see also Berchtold, 2016) could also be included.

Research Question 2: What is the average test-retest reliability in policy-capturing studies?

Potential Methodological Antecedents of Test-Retest Reliability

Another important question pertains to the extent to which test-retest reliability estimates in policy-capturing studies generalize (vs. vary appreciably) across numerous methodological choices made by authors of policy-capturing studies. We grouped these potential methodological antecedents to reliability (equivalent to moderators in validity generalization studies) into three sets: (a) general study and sample characteristics, (b) scenario characteristics, and (c) design characteristics.

General Study Characteristics. General study characteristics are factors common to almost all social science research. An example is the year of publication, which we use to determine if more recent studies—with the advantage of having more existing empirical studies to model and more published tutorials to follow—on average yield higher reliability than older studies. Other general study characteristics include sample type (i.e., student vs. nonstudent samples[2]), journal impact factor, and sample size.[3]

Scenario Characteristics. Scenario characteristics describe how the policy-capturing scenarios were constructed. The first subset of scenario characteristics involves those characteristics impacting study length. Concerns about the length of the survey are often motivated by the expectation that participants' cognitive resources or attention will decrease over the course of a survey. Research on vigilance decrement suggests that performance deteriorates as humans attempt to maintain continued attention on a specific task (Davies & Parasuraman, 1982; Mackworth, 1948). Similarly, ego depletion theory holds that performance deteriorates when regulatory resources are depleted (Baumeister & Heatherton, 1996; Hagger et al., 2010). In survey contexts, this expectation has been verified by self-reports of attention waning toward the middle or end of long surveys (Baer et al., 1997; Meade & Craig, 2012). The characteristics that determine the length of a policy- capturing study are factors that policy-capturing tutorials have identified as areas to which researchers should pay particular attention when designing studies. For example, with regard to recommendations regarding the ideal number of scenarios for a policy-capturing study, Rossi and Anderson (1982) recommended no more than 60 scenarios, whereas Cooksey (1996) suggested that up to 100 scenarios are acceptable. With regard to the number of cues per policy-capturing scenario, Aiman-Smith et al. (2002) recommended no more than five cues, whereas Karren and Barringer (2002) recommended no more than one fifth as many cues as scenarios. Based on the vigilance decrement effect and ego depletion theory, keeping track of a large number of scenarios or a large number of cues per scenario may reduce the test-retest reliability of policy-capturing judgments. However, the ego depletion phenomenon has recently been called into question (e.g., Carter et al., 2015; Hagger et al., 2016). Moreover, it is possible that as the number of scenarios in a study and/or the number of cues in a scenario increase, decision-makers compensate by paying attention to only a very small number of cues—thereby minimizing fatigue. We therefore examine the impact of the number of scenarios and the number of cues per scenario on reliability in an exploratory manner.

The second subset of scenario characteristics is scenario presentation characteristics. These characteristics have the potential to be related to reliability estimates because scenarios with different presentation characteristics require different levels of cognitive effort to process and under- stand them. Specifically, the fundamental cognitive experience of a policy-capturing study involves a decision-maker understanding the cues (and cue levels) in a scenario and forming a judgment related to that scenario. As such, the key components to determining the cognitive demand of a scenario are (a) how easy or difficult the cue presentation format is to process and (b) how easy or difficult the changes in cue levels from one scenario to the next are to identify. Accordingly, we focus on two scenario presentation characteristics: cue presentation format and attention to cue levels.

Regarding cue presentation formats, cues are generally presented as (a) images (e.g., pictures, drawings), (b) tables and/or graphs, and (c) text only.[4] Research on "scene gist" shows that individuals can rapidly (i.e., with minimal cognitive effort) and accurately extract meaning from a visual scene (Friedman, 1979; Li et al., 2002). Additionally, research on the "picture superiority effect" suggests that images are easier to remember and recall than written words because the dual encoding of images (i.e., images can be coded both as images and in verbal form) produces a more effective memory trace for response retrieval (Bevan & Steger, 1971; Paivio, 1969, 1971). Applied to policy-capturing studies, this means not only that decision-makers may understand what is being presented in an image-based scenario (e.g., cues presented in images, tables, or graphs) with less cognitive effort than if it were purely text-based but also that they may encode the information needed for judgment while using fewer cognitive resources than if the same scenario were purely text-based. Therefore, we examine whether cue presentation formats influence policy-capturing reliability.

Next, within the scenarios category (image, table/graph, or pure text), some formats specifically direct a decision-maker's attention to the cue levels that change from scenario to scenario. Research on attention suggests that such formats may require fewer resources than formats that do not direct attention to changes across scenarios (Egly et al., 1994; Rensink, 2002). Thus, making policy- capturing judgments in scenarios where the portions of the scenarios that change across scenarios are highlighted by the researchers (e.g., boldface font) or are inherently prominent (e.g., the only value in a table's row) may require less cognitive effort than when the change is not indicated. We therefore examine whether test-retest reliability varies as a function of whether changes across cue levels are highlighted.

Design Characteristics. Design characteristics describe methodological choices made by researchers in designing the study. One general example is survey medium—that is, pencil and paper or online— because there has been substantial debate as to the psychometric quality of data captured online (and not in front of the researcher) versus data captured in person via paper and pencil (Heerwegh, 2009; Marta-Pedroso et al., 2007). In-person paper-and-pencil studies conceivably exert pressure on decision-makers via a Hawthorne effect (Roethlisberger & Dickson, 1939), thus leading to more stable judgments. We therefore examine whether test-retest reliability differs for online versus paper-and-pencil studies.

An example specific to policy-capturing studies is study design. In a full factorial design, all factors (i.e., cues) are fully crossed and balanced. In contrast, confounded factorial designs, which include block designs and fractional factorial designs, involve systematically dividing the full factorial set into blocks (e.g., halves, quarters, eighths) and presenting each participant with one of the blocks (Graham & Cable, 2001; Karren & Barringer, 2002).[5] Compared to block designs and fractional factorial designs, the corresponding full factorial designs are longer. Karren and Bar- ringer (2002) emphasized the association between survey length and increases in participant stress and exhaustion, raising concerns about survey length resulting from a full factorial design (Graham & Cable, 2001). We therefore examine whether test-retest reliability differs across study designs.

Another study design characteristic is the time gap between the first and second iterations of the repeated scenarios. When the two iterations are in separate sessions with an intervening gap of days or weeks, decision-makers may deliberately or unwittingly (e.g., due to forgetfulness) change their judgment policy across sessions—thereby potentially leading to lower test-retest reliability estimates than in same-session designs. Conversely, however, compared to the corresponding same-session designs, the surveys in *each* session of different-session designs are shorter, thereby leading to lower cognitive load and, potentially, higher test-retest reliability. Thus, we explore whether test-retest reliability differs as a function of the time gap between iterations of repeated scenarios.

The last design characteristic is level of analysis for test-retest reliability. In policy-capturing studies, the test-retest reliability estimates can be calculated at the within-person (idiographic) level and/or at the between-person (nomothetic) level. To calculate within-person reliability, each participant's judgments are correlated across the initial and repeated versions of the repeated scenarios. Each participant has a test-retest reliability estimate, and the mean test-retest reliability across all participants is reported.[6] At the between-person level, test-retest reliability is calculated for each repeated scenario by separately correlating the scores on the initial and repeated version of the scenarios across participants. Then, the mean test-retest reliability across all repeated scenarios is reported. It is worth examining whether the level of analysis influences test-retest reliability for two reasons. First, test-retest reliability estimates at the within- and between-person levels of analysis provide different information (i.e., the reliability for each participant across repeated scenarios vs. the reliability for each repeated scenario across participants, respectively). Second, in general, there is considerable interest in the extent to which results from the between- person level of analysis generalize to the within-person level (Dalal et al., 2014; Molenaar & Campbell, 2009). Hence, we examine whether test-retest reliability is a function of levels of analysis.[7]

Overall, we examine the extent to which test-retest reliability in policy capturing generalizes (vs. varies appreciably) across the aforementioned methodological antecedents.

Research Question 3: Which methodological characteristics significantly influence test-retest reliability estimates in policy-capturing studies?

Potential Substantive Antecedents of Test-Retest Reliability

We also examine, in an exploratory manner, the extent to which test-retest reliability estimates in policy-capturing studies generalize (vs. vary appreciably) across two substantive antecedents: topic area and judgment type. Topic area includes organizational behavior and human resources (OBHR) research versus other—that is, non-OBHR—research. Judgment type consists of four types of judgment: (a) attitude (i.e., participants' latent disposition or tendency to respond with some degree of favorableness to a psychological object), (b) perception (i.e., the process of interpreting, selecting, and organizing objective information), (c) behavioral intention (i.e., indication of a person's readiness to perform a behavior), and (d) mixed or undeterminable (Fishbein & Ajzen, 2009).

Table 2. Identification of Studies and Selection Criteria.

Search Method	Search Scope	Search Terms
Database search	PsycINFO, ProQuest Dissertation Abstracts Online, ABI/INFORM COMPLETE online, and Web of Science databases	"policy capturing," "judgment analysis," and "judgement analysis"
Conference programs	The Academy of Management (AOM), American Educational Research Association (AERA), American Nurses Association (ANA), American Psychiatric Nurses Association (APNA), Association for Psychological Science (APS), British Psychological Society (BPS), Brunswik Society, Society for Industrial and Organizational Psychology (SIOP), Society for Judgment and Decision Making (SJDM), and Society for Personality and Social Psychology (SPSP)[a]	"policy capturing," "judgment analysis," and "judgement analysis"
Ancestry (backward) search	All coded articles from the aforementioned database and conference program searches as well as seven seminal articles: three seminal policy-capturing tutorials (i.e., Aiman-Smith et al., 2002; Graham & Cable, 2001; Karren & Barringer, 2002), one tutorial on experimental vignette studies (Aguinis & Bradley, 2014), two meta-analyses on the lens model (Karelaia & Hogarth, 2008; Kaufmann et al., 2013), and one review article on the test-retest reliability of professional judgment (Ashton, 2000)	"policy capturing," "judgment analysis," and "judgement analysis"
Descendent (forward) search	The aforementioned seven seminal articles	"reliability" or "retest"
Emails to listservs	AOM's Organizational Behavior division listserv, AOM's Research Methods division listserv, and the SJDM listserv	NA

Note: NA ¼ not applicable.
[a]We chose these conference programs based on the areas that emerged in the database search results. Specifically, we coded the research area of each included primary study from the database search. After identifying the areas where policy-capturing designs are used, we found the major conferences in each research area. Last, we searched the conference programs using the search terms in the table to find potential articles.

Research Question 4: Which substantive characteristics significantly influence test-retest reliability estimates in policy-capturing studies?

Method

Identification of Studies and Selection Criteria

To locate policy-capturing studies that reported test-retest reliability, we conducted a database search, a conference program search, and a search of articles that were "ancestors" and "descendants" of already located articles. Additionally, we sent emails to three listservs to request relevant unpublished research. See Table 2 for more details.

To evaluate an empirical study's relevance for the current meta-analysis, we examined the study's method section to determine whether the authors had in fact used a policy-capturing design. The flow diagram summarizing the study identification and evaluation process (Appelbaum et al., 2018; Moher et al., 2009) can be seen in Figure 2.

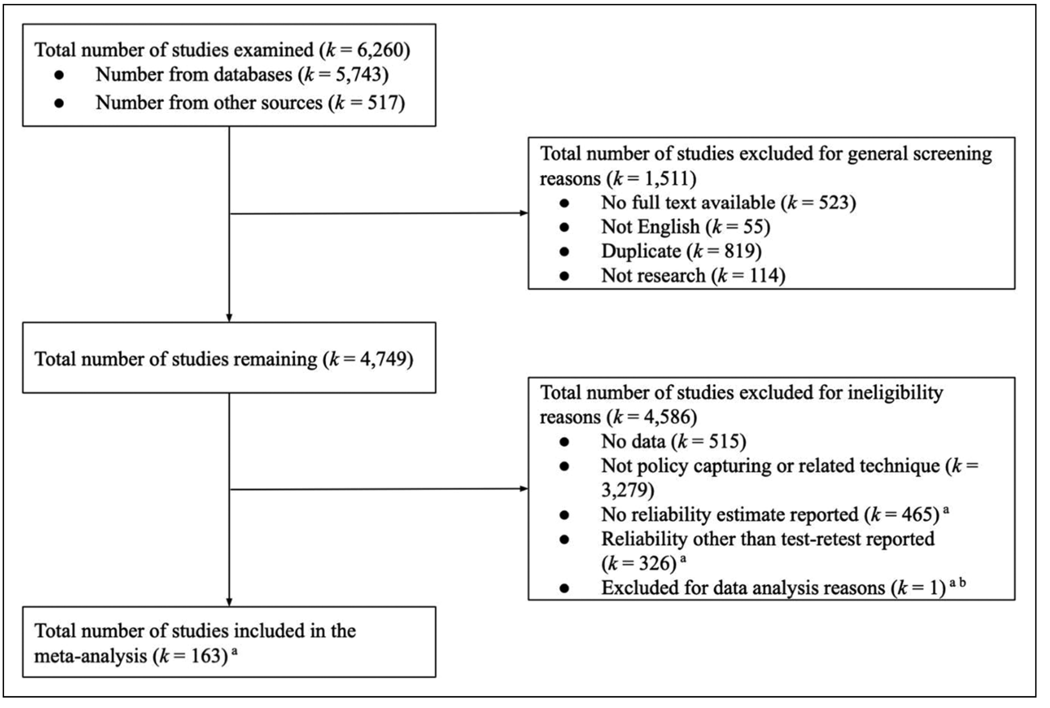

Figure 2. Flow diagram of studies through the meta-analysis inclusion process.

[a]The total number of empirical policy-capturing papers is 465 þ 326 þ 1 þ 163 955, which is used as the denominator to calculate the percentage of policy-capturing studies that reported a test-retest reliability estimate.

[b]One policy-capturing study (Ebert & Kruse, 1978) that reported a test-retest reliability estimate was not included in the meta-analysis due to the small sample size. The formula for calculating the inverse variance weight used in calculating the mean effect size and in the weighted least squares regression involves the N 3 standard error of the reliability correlation. The equation for the standard error is [illegible], where N is the sample size in the study. The sample size of the Ebert and Kruse (1978) study was three, meaning that the denominator in the standard error formula was zero and therefore that the standard error for this study was undefined. As a result, this study was excluded from further analysis. Ebert and Kruse reported a test-retest reliability correlation of .93.

Coding of Studies

We developed a coding manual to help us code test-retest reliability and potential antecedents. Interrater agreement for two independent coders was 85.44% across an initial set of 16 studies and 90.31% across a second set of 21 studies—the latter comparable to other meta-analyses (e.g., B. J. Hoffman et al., 2015; Knight & Eisenkraft, 2015). The coding manual was refined after each stage of the interrater agreement process, and a single coder used the final version (provided in the online supplementary materials) to code the remaining studies while consulting a second coder regarding particularly challenging coding decisions.

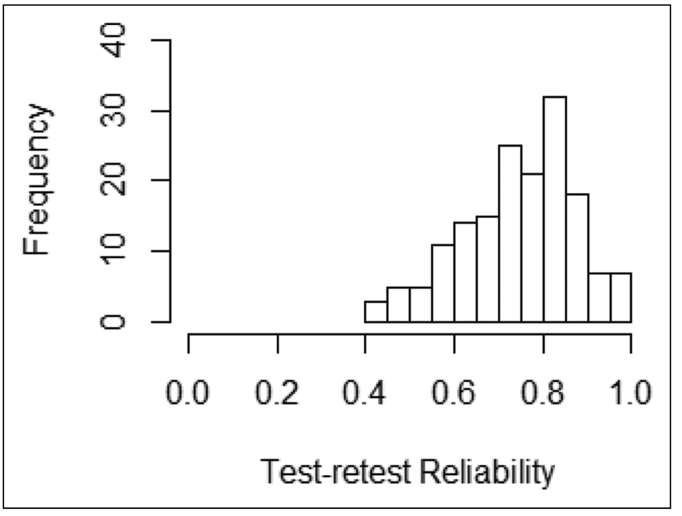

Figure 3. Histogram depicting the frequency of various levels of test-retest reliability.

Data Analysis

A random-effects meta-analysis model (Hedges & Olkin, 1985; Schmidt & Hunter, 2015) was used to estimate the population reliability of judgments made in policy-capturing studies. The effect size used in the analyses was the Fisher *z*-transformed test-retest reliability correlation between the original and repeated scenarios. The mean effect size analyses and the antecedent analyses were run in SPSS Version 19 (IBM Corp., 2010), using the MeanES and MetaReg macros (Lipsey & Wilson, 2001; Wilson, 2006), respectively, as well as the metafor package (Viechtbauer, 2010) in R 3.6.1. Weighted least squares (WLS) regression was applied for the antecedent analyses. For categorical (vs. continuous) antecedents, antecedent categories with fewer than 10 observed cases (see the rows marked “NA” in Table 4) were omitted prior to analysis. In practice, this resulted in all categorical antecedents except judgment type being reduced to dichotomous variables for the analyses. For example, for survey medium, the analysis compared only paper-and-pencil (in-person) text versus web-based (online) text because additional categories (e.g., audio and video) were excluded due to low observed frequencies.

Results

Test-Retest Reliability Reportage

Vis-a`-vis Research Question 1, we found that 17.17% (164 out of 955) of policy-capturing studies reported test-retest reliability estimates. Thus, although the number of policy-capturing studies reporting test-retest reliability estimates is large enough to conduct a reliability generalization study, the percentage of the total is low despite recommendations provided in policy-capturing tutorials (Aiman-Smith et al., 2002; Karren &

Barringer, 2002).

To examine trends in reportage over time, we furthermore calculated a point-biserial correlation between year of publication and whether test-retest reliability estimates were reported in the publication. Results showed a statistically significant but small tendency for more recent publications to be more likely to report test-retest reliability (r = .10, p= .002). Additionally, Table 3 depicts the frequencies (and percentages) of reportage as well as the means and standard deviations of test-retest reliability estimates reported in studies published before versus after the Aiman- Smith et al. (2002) and Karren and Barringer (2002) policy-capturing tutorials were published. The percentage of policy-capturing studies that actually reported test-retest reliability estimates was significantly higher after (20.39%) than before (14.23%) the tutorials were published: z = 2.70, p = .007. Interestingly, the mean test-retest reliability in studies published after the tutorials (M_{weighted} = 0.76, SD_{rho} =0.12) was actually lower than that in studies published before the tutorials (M_{weighted} = 0.80, SD_{rho}= 0.10), t(161) = –2.06, p = .041. Finally, the standard deviation of the test-retest reliability did not differ across studies published before versus after the tutorials, Levene's F(1,161) = 0.03. p =.864. These results suggest that the policy-capturing tutorials were associated with a beneficial effect on reportage of test-retest reliability per se. Moreover, the policy-capturing tutorials did not appear to lead authors to avoid reporting low test-retest reliability estimates: Doing so would presumably have manifested as higher reported means and truncated standard deviations in studies published after (vs. before) the tutorials, which is not what we found.

Mean Level of Test-Retest Reliability

Across 163 independent samples and 20,244 participants, the mean meta-analytic effect size for test- retest reliability was r = .78 (95% CI [0.75, 0.80], SE = 0.03), thereby addressing Research Question 2. Figure 3 shows the distribution of test-retest reliability estimates in primary studies. It should be noted that one policy-capturing study (Ebert & Kruse, 1978) that reported a test-retest reliability estimate was not included in the meta-analysis due to the small sample size (for details, see the note under Figure 2). For the meta-analysis, this reduced the k from 164 to 163.

Table 3. Frequencies, Proportions, Means, and Standard Deviations of Test-Retest Reliability in Articles Published Before Versus After the 2002 Policy-Capturing Tutorials (i.e., Aiman-Smith et al., 2002; Karren & Barringer, 2002).

Year	Number of Policy-Capturing Studies That Reported Test-Retest Reliability	Total Number of Policy-Capturing Studies	Percentage of Policy-Capturing Studies That Reported Test-Retest Reliability	Mean of Reported Test-Retest Reliability	Standard Deviation of Reported Test-Retest Reliability
Up to 2002 (including 2002)	71	499	14.23%	0.80	0.10
After 2002 (since 2003)	93	456	20.39%	0.76	0.12
Total (all years)	164[a]	955	17.17%	0.78	0.12

Note: The "Total" row provides the overall average test-retest reliability estimate for policy-capturing studies (mean r¼.78).
[a]One policy-capturing study (Ebert & Kruse, 1978) that reported a test-retest reliability estimate was included to compute the frequencies and percentages but was not included in the meta-analytic results (means and standard deviations) due to the small sample size. The formula for calculating the inverse variance weight used in calculating the mean effect size and in the weighted least sqaures regression involves the standard error of the reliability correlation. The equation for the standard error is $\frac{1}{\sqrt{N-3}}$ where N is the sample size in the study. The sample size of the Ebert and Kruse (1978) study was three, meaning that the denominator in the standard error formula was zero and therefore that the standard error for this study was undefined. As a result, this study was excluded from further analysis. Ebert and Kruse reported a test-retest reliability correlation of .93. For the meta-analytic results reported in subsequent tables, the exclusion of the Ebert and Kruse (1978) study reduced the k from 164 to 163.

Potential Antecedents to Test-Retest Reliability Estimates

We used three methods to test for heterogeneity across primary studies: (a) the omnibus Q test for heterogeneity (Lipsey & Wilson, 2001), (b) the standard deviation of the effect sizes, corrected for sampling error (i.e., SD_{rho}; Schmidt & Hunter, 2015), and (c) the 80% credibility interval (Koslowsky & Sagie, 1993). The omnibus Q test was statistically significant ($Q = 2{,}519.21, df = 162, p < .001$), the SD_{rho} was greater than zero ($SD_{rho} = 0.12$), and the 80% credibility interval (i.e., [0.63, 0.93]) was wider than the 0.11 rule of thumb proposed by Koslowsky and Sagie (1993). All three methods therefore suggested heterogeneity. Because of this and because we had specified the potential antecedents a priori (Schmidt & Hunter, 2015), we proceeded with antecedent analyses (equivalent to moderator analyses in validity generalization studies).

Examining Potential Antecedents to Test-Retest Reliability. To answer Research Questions 3 and 4, we conducted WLS regression analyses to test whether methodological and substantive characteristics influence reliability estimates in policy-capturing studies (see Tables 4 and 5). However, with one exception (described in the following paragraph), we tested each antecedent separately—that is, one at a time—to maximize the k (number of independent samples) in each analysis because the k varied dramatically across antecedents and because the listwise k in a model containing multiple antecedents was often quite low.

Table 4. Descriptive and Meta-Analytic Statistics for Categorical Putative Antecedents to Test-Retest Reliability.

Antecedent Type	Categorical Putative Antecedent	Category	Frequency (Percentage)	Mean Test-Retest Reliability Estimate (*SE*)[a]	95% CI	SD_{rho}[b]	80% CV[b]	b (*SE*)[c]
General study and sample characteristics	Sample type[d]	Student	53 (32.50%)	0.77 (0.05)	[0.73, 0.81]	0.14	[0.59, 0.95]	
		Nonstudent	101 (62.00%)	0.78 (0.04)	[0.75, 0.81]	0.11	[0.64, 0.92]	
		Mixed[e]	9 (5.50%)	NA	NA	NA	NA	
		Total	163 (100.00%)					0.04 (0.06)
Scenario characteristics	Cue presentation format[d]	Text	94 (63.10%)	0.78 (0.04)	[0.75, 0.81]	0.12	[0.63, 0.93]	
		Table and/or graph	50 (33.60%)	0.78 (0.05)	[0.73, 0.81]	0.12	[0.63, 0.93]	
		Image	4 (2.70%)	NA	NA	NA	NA	
		Video	1 (0.70%)	NA	NA	NA	NA	
		Total	149 (100.00%)					–0.01 (0.06)
	Attention to cue levels	Draws attention to changing cue levels	72 (53.30%)	0.78 (0.04)	[0.75, 0.81]	0.12	[0.63, 0.93]	
		Does not draw attention to changing cue levels	63 (46.70%)	0.77 (0.05)	[0.73, 0.80]	0.12	[0.62, 0.92]	
		Total	135 (100.00%)					–0.04 (0.06)
Study design characteristics	Survey medium[d]	Paper-and-pencil (in-person) text	90 (63.40%)	0.80 (0.04)	[0.77, 0.82]	0.12	[0.65, 0.95]	
		Web-based (online) text	43 (30.30%)	0.74 (0.05)	[0.69, 0.78]	0.10	[0.61, 0.87]	
		Audio	1 (0.70%)	NA	NA	NA	NA	
		Web-based (online) video	1 (0.70%)	NA	NA	NA	NA	
		Other[f]	7 (4.90%)	NA	NA	NA	NA	
		Total	142 (100.00%)					–0.20* (0.06)
	Study design[d]	Full-factorial (orthogonal) design	81 (49.70%)	0.78 (0.04)	[0.74, 0.81]	0.12	[0.63, 0.93]	
		Fractional design	45 (27.60%)	0.80 (0.05)	[0.76, 0.84]	0.12	[0.65, 0.95]	
		Block design	9 (5.50%)	NA	NA	NA	NA	
		Other[g]	28 (17.20%)	NA[h]	NA	NA	NA	
		Total	163 (100.00%)					0.08 (0.07)
	Time gap	Same session	147 (90.20%)	0.78 (0.04)	[0.76, 0.80]	0.12	[0.62, 0.93]	
		Separate sessions	16 (9.80%)	0.73 (0.10)	[0.62, 0.81]	0.09	[0.61, 0.85]	
		Total	163 (100.00%)					–0.08 (0.11)

(continued)

Table 4. (continued)

Antecedent Type	Categorical Putative Antecedent	Category	Frequency (Percentage)	Mean Test-Retest Reliability Estimate (*SE*)[a]	95% CI	SD_{rho}[b]	80% CV[b]	b (*SE*)[c]
	Test-retest reliability level of analysis[d]	Within-person	75 (57.30%)	0.79 (0.04)	[0.75, 0.81]	0.13	[0.62, 0.96]	
		Between-persons	51 (38.90%)	0.76 (0.05)	[0.72, 0.80]	0.11	[0.62, 0.90]	
		Both[i]	5 (3.80%)	NA	NA	NA	NA	
		Total	131 (100.00%)					–0.09 (0.06)
Substantive characteristics	Topic area	OBHR	82 (50.31%)	0.77 (0.04)	[0.74, 0.80]	0.12	[0.62, 0.92]	
		Other (i.e., non-OBHR) area	81 (49.69%)	0.78 (0.04)	[0.75, 0.81]	0.11	[0.64, 0.92]	
		Total	163 (100.00%)					0.07 (0.06)
	Judgment type	Attitude	36 (22.09%)	0.76 (0.06)	[0.71, 0.81]	0.09	[0.64, 0.88]	0.12 (0.09)[j]
		Perception	47 (28.83%)	0.79 (0.05)	[0.74, 0.82]	0.10	[0.66, 0.92]	0.20 (0.08)[j]
		Behavioral intention	52 (31.90%)	0.80 (0.05)	[0.76, 0.83]	0.13	[0.63, 0.97]	0.26* (0.08)[j]
		Mixed/undetermined	28 (17.18%)	0.72 (0.06)	[0.66, 0.78]	0.12	[0.57, 0.87]	
		Total	163 (100.00%)					–0.06 (0.03)

Note. We tested each putative antecedent to reliability (equivalent to a moderator in a validity generalization analysis) separately—that is, one at a time—to maximize the *k* (number of independent samples) in each analysis because the *k* varies dramatically across antecedents and because the listwise *k* in a model containing multiple antecedents was often quite low. As can be seen in the online supplementary materials, most (although not all) of these results were replicated when the data were analyzed separately at the within-person versus between-person levels of analysis. OBHR ¼organizational behavior and human resources.
[a]Mean test-retest reliability is reported for each category (subgroup) of each antecedent. We only reported mean test-retest reliability for categories with enough data. Categories with insufficient data are marked as NA. [b]SD_{rho} and 80% credibility interval are computed using Schmidt and Hunter's (2015) method. [c]Standardized regression coefficients from metaregression equations where the Fisher z_r-transformed test-retest correlation is regressed onto each predictor are reported. Because we examined the effect of each putative antecedent separately (except for judgment type), there is only one predictor in each metaregression equation. [d]Categories with fewer than 10 effect sizes were omitted from analyses; these categories are represented by NA in the table. Each putative antecedent except judgment type was ultimately treated as dichotomous in the analyses, whereas judgment type was treated as a set of three dummy variables. [e]Samples in this category included both students and nonstudents. [f]Either a computer was used to project the policy-capturing scenarios while participants provided their judgments on paper or else some participants used web-based surveys and other participants used paper-and-pencil surveys. [g]The "Other" study design category includes either nonorthogonal designs or orthogonal designs in which each participant receives a random set of scenarios. [h]The "Other" study design category includes different designs (e.g., random subsets of scenarios, designs that do not specify cue levels). Given the difficulty in interpreting the mean test-retest reliability for this category, we omit the mean test-retest reliability for this category. [i]Five studies reported both the within-person and between-person test-retest reliability estimates. In the analysis to compute the mean meta-analytic test-retest reliability estimate and the analyses for all antecedents except one (see next sentence), we took the average of the within-person and between-person test-retest reliability estimates for each of these primary studies. These five studies were, however, not included in the analysis for the antecedent of levels of analysis. [j]Standardized regression coefficients from a multiple metaregression equation where the Fisher z_r-transformed test-retest correlation is regressed onto the three dummy variables, each with the corresponding category coded as 1 and the rest coded as 0. This analysis revealed that behavioral intention judgments exhibited significantly higher test-retest reliability than did the set of other types of judgments (i.e., attitudinal, perceptual, and mixed/undetermined). *$p < .05$.

Table 5. Descriptive and Meta-Analytic Statistics for Continuous Putative Antecedents to Test-Retest Reliability.

Antecedent Type	Continuous Putative Antecedent	*k*	*M*	*SD*	*Min*	*Max*	b^a	*SE*
General study and sample characteristics	Corrected year of publication[b]	163	2003.90	11.17	1968	2021	-0.06	0.00
	Sample size[c]	163	124.20	147.48	4	864	-0.03	0.01
	Journal impact factor[c]	163	6.58	11.91	2	13.25	-0.07	0.07
Scenario characteristics	Number of cues[c]	163	6.58	11.91	2	150	-0.08	0.03
	Number of unique scenarios[c]	162	45.97	44.44	8	390	0.01	0.01
	Number of repeated scenarios[c]	160	11.08	15.84	1	100	-0.01	0.02
	Total number of scenarios[c]	159	56.61	54.71	9	480	0.00	0.01

Note. We tested each putative antecedent to reliability (equivalent to a moderator in a validity generalization analysis) separately—that is, one at a time—to maximize the *k* (number of independent samples) in each analysis because the *k* varies dramatically across antecedents and because the listwise *k* in a model containing multiple antecedents was often quite low. As can be seen in the online supplementary materials, most (although not all) of these results were replicated when the data were analyzed separately at the within-person versus between-person levels of analysis.

[a]Standardized regression coefficients from metaregression equations where the Fisher z_r-transformed test-retest correlation is regressed onto each predictor are reported. Because we examined the effect of each putative antecedent separately, there is only one predictor in each metaregression equation. [b]To correct for time spent on the future publication process, the corrected publication year for doctoral dissertations, master's theses, and other theses is equal to (Original Year 2), and the corrected publication year for conference presentations is equal to (Original Year 1); in contrast, the corrected publication year for journal articles is equal to the original publication year (Evans et al., 2018; cf. Wegman et al., 2018). It should be noted that the analysis described in text and in Table 3 contrasted policy-capturing studies conducted before versus after the 2002 policy-capturing tutorials (i.e., Aiman-Smith et al., 2002; Karren & Barringer, 2002), whereas the current analysis examined corrected publication year as a continuous antecedent variable. See also, in this regard, the analogous analysis in the online supplementary materials in which influential studies were removed. [c]Due to high skewness, this variable was square-root-transformed before being used in the meta-analysis.

Based on these analyses, among the methodological antecedents, only the survey medium was significantly associated with test-retest reliability: reliability estimates were higher for paper-and- pencil surveys (k 90, N 8,746, $M_{weighted}$ 0.80) than for online surveys (k = 43, N = 7,648, $M_{weighted}$ = 0.74; b = –0.20, SE = 0.06, p = .020). According to the Common Language Effect Size Indicator (McGraw & Wong, 1992), in 55 out of 100 comparisons, a study involving a paper-and- pencil policy-capturing survey will have a higher test-retest reliability than a study involving an online survey. Among the substantive antecedents, the test-retest reliability estimates were higher from studies using behavioral intention judgments (k = 52, N = 6,482, $M_{weighted}$ = 0.80) versus nonbehavioral intention judgments (k = 111, N = 13,762, $M_{weighted}$ = 0.76) while controlling for other judgment types (i.e., two additional dummy variables for the attitudinal and perceptual judgment types, with the mixed/ undeterminable" judgment type serving as the reference group); b for the behavioral intention judgments dummy variable = 0.26, *SE* = 0.08, *p* = .017. Using the Common Language Effect Size Indicator, in 53 out of 100 random comparisons, a policy- capturing study with behavioral intention judgments will have a higher level of test-retest reliability than a policy-capturing study with nonbehavioral intention judgments.

These results were largely replicated across an array of robustness checks described further in the online supplementary materials: (a) four sets of sensitivity analyses, (b) a multilevel meta-analysis that accounted for the nesting of studies within authors, and (c) an exploratory metaregression model containing the two predictors that were significant in the focal analyses—survey medium and the dummy variable for behavioral intention judgments—while controlling for the dummy variables for attitudinal and perceptual judgments.

Discussion

As we demonstrate in this article, the percentage of policy-capturing studies reporting test-retest reliability is very low: 17.17% (164 out of 955). As such, the most important conclusion from this review is a reiteration of previous advice published in *Organizational Research Methods* (Aiman-Smith et al., 2002; Karren & Barringer, 2002) regarding the need to measure and report test-retest reliability correlations in policy-capturing studies. Reliability estimates are often required for the publication of other social science measures, and policy-capturing measures should be no different. In the case of policy capturing, test-retest reliability is necessary, although not sufficient, for a valid judgment policy. It is worth noting, however, that the available evidence suggests that the afore- mentioned tutorials may have had a modest positive effect on *overall* reliability reportage while seemingly having little effect on *selective* reliability reportage (i.e., failure to report low reliability). Of primary importance, this meta-analysis was conducted to determine the average test-retest reliability estimate reported in policy-capturing studies and the extent to which this average estimate generalizes across various factors. Results support the conclusion that on the whole, policy capturing is a relatively reliable way to assess the factors decision-makers use to make judgments (mean *r*.78). Furthermore, the test-retest reliability of policy capturing generalizes across several methodological choices made by primary study authors. An exception was survey medium. Specifically, test-retest reliability estimates were higher for paper-and-pencil surveys than for web-based (online) surveys. It may be the case that in-person paper-and-pencil studies put external pressure on decision-makers via a Hawthorne effect, leading to more stable judgments and thus higher test-retest reliability. For substantive factors, although reliability generalizes across OBHR versus non-OBHR studies, it varies across studies with different types of judgments. Specifically, individuals make more stable behavioral intention than nonbehavioral intention (e.g., attitudinal, perceptual) judgments.

Limitations, Implications, and Suggestions for Future Research

This meta-analysis, like any other, has several limitations. Primary among the current limitations is that prior advice to the contrary (Aiman-Smith et al., 2002; Karren & Barringer, 2002) notwithstanding, the modal policy-capturing study did not report a test-retest reliability estimate—and therefore could not be included in the meta-analysis. Although many primary study authors may well have been inattentive to test-retest reliability because it was not the focus of their research, we cannot rule out the possibility that some authors may have found test-retest reliability to be low and then opted not to report it. Importantly, selective nonreportage may mask (i.e., attenuate) the impact of antecedents on test-retest reliability. Therefore, although meta-analyses frequently have a chilling effect on subsequent primary studies, we encourage continued research on antecedents to test-retest reliability in policy-capturing studies. Moreover, we encourage journal editors and reviewers to insist that authors report test-retest reliability estimates in policy-capturing studies.

Second, although the total number of independent samples that did report a test-retest reliability estimate was more than sufficient to estimate an average effect size estimate, it did somewhat constrain our ability to examine potential methodological and substantive antecedents to test-retest reliability. Specifically, although we were able to examine the impact of each antecedent separately, we were unable (due to often low listwise-deleted *k*) to examine the impact of multiple antecedents simultaneously or interactions between antecedents. Future primary studies could there- fore manipulate and examine interactions between conceptually meaningful combinations of antecedents (e.g., number of cues per scenario, scenario design, and time gap) to determine the importance of making tradeoffs to control the length of the policy-capturing component of the survey (e.g., compensating for a high number of cues per scenario by using a block vs. full orthogonal design).

Third, although we proposed that the number of scenarios and the number of cues per scenario may have had an effect on test-retest reliability as a function of survey length, we were unable to directly assess survey length. Specifically, survey length in a given primary study is attributable not only to the policy-capturing component but also to other components (e.g., Likert-type self-report measures). Future primary studies could therefore manipulate and examine interactions between the lengths of the policy-capturing component and other components of the survey to determine the importance of making tradeoffs to control the length of the overall survey (e.g., compensating for a long policy-capturing measure by shortening the non-policy-capturing components of the survey). Fourth, although we discussed vigilance decrement and ego depletion as possible reasons why certain methodological factors may exert effects on test-retest reliability, we were unable to actually measure vigilance decrement or ego depletion directly. Given recent concerns regarding the replicability of the ego-depletion phenomenon (e.g., Hagger et al., 2016), it seems important for future policy-capturing research to measure this phenomenon directly, ideally using preregistered studies. Preregistration can make a clear distinction between a priori and post hoc analyses, thus promoting transparency and reducing “opportunistic researcher degrees of freedom” (Toth et al., 2020).

Fifth, we were unable to examine some potentially important methodological antecedents because primary studies rarely reported this information. In particular, primary studies rarely reported either the extremity (in terms of cue values) or the order (i.e., location or serial position within the set of scenarios) of the first iteration of the repeated scenarios. For example, based on the “peak-end rule” (Fredrickson & Kahneman, 1993), scenarios with all positive (negative) cue values and scenarios located at the end of the original set of scenarios may be most salient to decision- makers such that test-retest reliability may be higher if these scenarios are repeated. Future primary studies could test such assertions.

Finally, because the survey medium was the sole statistically significant methodological antecedent to test-retest reliability, this methodological factor deserves further attention. Future primary studies could identify additional (beyond the Hawthorne effect) explanations for this effect and could then manipulate each potential explanation independently from survey medium. Moreover, due to the observed frequencies of primary studies in categories within the survey medium factor, this factor was ultimately examined as a comparison between paper-and-pencil studies and web- based (online) studies. More primary studies are needed that use audio, video, and other media.

Similarly, because the behavioral intention versus nonbehavioral intention judgment type was the sole statistically significant substantive antecedent to test-retest reliability, this substantive factor deserves further attention. For many policy-capturing studies, researchers may have some flexibility regarding which type of judgments to use. For example, in a policy-capturing study on job choice, researchers can either use a behavioral intention question, by asking about one’s likelihood of accepting a job offer, or an attitudinal question, by asking about the favorableness of a job offer. When theory does not dictate the judgment type, researchers may wish to maximize reliability by using behavioral intention judgments.

Our findings have implications for stimulus-material adaptations in policy-capturing studies. The current meta-analysis demonstrates that test-retest reliability in policy-capturing designs generalizes across observed variation in most of the methodological factors examined. This in turn suggests that vis-a`-vis reliability, researchers do have some leeway in adapting stimulus materials. Less certainty, however, exists with regard to adapting the number of cues and scenarios in particular: Although the focal meta-analytic results as well as the ancillary within-person results in the online supplementary materials suggest that test-retest reliability generalizes across observed variation in the number of cues and several operationalizations of the number of scenarios, the ancillary between-person results in the online supplementary materials suggest that cutting the number of cues or unique scenarios reduces reliability. Cuts to the number of cues or unique scenarios (e.g., in an effort to shorten an existing policy-capturing measure) should therefore be made judiciously.

Best-Practice Recommendations for Future Policy-Capturing Studies

It seems only fitting to end this article by providing guidance for future policy-capturing studies, drawn from what was—and what was unable to be—examined in this meta-analysis. In particular, we provide recommendations associated with (a) reporting reliability, (b) designing policy- capturing studies for the reportage of reliability, and (c) interpreting reliability. Table 6 provides a summary of our recommendations.

Reporting Reliability. First, we repeat previous recommendations (e.g., Aiman-Smith et al., 2002; Karren & Barringer, 2002) that future policy-capturing studies should routinely report test-retest reliability estimates. We moreover encourage researchers to report both within-person and between- person reliability estimates because these estimates provide unique information. Moreover, we suggest that researchers report the standard deviation of within-person reliability estimates (across people, perhaps as a function of individual differences such as conscientiousness that may result in higher vs. lower reliability) and the standard deviation of between-person reliability estimates (across scenarios that might be expected to exhibit higher or lower reliability due, e.g., to the peak-end rule).

We moreover recommend that researchers take action to improve the reliability of their policy- capturing measures. As noted previously, for instance, in cases where the type of judgment is not dictated by theory, researchers have the potential to improve test-retest reliability by using behavioral intention judgments rather than, say, attitudinal judgments. Having said this, we acknowledge that researchers have limited options to increase reliability due to the fact that reliability generalized across virtually all the antecedents we examined.

Designing Policy-Capturing Studies for the Reportage of Reliability. To compute test-retest reliability, researchers need to repeat some scenarios and correlate the scores across the first and second iterations of these scenarios. Therefore, researchers need to consider the number of scenarios to repeat and which scenarios to repeat. We therefore provide some best-practice recommendations on these issues.

Number of scenarios to repeat. The ideal number of scenarios to repeat presents a tradeoff. On the one hand, controlling the length of a policy-capturing measure to reduce any vigilance decrement effect is generally a concern. On the other hand, a test-retest reliability correlation coefficient, like any correlation coefficient, is calculated by correlating two vectors of data points. If each vector contains only a few values, the correlation coefficient is likely to be very unstable (Schönbrodt & Perugini, 2013). Therefore, especially when reporting within-person test-retest correlations in policy-capturing studies, we recommend repeating at least 10 scenarios (i.e., the median number of repeated scenarios in the within-person studies in our meta-analysis[8]).

Which scenarios to repeat. We recommend that researchers consider the potential extremity effects and order effects mentioned previously (e.g., the peak-end rule). Future research can not only facilitate the empirical examination of such effects but also, at a practical level, can ensure that the selection of scenarios is systematic and therefore more comparable across studies. Thus, we recommend that researchers repeat salient scenarios under extremity and order effects along with randomly selected scenarios.

Interpreting Reliability. We discourage policy-capturing researchers from using hard reliability cutoffs (and then selectively underreporting reliability estimates that fall below these cutoffs). Our recommendation is consistent with guidelines in other quantitative research areas (e.g., Greco et al., 2018; Williams et al., 2020). Moreover, researchers should carefully contextualize reliability estimates (see Table 6 for details).

Table 6. Recommendations for Future Policy-Capturing Studies.

Category	Domain	Recommendation(s)	Reason(s)
Reportage	1. Reporting test-retest reliability	• Researchers should report test-retest reliability! • Journal reviewers should require that researchers report test-retest reliability! • When the outcome variables are continuous rather than nominal, we suggest that researchers report the test-retest Pearson product-moment correlation coefficient.	• In policy capturing, as elsewhere, reliability is necessary but not sufficient for validity. • The primary reliability-based concern in policy-capturing measures is the extent to which the decision-maker is using a temporally stable judgment policy across scenarios in the measure. • Almost all the primary studies in our meta-analysis involved judgments (vs. choices). Therefore, the outcome variables (e.g., scores on decision-makers' responses to the two iterations of the repeated scenarios) are generally continuous rather than nominal variables. In this case, Pearson's product-moment correlation coefficient, a pure estimate of *reliability* (LeBreton & Senter, 2008), is the appropriate test-retest reliability statistic.
	2. Reporting additional types of reliability (or agreement)	• If researchers are interested in absolute agreement in addition to reliability, they can *additionally* report a test-retest equivalent of r_{wg} (LeBreton & Senter, 2008; see also Berchtold, 2016). • In the more infrequent cases where multiple judgments are elicited per policy-capturing scenario, researchers can *additionally* report Cronbach's alpha.	• Test-retest reliability refers to the relative stability of scores across iterations, but test-retest agreement, which refers to the absolute agreement of scores across iterations, may also frequently be of interest. • Internal consistency (assessed by Cronbach's alpha) is likely to be of interest only in the more infrequent cases when multiple

(continued)

Table 6. (continued)

Category	Domain	Recommendation(s)	Reason(s)
			judgments (rather than one) are elicited per policy-capturing scenario—and when these judgments are moreover assumed to be indicators of the same underlying construct.
	3. Reporting test-retest reliability at multiple levels of analysis	. Researchers should report both between-person and within-person reliability estimates.	. Between-person and within-person reliability estimates provide unique—and complementary—information. Between-person test-retest reliability captures the stability of judgments across participants (nomothetic). Within-person test-retest reliability captures the stability of judgments across policy-capturing scenarios (idiographic). . Both similarities and differences in obtained estimates across the between-person and within-person levels are noteworthy (Dalal et al., 2014). We found similar reliability estimates across the two levels, but future research that follows the subsequent recommendations in this table (e.g., recommendations involving the number and nature of scenarios that should be repeated) should continue to report—and compare—reliability estimates at both levels. The
	4. Reporting sufficient information about test-retest reliability	. Researchers should report not only the mean reliability estimate but also the standard deviation of estimates	. standard deviation of within-person reliability estimates across people provides information regarding

(continued)

Table 6. (continued)

Category	Domain	Recommendation(s)	Reason(s)
		across scenarios (for between-person reliability estimates) and across people (for within-person reliability estimates).	the extent to which judgments differ in test-retest reliability across people. A large standard deviation across people would suggest the need to examine the impact of traits (e.g., general mental ability, conscientiousness, self-monitoring) and/or states (e.g., positive and negative mood). . The standard deviation of between-person reliability estimates across scenarios shows the differences in reliability across scenarios. A large standard deviation across scenarios would suggest the need to examine the impact of scenario or cue characteristics (e.g., scenario serial position, cue extremity).
Designing policy-capturing studies for the reportage of reliability	1. Including enough repeated scenarios	. Researchers should include enough scenarios (i.e., at least 10 scenarios) to estimate within-person test-retest reliability.	. When estimating within-person reliability, we recommend repeating at least 10 scenarios (i.e., the median number of repeated scenarios in the within-person reliability studies). . A test-retest reliability correlation coefficient, like any correlation coefficient, is calculated by correlating two vectors of data points. If each vector contains only a few values, the correlation coefficient is likely to be very unstable (Schönbrodt & Perugini, 2013).[a]

(continued)

Table 6. (continued)

Category	Domain	Recommendation(s)	Reason(s)
	2. Choosing the repeated scenarios	. We recommend that researchers repeat the scenarios that would be most salient under extremity and order effects (i.e., the last scenario, the first scenario, the scenario with the highest possible value on all cues, and the scenario with the lowest possible value on all cues) along with at least 6 other randomly selected scenarios.	. We speculate that extremity and/or order effects may influence test-retest reliability estimates.[b] . Higher reliability estimates may be found for scenarios that contain extreme (i.e., lowest or highest) levels of all cues compared to randomly selected scenarios. . Higher reliability estimates may be found for scenarios whose first iteration occurs at the end or beginning of the set of scenarios compared to randomly selected scenarios.
Interpreting reliability	1. Avoiding hard reliability cutoffs and selective reporting of reliability	. Researchers should neither use hard reliability cutoffs in policy-capturing studies nor selectively underreport reliability estimates that fall below hard cutoffs. . Instead, researchers should interpret obtained reliability estimates in light of the 95% confidence interval from the current meta-analysis (i.e., .75 to .80) as well as in light of the specific domain being studied, the stage of scale validation, and the purpose of the study (e.g., basic research vs. high-stakes decisions in applied settings).	. We discourage policy-capturing researchers from using hard reliability cutoffs because: (a) researchers may then underreport reliability estimates that fall below these cutoffs, and (b) reliability estimates should be contextualized.

[a]Our recommendation to repeat at least 10 scenarios exceeds that of Aiman-Smith et al. (2002), who recommended repeating "4 to 5" scenarios (p. 409) even for within-person test-retest correlations. Our recommendation is different because the stability of within-person test-retest correlations is likely to be appreciably higher when repeating 10 rather than four to five scenarios (Scho¨nbrodt & Perugini, 2013). The downside of repeating 10 or more scenarios is that doing so increases the length of the study, potentially leading to lower quality responses. However, the burden on decision-makers is unlikely to be appreciably higher when repeating 10 rather than four or five scenarios. Moreover, if needed, researchers can divide the study into two sessions. [b]Although we speculate that extremity and/or order effects may influence test-retest reliability estimates, we could not test such effects meta-analytically because very few primary studies reported the cue levels in each repeated scenario or the location of the first iteration of each repeated scenario within the set of scenarios. Our recommendation, if followed, would, by definition, increase such reportage.

Acknowledgements

The authors thank Jeremy Wong for his help with the literature search; Jesse Olsen, Hannes Zacher, Joel Marcus, and Nadine Raaphorst for conducting additional analyses on their data sets for us; and Mark Reynolds, Esther Kaufmann, and Una Adderley for providing us with unpublished papers or data sets.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or pub- lication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Ze Zhu https://orcid.org/0000-0003-1076-3215
Reeshad S. Dalal https://orcid.org/0000-0003-4040-4686 Shea Fyffe https://orcid.org/0000-0003-0312-7915

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Due to space constraints, this Short Report does not aim to provide readers with a comprehensive under- standing of the policy-capturing technique. For more detailed information, we recommend Cooksey (1996), Aiman-Smith et al. (2002), and Karren and Barringer (2002).
2. In practice, nonstudents in policy-capturing studies are often (but not always) domain experts. Research has shown that novices and experts approach judgment tasks differently (Aiman-Smith et al., 2002; Hardiman et al., 1989; Mackay et al., 1992), and some researchers have suggested that the results from judgment tasks that use students are unlikely to generalize to judgment tasks requiring domain expertise (Barr & Hitt, 1986; Shanteau & Stewart, 1992).
3. Mathematically, assuming the sample size is large enough to compute a

test-retest reliability coefficient, sample size should affect the standard error and therefore the width of the confidence interval associated with the reliability estimate rather than affecting the size of the point estimate itself.

4. It is also possible for policy-capturing studies to use auditory scenarios or video-based scenarios, and thus we also coded for these cue presentation formats.
5. Whereas block designs use all possible blocks (i.e., each group of participants receives a different block of scenarios), fractional factorial designs include only one subset (i.e., all participants receive the same block of scenarios).
6. It is worth noting that the desirability (or feasibility) of computing within-person test-retest reliability depends on the number of repeated scenarios. Mathematically, a correlation coefficient can be estimated and moreover can exhibit values other than –1 and þ1 with at least three data points—here, three repeatedpolicy-capturing scenarios. However, even beyond this minimum number of data points, researchers should be concerned about the potential for sampling error and departures from normality.
7. In addition to examining level of analysis as a potential antecedent to test-retest reliability, we conducted analyses separately at each level of analysis to examine whether the impact of the other potential antecedents was similar at both levels. See the online supplementary materials for these additional analyses.
8. In contrast, the median number of repeated scenarios in the between-person studies in our meta-analysis was two.

References

References marked with an asterisk indicate primary studies included in the meta-analysis.

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*, 351-371.

*Aiman-Smith, L., Bauer, T. N., & Cable, D. M. (2001). Are you attracted? Do you intend to pursue? A recruiting policy-capturing study. *Journal of Business and Psychology*, *16*, 219-237.

Aiman-Smith, L., Scullen, S., & Barr, S. (2002). Conducting studies of decision making in organizational contexts: A tutorial for policy-capturing and other regression-based technologies. *Organizational Research Methods*, *5*, 388-414.

*Albers-Miller, N. D. (1999). Consumer misbehavior: Why people buy illicit goods. *Journal of Consumer Marketing*, *16*, 273-287.

*Andersen, J., Larsen, J. K., Schultz, V., Nielsen, B. M., Kørner, A., Behnke, K., Munk-Andersen, E., Butler, B., Allerup, P., .. . Bech, P. (1989). The brief psychiatric rating scale. *Psychopathology*, *22*, 168-176.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*, 3-25.

*Athanasou, J., & Cooksey, R. (2001). Judgment of factors influencing interest: An Australian study. *Journal of Vocational Education Research*, *26*, 77-96.

Ashton, R. H. (2000). A review and analysis of research on the test–retest reliability of professional judgment. *Journal of Behavioral Decision Making*, *13*, 277-294.

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment*, *68*, 139-151.

*Baker, S., & Thompson, C. (2012). Initiating artificial nutrition support: A clinical judgement analysis. *Journal of Human Nutrition and Dietetics*, *25*, 427-434.

*Barley, E. A., & Jones, P. W. (2005). Analysis of the cues used by patients when making assessments of their asthma severity. *European Respiratory Journal*, *25*, 671-676.

Barr, S., & Hitt, M. (1986). A comparison of selection decision models in manager versus student samples. *Personnel Psychology*, *39*, 599-617.

Baumeister, R. F., & Heatherton, T. F. (1996). Self-regulation failure: An overview. *Psychological Inquiry*, *7*, 1-15.

*Bech, P., Haaber, A., & Joyce, C. R. B. (1986). Experiments on clinical observation and judgement in the assessment of depression: Profiled videotapes and judgement analysis. *Psychological Medicine*, *16*, 873-883.

*Beckstead, J. W., Pezzo, M. V., Beckie, T. M., Shahraki, F., Kentner, A. C., & Grace, S. L. (2014). Physicians' tacit and stated policies for determining patient benefit and referral to cardiac rehabilitation. *Medical Decision Making*, *34*, 63-74.

*Beckstead, J. W., & Stamp, K. D. (2007). Understanding how nurse practitioners estimate patients' risk for coronary heart disease: A judgment analysis. *Journal of Advanced Nursing*, *60*, 436-446.

Berchtold, A. (2016). Test-retest: Agreement or reliability? m*ethodological Innovations*, *9*, 1-7. Bevan, W., & Steger, J. A. (1971). Free recall and abstractness of stimuli. *Science*, *172*, 597-599.

*Blume, B. D., Baldwin, T. T., & Rubin, R. S. (2009). Reactions to different types of forced distribution performance evaluation systems. *Journal of Business and Psychology*, *24*, 77-91.

*Bolton, K. R. (2005). *Organizational characteristics, goal orientation, and organizational attractiveness: A person-organization fit perspective* [Publication No. 1427916] [Master's thesis, University of Houston]. ProQuest Dissertations & Theses Global.

*Bonaccio, S. (2006). *Determining the relative importance of antecedents to advice utilization during decision- making with and without missing information* [Publication No. 3251589] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Bonaccio, S., & Dalal, R. S. (2010). Evaluating advisors: A policy-capturing study under conditions of complete and missing information. *Journal of Behavioral Decision Making*, *23*, 227-249.

*Bradshaw, G. L., & Shaw, D. (1992). Forecasting solar flares: Experts and artificial systems. *Organizational Behavior and Human Decision Processes*, *53*, 135-157.

*Braspenning, J., & Sergeant, J. (1994). General practitioners' decision making for mental health problems: Outcomes and ecological validity. *Journal of Clinical Epidemiology*, *47*, 1365-1372.

Brown, T. R. (1972). A comparison of judgmental policy equations obtained from human judges under natural and contrived conditions. *Mathematical Biosciences*, *15*, 205-230.

*Brown, T. J., & Allgeier, E. R. (1996). The impact of participant characteristics, perceived motives, and job behaviors on co-workers' evaluations of workplace romances. *Journal of Applied Social*

Psychology, *26*, 577-595.
*Browne, J. P., O'Boyle, C. A., McGee, H. M., McDonald, N. J., & Joyce, C. B. (1997). Development of a direct weighting procedure for quality of life domains. *Quality of Life Research*, *6*, 301-309.
*Brundin, E., & Gustafsson, V. (2013). Entrepreneurs' decision making under different levels of uncertainty: The role of emotions. *International Journal of Entrepreneurial Behavior & Research*, *19*, 568-591.
*Cable, D. M., & Graham, M. E. (2000). The determinants of job seekers' reputation perceptions. *Journal of Organizational Behavior*, *21*, 929-947.
*Cable, D. M., & Judge, T. A. (1994). Pay preferences and job search decisions: A person-organization fit perspective. *Personnel Psychology*, *47*, 317-348.
*Carkenord, D. M., & Stephens, M. G. (1994). Understanding student judgments of teaching effectiveness: A "policy capturing" approach. *The Journal of Psychology*, *128*, 675-682.
*Carson, K. P., Cardy, R. L., & Dobbins, G. H. (1991). Performance appraisal as effective management or deadly management disease two initial empirical investigations. *Group & Organization Management*, *16*, 143-159.
Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of metaanalytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*, 796-815.
*Carvalho, M. A. (2005). *Accessing legislative voting behavior: A multimethod research approach* [Publication No. 3196120] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment*, *20*, 333-346.
*Chen, W. Y., Wang, M. L., & Hsu, B. F. (2013). When PJ fit and PO fit meet guanxi in a Chinese selection context. *Journal of Technology Management in China*, *8*, 174-189.
*Cheyne, H., Dalgleish, L., Tucker, J., Kane, J., Shetty, Ashalatha., McLeod, S., & Niven, C. (2012). Risk assessment and decision making about in-labour transfer from rural maternity care: A social judgment and signal detection analysis. *BMC Medical Informatics and Decision Making*, *12*, Article 122. https://doi.org/ 10.1186/1472-6947-12-122
*Chhinzer, N. N. (2007). *Evaluating layoff techniques: A policy-capturing*

study of voluntary versus involuntary layoffs [Publication No. 978-0-494-28125-3] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Christie, L. A. (1996). *Expertise in nurses' clinical judgments: The role of cognitive variables and experience* [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. Academic Press.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98-104.

*Crocker, R. K., & Banfield, H. (1986). Factors influencing teacher decisions on school, classroom, and curriculum. *Journal of Research in Science Teaching*, *23*, 805-816.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Erlbaum.

*Cuguero´-Escofet, N., Fortin, M., & Canela, M. A. (2014). Righting the wrong for third parties: How monetary compensation, procedure changes and apologies can restore justice for observers of injustice. *Journal of Business Ethics*, *122*, 253-268.

*Dahling, J. J., & Thompson, M. N. (2010). Contextual supports and barriers to academic choices: A policy- capturing analysis. *Journal of Vocational Behavior*, *77*, 374-382.

Dalal, R.S., Bhave, D.P., & Fiset, J. (2014). Within-person variability in job performance: A theoretical review and research agenda. *Journal of Management*, *40*, 1396-1436.

*Dalal, R. S., & Bonaccio, S. (2010). What types of advice do decision-makers prefer? *Organizational Behavior and Human Decision Processes*, *112*, 11-23.

*Davies, M. L. (2011). *Detecting and preventing financial abuse of older adults: Examining decision making by health, social care and banking professionals*. [Publication No. U578889] [Doctoral dissertation]. Dissertations & Thesis Global.

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. Academic Press.

*de Cueto, J. F. (2012). *Relative importance of false positives in the selection process* [Publication No. 3517008] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*De Goede, M. E., Van Vianen, A. E., & Klehe, U. C. (2013). A tailored policy-capturing study on po fit perceptions: The ascendancy of attractive over aversive fit. *International Journal of Selection and Assessment*, *21*, 85-98.

*Deshpande, S. P., & Joseph, J. (1994). Variation in compensation decisions by managers: An empirical investigation. *The Journal of Psychology*, *128*, 41-50.

*Deshpande, S. P., & Schoderbek, P. P. (1992). Pay allocations by hospital administrators: An empirical analysis. *Journal of Healthcare Management*, *37*, 321-332.

*Deshpande, S. P., & Schoderbek, P. P. (1993). Pay-allocations by managers: A policy-capturing approach. *Human Relations*, *46*, 465-479.

*DeTienne, D. R., Shepherd, D. A., & De Castro, J. O. (2008). The fallacy of "only the strong survive": The effects of extrinsic motivation on the persistence decisions for under-performing firms. *Journal of Business Venturing*, *23*, 528-546.

*Di Stasio, V. (2014). *Why education matters to employers: A vignette study in Italy, England and the Netherlands*. [Unpublished doctoral dissertation]. Universiteit van Amsterdam.

*Dineen, B. R., Noe, R. A., & Wang, C. (2004). Perceived fairness of web-based applicant screening procedures: Weighing the rules of justice and the role of individual differences. *Human Resource Management*, *43*, 127-145.

*Dorsey, D. W. (1997). *Managerial merit pay allocation: An analysis of alternative judgment models using regression and fuzzy expert system techniques* [Publication No. 9724003] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Drenth, D. J. (2002). *Differences in work preferences as a function of gender and ethnicity: A policy capturing approach* [Publication No. 3052641] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Drover, A. W. (2014). *The influence of angel investor characteristics on venture capitalist decision making* [Publication No. 3640752] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.

*Dulebohn, J., & Martocchio, J. J. (1998). Employees' perceptions of the distributive justice of pay raise decisions: A policy capturing approach. *Journal of Business and Psychology*, *13*, 41-64.

Ebert, R. J., & Kruse, T. E. (1978). Bootstrapping the security analyst.

Journal of Applied Psychology, *63*, 110-119.
Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*, 161-177.
Evans, S. C., Amaro, C. M., Herbert, R., Blossom, J. B., & Roberts, M. C. (2018) “Are you gonna publish that?” Peer-reviewed publication outcomes of doctoral dissertations in psychology. *PLOS ONE*, *13*(2), Article e0192219. https://doi.org/10.1371/journal.pone.0192219
*Fehr, R. (2007). *When apologies work: The benefits of matching apology content to victims and context* [Publication No. 1450214] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
Fishbein, M., & Ajzen, I. (2009). *Predicting and changing behavior: The reasoned action approach*. Taylor & Francis.
Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*, 45-55.
Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*, 316-355.
Garb, H. N. (1999). Call for a moratorium on the use of the Rorschach Inkblot Test in clinical and forensic settings. *Assessment*, *6*, 313-317.
*Gebing, T. A. (2000). *What is she thinking about? A policy capturing investigation of battered women’s decision to stay in violent relationships* [Publication No. 9995031] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Geller, S. E. (1994). *The influence of psychosocial factors on heart transplantation* [Publication No. 9424974] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*German, H. (2011). *Capturing the justice judgment: An application of the theory of representative design in two policy capturing studies in organizational justice* [Unpublished doctoral dissertation]. Durham University.
*German, H., Fortin, M., & Read, D. (2016). Justice judgments: Individual self-insight and between-and within-person consistency. *Academy of Management Discoveries*, *2*, 33-50.
*Golay, L. M. (2016). *A judgment analysis of psychological contracts: Priorities of part-time and full-time employees in relation to fulfillment and obligation to stay* [Unpublished doctoral dissertation]. University of Connecticut.

*Gorman, C. D., Clover, W. H., & Doherty, M. E. (1978). Can we learn anything about interviewing real people from “interviews” of paper people? Two studies of the external validity of a paradigm. *Organizational Behavior and Human Performance*, *22*, 165-192.
Graham, M. E., & Cable, D. M. (2001). Consideration of the incomplete block design for policy-capturing research. *Organizational Research Methods*, *4*, 26-45.
Greco, L. M., O’Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-analysis of coefficient alpha: A reliability generalization study. *Journal of Management Studies*, *55*, 583-618.
Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., & Zwienenberg, M. (2016). A multi-lab pre-registered replication of the ego depletion effect. *Perspectives on Psychological Science*, *11*, 546-573.
Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, *136*, 495-525.
*Hall, J. A. (1979). *On the adequacy of current authoritative guidelines for the review and evaluation of systems development controls* [Publication No. 8013017] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Hanisch, D. N., & Rau, S. B. (2014). Application of metric conjoint analysis in family business research. *Journal of Family Business Strategy*, *5*, 72-84.
Hardiman, P., Dufresne, R., & Mestre, J. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition*, *17*, 627-638.
*Harries, C. (1995). *Judgement analysis of patient management: General practitioners’ policies and self- insight* [Publication No. U072152] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Harries, C., Evans, J. St. B. T., Dennis, I., & Dean, J. (1996). A clinical judgement analysis of prescribing decisions in general practice. *Le Travail Humain*, *59*, 87-109.
*Harries, P. A. (2004). *Occupational therapists’ judgement of referral priorities: Expertise and training*. [Publication No. 10052525] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.
*Harries, P. A., & Gilhooly, K. (2003). Identifying occupational therapists’ referral priorities in community health. *Occupational Therapy International*, *10*, 150-164.

*Harries, P., & Gilhooly, K. (2011). Training novices to make expert, occupationally focused, community mental health referral decisions. *British Journal of Occupational Therapy*, *74*, 58-65.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, *21*, 111-121.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

*Hergert, J. (2016). *Personality, situation, and infidelity in romantic relationships*. [Unpublished doctoral dissertation]. University of Hagen.

*Heron, R. L. (1997). *Intuition and deliberation: A study of women's sexual choices* [Publication No. 9840305] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology*, *100*, 1143-1168.

*Hoffman, P. J., Slovic, P., & Rorer, L. G. (1968). An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin*, *69*, 338-349.

*Holcomb, R. C. (2011). *Improving judgment performance and cognitive systems engineering by examining the relationship between task properties and cognitive mode* [Publication No. 3455046] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Hollenbeck, J. R. (1984). *Control theory, self-focus and behavior in organizations (self-consciousness, self- awareness, retail sales)* [Publication No. 8241599] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.

*Holtz, B. C., Hu, B., & Han, S. (2017). Resource foci, valence, and distributive justice effects: A meta-analysis and policy capturing study. In *Academy of Management Proceedings* (Vol. *2017*, No. 1, p. 11586). Academy of Management.

*Hoye, G. V., Weijters, B., Lievens, F., & Stockman, S. (2016). Social influences in recruitment: When is word-of-mouth most effective? *International Journal of Selection and Assessment*, *24*, 42-53.

IBM Corp. 2010. *IBM SPSS Statistics for Macintosh, Version 19.0*.

*Ikomi, P. A., & Guion, R. M. (2000). The prediction of judgment in realistic tasks: An investigation of self- insight. *The International Journal of*

Aviation Psychology, *10*, 135-153.

*Irish, P. A., III. (1987). *Capturing the policies of time-constrained decision makers: The effects of deadline control, cue structure and individual difference variables* [Accession Order No. 8728022] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Jacklin, R., Svedalis, N., Harries, C., Darzo, A., & Vincent, C. (2008). Judgment analysis: A method for quantitative evaluation of trainee surgeons' judgments of surgical risk. *The American Journal of Surgery*, *195*, 183-188.

*Kaplan, S., Engelsted, L., Lei, X., & Lockwood, K. (2018). Unpackaging manager mistrust in allowing telework: Comparing and integrating theoretical perspective. *Journal of Business and Psychology*, *33*, 365-382.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*, 404-426.

Karren, R. J., & Barringer, M. W. (2002). A review and analysis of the policy-capturing methodology in organizational research: Guidelines for research and practice. *Organizational Research Methods*, *5*, 337-361.

*Karver, M. S. (2000). *The dangerousness to self judgment: Help line counselors' accuracy, agreement, and insight into judgment* [Publication No. 9996254] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

Kaufmann, E., Reips, U. D., & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PLOS ONE*, *8*(12), Article e83528. https://doi.org/10.1371/journal.pone.0083528

*Kennedy, D. J. (1999). *On the road again: An investigation of the situational and intentional antecedents of job relocation decisions in the service sector* [Publication No. 9914661] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Kight, W. D. (2010). *An analysis of reasonableness models for research assessments* [Publication No.3396317] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

Knight, A. P., & Eisenkraft, N. (2015). Positive is usually good, negative is not always bad: The effects of group affect on social integration and task performance. *Journal of Applied Psychology*, *100*, 1214-1227.

Koslowsky, M., & Sagie, A. (1993). On the efficacy of credibility intervals as indicators of moderator effects in meta-analytic research. *Journal of*

Organizational Behavior, *14*, 695-699.
*Kroecker, T. S. (1993). *Situational and individual influences upon ethical decision-making in organizations: A policy capturing approach* [Publication No. 9324225] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Kung, Y. H. (1997). *The enhancement of self-insight through frequency count on cue usage* [Publication No. 9815959] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Kutcher, E. J., III. (2007). *Assessing fit in the interview: How candidates consider content and context cues to person organization fit* [Publication No. 3310188] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815-852.
*Lee, H., & Dalal, R. S. (2011). The effects of extremities on ratings of dynamic performance appraisal. *Human Performance*, *24*, 99-118.
Leigh, D. E. (1986). Union preferences, job satisfaction, and the union-voice hypothesis. *Industrial Relations*, *25*, 65-71.
*Lelchook, A. (2010). *The use of humor by leaders in response to situational stressors* [Publication No.1474720] [Master's thesis]. ProQuest Dissertations & Theses Global.
*Levi, K. (1989). Expert systems should be more accurate than human experts: Evaluation procedures from human judgment and decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*, 647-657.
Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*, 9596-9601.
*Lievens, F., Highhouse, S., & Corte, W. (2005). The importance of traits and abilities in supervisors' hirability decisions as a function of method of assessment. *Journal of Occupational and Organizational Psychology*, *78*, 453-470.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. *49*). SAGE.
*Liu, S. H. (1993). *Priority of criteria used for occupational ratings: A study of Gottfredson's theory of circumscription and compromise* [Publication No. 9334387] [Doctoral dissertation]. ProQuest Dissertations & Theses

Global.
*Lopina, E. C. (2015). *Understanding older workers' decisions to participate in voluntary training opportunities* [Publication No. 3721050] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.
*MacDonald, H. A. (2008). *Motivational differences in feedback-seeking intentions: A cultural analysis* [Publication No. 978-0-494-43310-2] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*MacDonald, H. A., Sulsky, L. M., Spence, J. R., & Brown, D. J. (2013). Cultural differences in the motivation to seek performance feedback: A comparative policy-capturing study. *Human Performance*, *26*, 211-235.
Mackay, J., Barr, S., & Kletke, M. (1992). An empirical investigation of the effects of decision aids on problem-solving processes. *Decision Sciences*, *23*, 648-672.
Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*, 6-21.
*Marcus, J., MacDonald, H. A., & Sulsky, L. M. (2015). Do personal values influence the propensity for sustainability actions? A policy-capturing study. *Journal of Business Ethics*, *127*, 459-478.
*Marron, G. F. (1994). *Work and family issues: The impact of eldercare on work force policies and job choice decisions* [Publication No. 9501263] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
Marta-Pedroso, C., Freitas, H., & Domingos, T. (2007). Testing for the survey mode effect on contingent valuation data quality: A case study of web based versus in-person interviews. *Ecological Economics*, *62*, 388-398.
*McDonough, T. A. (2010). A policy capturing investigation of battered women's decisions to stay in violent relationships. *Violence and Victims*, *25*, 165-184.
*McGee, H. M., O'Boyle, C. A., Hickey, A., O'Malley, K., & Joyce, C. R. B. (1991). Assessing the quality of life of the individual: The SEIQoL with a healthy and a gastroenterology unit population. *Psychological Medicine*, *21*, 749-759.
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361-365.
*McIlroy, J. M. H. (2005). *The effect of decision condition in a judgmental policy capturing exercise* [Publication No. 0-494-07626-7] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in

survey data. *Psychological Methods*, *17*, 437-455.
*Meixner, W. F. (1985). *The effect of the organizational environment on judgment consensus: The case for professional government auditors in the state of Texas* [Publication No. 8514267] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Mellewigt, T., Thomas, A., Weller, I., & Zajac, E. J. (2017). Alliance or acquisition? A mechanisms-based, policy-capturing analysis. *Strategic Management Journal*, *38*, 2353-2369.
*Miller, T. L., & Wesley, C. L. (2010). Assessing mission and resources for social change: An organizational identity perspective on social venture capitalists 'decision criteria. *Entrepreneurship Theory and Practice*, *34*, 705-733.
*Mishra, V., & Roch, S. G. (2017). Do all raters value task, citizenship, and counterproductive behaviors equally: An investigation of cultural values and performance evaluations. *Human Performance*, *30*, 193-211.
Moher, D., Liberati, A., Tetzalaff, J., & Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097
Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*, 112-117.
*Monahan, C. J., & Muchinsky, P. M. (1985). Intrasubject predictions of vocational preference: Convergent validation via the decision theoretic paradigm. *Journal of Vocational Behavior*, *27*, 1-18.
*Murnieks, C. Y., Cardon, M. S., Sudek, R., White, T. D., & Brooks, W. T. (2016). Drawn to the fire: The role of passion, tenacity and inspirational leadership in angel investing. *Journal of Business Venturing*, *31*, 468-484.
*Murphy, K. R. (1979). *Convergent and discriminant validity of subjectively weighted models and regression models of decision making processes* [Publication No. 8006031] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*O'Boyle, C. A. (1994). The schedule for the evaluation of individual quality of life (SEIQoL). *International Journal of Mental Health*, *23*, 3-23.
*Oberoi, D. V., Jiwa, M., Mcmanus, A., & Parsons, R. (2016). Do men know which lower bowel symptoms warrant medical attention? A web-

based video vignette survey of men in western Australia. *American Journal of Men's Health*, *10*, 474-486.

*Ohme, M., & Zacher, H. (2015). Job performance ratings: The relative importance of mental ability, conscientiousness, and career adaptability. *Journal of Vocational Behavior*, *87*, 161-170.

*Okopny, D. R. (1985). *Planning an international audit: An empirical investigation of internal auditor judgment* [Publication No. 8526357] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Olsen, J. E., & Martins, L. L. (2016). Racioethnicity, community makeup, and potential employees' reactions to organizational diversity management approaches. *Journal of Applied Psychology*, *101*, 657-672.

*Ostgaard, D. (1998). *Team characteristics: Preferences and linkages to work attitudes* [Publication No. 9826847] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

Paivio, A. (1969). Mental imagery in associative learning and memory. P*sychological Review*, *76*, 241-263. Paivio, A. (1971). *Imagery and verbal processes*. Holt, Rinehart, and Winston.

*Parker, D. L. (1989). *Organizational culture and support for innovation: A policy-capturing approach* [Publication No. 9023203] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Paterson, B., Dowding, D., Harries, C., Cassells, C., Morrison, R., & Niven, C. (2008). Managing the risk of suicide in acute psychiatric inpatients: A clinical judgement analysis of staff predictions of imminent suicide risk. *Journal of Mental Health*, *17*, 410-423.

*Pearson, A. M. (2009). *An exploratory examination of the impact of web functionality across the customer service life cycle: A multi-criteria approach* [Publication No. 3358703] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Phelps, R. H. (1978). *Expert livestock judgment: A descriptive analysis of the development of expertise* [Publication No. 78-2413] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Polk, K. M. (2017). *Can interview structure be manipulated to enhance applicant reactions* [Publication No. 13837215] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.

*Porter, C. M., Parrigon, S. E., Woo, S. E., Saef, R. M., & Tay, L. (2016). Cultural and intellectual openness differentially relate to social judgments of potential work partners. *Journal of Personality*, *85*, 632-642.

*Qureshi, A. A. (1993). *An investigation of auditor judgment in the evaluation of contingent legal liabilities* [Publication No. 9416553] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Raaphorst, N., Groeneveld, S., & Van de Walle, S. (2018). Do tax officials use double standards in evaluating citizen-clients? A policy-capturing study among Dutch frontline tax officials. *Public Administration*, *96*, 134-153.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology*, *53*, 245-277.

*Reuer, J. J., Tyler, B. B., Tong, T. W., & Wu, C. W. (2012). Executives' assessments of international joint ventures in China: A multi-theoretical investigation. *Management and Organization Review*, *8*, 311-340.

*Reynolds, M. (2019). *Predictors of interest in voluntary participation for mindfulness training* [Unpublished master's thesis]. San Diego State University.

Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker*. Harvard University Press.

Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments* (pp. 15-67). SAGE.

*Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology*, *87*, 66-80.

*Russell, C. J., & Van Sell, M. (2012). A closer look at decisions to quit. *Organizational Behavior and Human Decision Processes*, *117*, 125-137.

Rynes, S. L., Schwab, D. P., & Heneman, H. G. (1983). The role of pay and market pay variability in job application decisions. *Organizational Behavior and Human Performance*, *31*, 353-364.

*Scheuch, G. (1983). *A social judgment theory analysis of pregnancy resolution policies among college students and counselors (abortion)* [Publication No. 8417999] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). SAGE.

Scho¨nbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609-612.

*Sekiguchi, T. (2003). *The role of person–organization fit and person–job* *fit in managers' hiring decisions: The effects of work status and occupational characteristics of job openings* [Publication No. 3102714] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Sekiguchi, T., & Huber, V. L. (2011). The use of person–organization fit and person–job fit information in making selection decisions. *Organizational Behavior and Human Decision Processes*, *116*, 203-216.

Shanteau, J., & Stewart, T. R. (1992). Why study expert decision making? Some historical perspectives and comments. *Organizational Behavior and Human Decision Processes*, *53*, 95-106.

*Shepherd, D. A. (1999). Venture capitalists' assessment of new venture survival. *Management Science*, *45*, 621-632.

*Shepherd, D. A., & Zacharakis, A. (2003). A new venture's cognitive legitimacy: An assessment by customers. *Journal of Small Business Management*, *41*, 148-167.

*Sherer, P. D., Schwab, D. P., & Heneman, H. G. (1987). Managerial salary-raise decisions: A policy-capturing approach. *Personnel Psychology*, *40*, 27-38.

*Silva-Castan, J. R. (1996). *The selection of institutions of higher education as consultants by small business owner/managers: A policy-capturing approach* [Publication No. 9704935] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Sinclair, A. L. (2003). *Disentangling contributions of process elements to the fair process effect: A policy- capturing approach* [Publication No. 3087505] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger (Editor-in-Chief) *APA handbook of testing and assessment in psychology* (Vol. 1, pp. 61-84). American Psychological Association.

*Soltani, B. (2013). *Organizational citizenship behavior, gender, and performance ratings: Does the rater matter?* [Unpublished doctoral dissertation]. San Diego State University.

*Stewart, T. R., Middleton, P., Downton, M., & Ely, D. (1984). Judgments of photographs vs. field observations in studies of perception and judgment of the visual environment. *Journal of Environmental Psychology*, *4*, 283-302.

*Stewart, T. R., Moninger, W. R., Brady, R. H., Merrem, F. H., Stewart, T. R., & Grassia, J. (1989). Analysis of expert judgment in a hail forecasting experiment. *Weather and Forecasting*, *4*, 24-34.

*Strawser, J. R. (1985). *An empirical investigation of auditor judgment: Factors affecting perceived audit risk* [Publication No. 8605222] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Teas, R. K. (1985). An analysis of the temporal stability and structural reliability of metric conjoint analysis procedures. *Journal of the Academy of Marketing Science*, *13*, 122-142.

*Tews, M. J., Stafford, K., & Zhu, J. (2009). Beauty revisited: The impact of attractiveness, ability, and personality in the assessment of employment suitability. *International Journal of Selection and Assessment*, *17*, 92-100.

*Thomas, F. D., III. (2009). *Using judgment analysis to identify at-risk drivers and to evaluate the effectiveness of training for changing drivers' perceptions of crash risk* [Publication No. 3383930] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Thompson, N. J. (2013). *Leader effectiveness in the eye of the beholder: Self-affirming implicit policies in leader perception* [Publication No. 10598026] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.

*Tomassetti, A. J. (2017). *On the reliability of judgments made in policy capturing measures: A meta-analytic review and experimental analysis* [Unpublished doctoral dissertation]. George Mason University.

*Tomassetti, A. J., Dalal, R. S., & Kaplan, S. A. (2016). Is policy capturing really more resistant than traditional self-report techniques to socially desirable responding? *Organizational Research Methods*, *19*, 255-285.

*Tomkovick, C. L. (1993). *An empirical investigation of product specific and product non-specific factors in making product continuance/discontinuance decisions* [Publication No. 9328830] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J., .. .Borns, J. (2020). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*. Advance online publication. https://doi.org/10.1007/s10869-020-09695-3

*Trouba, E. J. (2007). *A person-organization fit study of the Big Five personality model and attraction to organizations with varying compensation system characteristics* [Publication No. 3321996] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Turner, S. R. (2008). *The impact of role conceptualization on the process*

and outcomes of decision making in an educational context [Unpublished doctoral dissertation]. The University of Tennessee.
*Tye, M. G. (2005). *Selection of expatriates: A study of decision-making models* [Publication No. 3200703] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Vega, R. P. (2015). *Why use flexible work arrangements? A policy capturing study examining the factors related to flexible work arrangement utilization* [Publication No. 3706954] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48.
*Viswesvaran, C., & Barrick, M. R. (1992). Decision-making effects on compensation surveys: Implications for market wages. *Journal of Applied Psychology*, *77*, 588-597.
Viswesvaran, C., Schmidt, F. L., & Deshpande, S. P. (1994). A meta-analytic method for testing hypotheses about clusters of decision makers. *Organizational Behavior and Human Decision Processes*, *58*, 304-321.
*Voegel, J. A. (2016). *An exploratory examination of the ethical decision making process of entrepreneurs through the theory of planned behavior lens: A policy-capturing approach* [Publication No. 10163453] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.
*Wang, J. (2005). *Factors determining Taiwanese nurses' clinical judgments about hospitalized elderly patients with acute confusion* [Publication No. 3208335] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Warnick, B. J., Murnieks, C. Y., Mcmullen, J. S., & Brooks, W. T. (2018). Passion for entrepreneurship or passion for the product? A conjoint analysis of angel and VC decision-making. *Journal of Business Venturing*, *33*, 315-332.
*Webber, S. A. (1994). *A comparison of judgmental decision modeling techniques: Models of internal control procedure evaluation using the analytic hierarchy process and policy-capturing ANOVA* [Publication No. 9509744] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
Wegman, L. A., Hoffman, B. J., Carter, N. T., Twenge, J. M., & Guenole, N. (2018). Placing job characteristics in context: Cross-temporal meta-analysis of changes in job characteristics since 1975. *Journal of Management*, *44*, 352-386.

*Werner, S., & Ones, D. S. (2000). Determinants of perceived pay inequities: The effects of comparison other characteristics and pay-system communication. *Journal of Applied Social Psychology*, *30*, 1281-1309.

*Wexler, B. R. (2017). *Beauty is beneficial: An examination of candidate facial attractiveness, gender, qualification, and customer visibility on online recruitment intentions* [Publication No. 10270836] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.

*Whorton, R. (2015). *Marked: A policy capturing investigation of job applicant tattoos as stigmatizing marks in blue and white collar employment* [Publication No. 3710439] [Doctoral dissertation]. ProQuest Dissertations & Thesis Global.

Williams, L. J., O'Boyle, E. H., & Yu, J. J. (2020). Condition 9 and 10 tests of model confirmation: A review of James, Mulaik, and Brett (1982) and contemporary alternatives. *Organizational Research Methods*, *23*, 6-29.

Wilson, D. B. (2006). *Meta-analysis macros for SAS, SPSS, and Stata*. http://mason.gmu.edu/~dwilsonb/ma. html

*Wood, M. S., Bylund, P., & Bradley, S. (2016). The influence of tax and regulatory policies on entrepreneurs' opportunity evaluation decisions. *Management Decision*, *54*, 1160-1182.

*Wood, M. S., & Williams, D. W. (2014). Opportunity evaluation as rule-based decision making. *Journal of Management Studies*, *51*, 573-602.

*Yang, H., Thompson, C., Hamm, R. M., Bland, M., & Foster, A. (2013). The effect of improving task representativeness on capturing nurses' risk assessment judgements: A comparison of written case simulations and physical simulations. *BMC Medical Informatics and Decision Making*, *13*, 62-73.

*Yarab, P. E. (1999). *Not all rivals are created equal: A policy capturing approach to examining jealousy and threat based on the characteristics of a rival* [Publication No. 9950994] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.

*Zacharakis, A. L., McMullen, J. S., & Shepherd, D. A. (2007). Venture capitalists' decision policies across three countries: an institutional theory perspective. *Journal of International Business Studies*, *38*, 691-708.

*Zacher, H., Dirkers, T. B., Korek, S., & Hughes, B. (2017). Age-differential effects of job characteristics on job attraction: A policy-capturing study. *Frontiers in Psychology*, *8*, Article 1124. https://doi.org/10.3389/

fpsyg.2017.01124
Zedeck, S. (1977). An information processing model and approach to the study of motivation. *Organizational Behavior and Human Performance*, *18*, 47-77.
*Zhang, L. (2008). *Corporate social responsibility, applicants' ethical predispositions, and organizational attraction: A person–organization fit perspective* [Publication No. 3288923] [Doctoral dissertation]. ProQuest Dissertations & Theses Global.
*Zhang, L., & Gowan, M. A. (2012). Corporate social responsibility, applicants' individual traits, and organizational attraction: A person–organization fit perspective. *Journal of Business and Psychology*, *27*, 345-362.

Author Biographies

Ze Zhu received her PhD in industrial and organizational psychology from George Mason University. Her research interests include employee well-being and research methods.

Alan J. Tomassetti, PhD, is a Senior Human Resource Consultant with CPS HR Consulting's licensure and certification unit. He graduated from George Mason University in 2017.

Reeshad S. Dalal received his PhD in 2003 from the University of Illinois at Urbana-Champaign and is currently a professor of industrial and organizational *psychology* at George Mason University (Fairfax, VA, U.S.A). He is an associate editor at the *Journal of Business and Psychology*, is or has been on the editorial boards of several leading journals, and is a Fellow of the Association for Psychological Science and the Society for Industrial and Organizational Psychology. His life goal is none other than to replicate the findings from the “Lady Tasting Tea” experiment.

Shannon W. Schrader is currently a fifth-year clinical psychology doctoral student at George Mason Uni- versity. Shannon’s program of research examines the multidimensionality of self-control within individuals involved in the criminal legal system. Her research is funded by a Ruth L. Kirschstein NRSA F-31 award through the National Institute on Drug Abuse.

Kevin Loo is a first-year doctoral student in the industrial and organizational psychology program at George Mason University. His research interests include various topics in occupational health psychology (e.g. employee well-being and work-related stress) and organizational citizenship behaviors.

Isaac E. Sabat, PhD, is an assistant professor in industrial and organizational psychology and Diversity Sciences at Texas A&M University. His program of research broadly focuses on understanding and improving the working lives of stigmatized employees. This work has been published in *Journal of Business and Psychol- ogy*, *Journal of Organizational Behavior*, *Journal of Vocational Behavior*, and *Harvard Business Review*.

Balca Alaybek received her PhD in industrial and organizational psychology from George Mason University and MBA in management of technology from New Jersey Institute of Technology. Her research interests include judgment and decision-making, job performance, and research methods.

You Zhou is currently a doctoral student in the industrial and organizational psychology program at the University of Minnesota, Twin Cities. Her major research interests are performance, motivation and personnel selection methods.

Chelsea Jones received her bachelor’s degree from North Carolina State University in Economics. She currently works in information technology consulting.

Shea Fyffe is a doctoral student in industrial and organizational psychology at George Mason University. His current research interests include alternative forms of personality assessment, psychological measurement, natural language processing, machine learning, research methods and applications of programming to the social sciences.