Journal of Applied Psychology
2017, Vol. 102, No. 3, 274–290

0021-9010/17/$12.00 http://dx.doi.org/10.1037/apl0000163

Twilight of Dawn or of Evening? A Century of Research Methods in the *Journal of Applied Psychology*

Jose M. Cortina
George Mason University

Herman Aguinis
George Washington University

Richard P. DeShon
Michigan State University

We offer a critical review and synthesis of research methods in the first century of the *Journal of Applied Psychology*. We divide the chronology into 6 periods. The first emphasizes the first few issues of the journal, which, in many ways, set us on a methodological course that we sail to this day, and then takes us through the mid-1920s. The second is the period through World War II, in which we see the roots of modern methodological concepts and techniques, including a transition from a discovery orientation to a hypotheticodeductive model orientation. The third takes us through roughly 1970, a period in which many of our modern-day practices were formed, such as reliance on null hypothesis significance testing. The fourth, from 1970 through 1989, sees an emphasis on the development of measures of critical constructs. The fifth takes us into the present, which is marked by greater plurality regarding data-analytic approaches. Finally, we offer a glimpse of possible and, from our perspective, desirable futures regarding research methods. Specifically, we highlight the need to conduct replications; study the exceptional and not just the average; improve the quality of the review process, particularly regarding methodological issues; emphasize design and measurement issues; and build and test more specific theories.

Keywords: Research methods, historical review, theory

Supplemental materials: http://dx.doi.org/10.1037/apl0000163.supp

> Psychology is just beginning, the best things are yet to be found out . . . its difficulties and obscurities are the twilight of dawn and not that of evening.
>
> —G. Stanley Hall, "Practical Relations Between Psychology and the War"

When G. Stanley Hall, John Wallace Baird, L. R. Geissler, and the other fathers of the *Journal of Applied Psychology* (*JAP*) were attempting to chart and communicate a course for the journal, they seemed keen to make the point that there were many real-world problems for which psychological science would be useful. They also wanted to make clear that we as a field knew very little about what the solutions to those problems would be or how we should go about finding them. Nevertheless, they were optimistic. As Hall put it, our vision was obscured in the "twilight of dawn" rather than the twilight of evening (Hall, 1917, p. 10).

Those early writers also made a case that ours needed to be a field that embraced more sophisticated and objective research methods than did other areas of psychology. We were to address issues of relevance—to the war effort, to private industry, to government—using something more than the opinions of educated men. As a result, a considerable percentage of the papers published in the early years of *JAP* were methodological in nature, from comparisons of different intelligence measures by Yerkes (1917) and by Wells (1917), to the norming of equilibrium tests by Burtt (1918), to item writing (Wembridge, 1918), to rater bias (Thorndike, 1920), and on and on. Indeed, the first empirical paper published in the journal was a validation study (Terman et al., 1917).

This emphasis on rigorous empiricism set *JAP* apart from other psychology journals of its time (e.g., *Psychological Bulletin*, *Psychological Review*), a distinction that, in many ways, continues to this day. The importance attached to rigorous methodology promulgated training in methods, which led to a virtuous cycle of each generation building on the methodological savvy of the previous. This led, over time, to developments in the understanding of measurement error, validation, meta-analysis, rater agreement and aggregation, item response theory (IRT), moderating effects, multilevel modeling, and many other methodological areas, with clear implications for substantive theories and research in many domains.

The purpose of our article is to trace the history and also influence the future of research methods in the journal as well as

This article was published Online First February 2, 2017.

Jose M. Cortina, Department of Psychology, George Mason University; Herman Aguinis, Department of Management, School of Business,, George Washington University; Richard P. DeShon, Department of Psychology, Michigan State University.

We thank Gilad Chen, Steve W. J. Kozlowski, and Eduardo Salas for comments on previous drafts.

Correspondence concerning this article should be addressed to Jose M. Cortina, Department of Psychology, George Mason University, 4400 University Drive, 3F5, Fairfax, VA 22030. E-mail: jcortina@gmu.edu

in applied psychology and related fields. We divide the chronology into six periods, and a summary of this chronology is included in Figures 1 and 2, with Figure 1 focusing on major topics and Figure 2 focusing on a selective subset of influential articles (the online supplemental materials include additional information regarding the rationale for structuring our manuscript as shown in Figure 1, and the selection of articles included in Figure 2). The first section of the paper takes us through the mid-1920s, which, in many ways, set us on a methodological course that we sail to this day. The second is the period through World War II, in which we see the roots of modern methodological concepts and techniques. The third takes us through roughly 1970, which is when many of our modern-day methodological concepts and techniques were formed. The fourth, from 1970 through 1989, sees an emphasis on the development of good measures of critical constructs. The fifth takes us through the recent past, which is marked by a plurality regarding data-analytic approaches. The sixth offers a glimpse of possible and, from our perspective, desirable futures.

We arrived at these particular periods after a review and content analysis of all articles published in *JAP* since 1917. Although other cutoff points could be used, the ones that we chose allowed us to partition the past century of research into pseudocategorical blocks that can be distinguished from each other.

1917–1925: Who Are You?

At its inception, *JAP* was the first outlet for work in a variety of areas of psychological science devoted to solving real-world problems, none of which had their own labels, much less their own journals. Instead, we all fell under the heading Applied Psychology, and *JAP* was our home. Although some of the distinctions that the founders of applied psychology wished to draw related to the questions that were asked, another part of the distinction came from the methods used to answer those questions. For example, in the third paragraph of their foreword to the new journal, Hall, Baird, and Geissler (1918) suggested that there are many psychologists who are "clamoring for more effective methods of diagnosing character and intellectual equipment" (p. 6), an observation that was certainly borne out in the early emphasis on testing.

From a research methods standpoint, the first few issues of *JAP* show a desire for objectivity, not as an end of itself but as a means to valid conclusions. It was easy enough, in retrospect, to see how absurd it had been for Goddard to base conclusions about the intelligence of immigrants on their answers to questions like "Who is Christy Matthewson?" (a great pitcher, long since gone) and "What is Crisco?" (a mysterious, artery-clogging substance that is, strangely, still around), an approach that led him to classify 80% of Ellis Island's arrivals as "feebleminded" (Hothersall, 1990). We needed methods that would lead us to appropriate and useful conclusions. And, indeed, we see authors of articles in the first few issues of *JAP* addressing this need.

In particular, we would call attention to two emphases. The first is cognitive ability testing. Thus, we see a great deal of early work on ability testing such as Terman et al. (1917) developing and norming ability tests for police and firefighters, and Yerkes (1917)

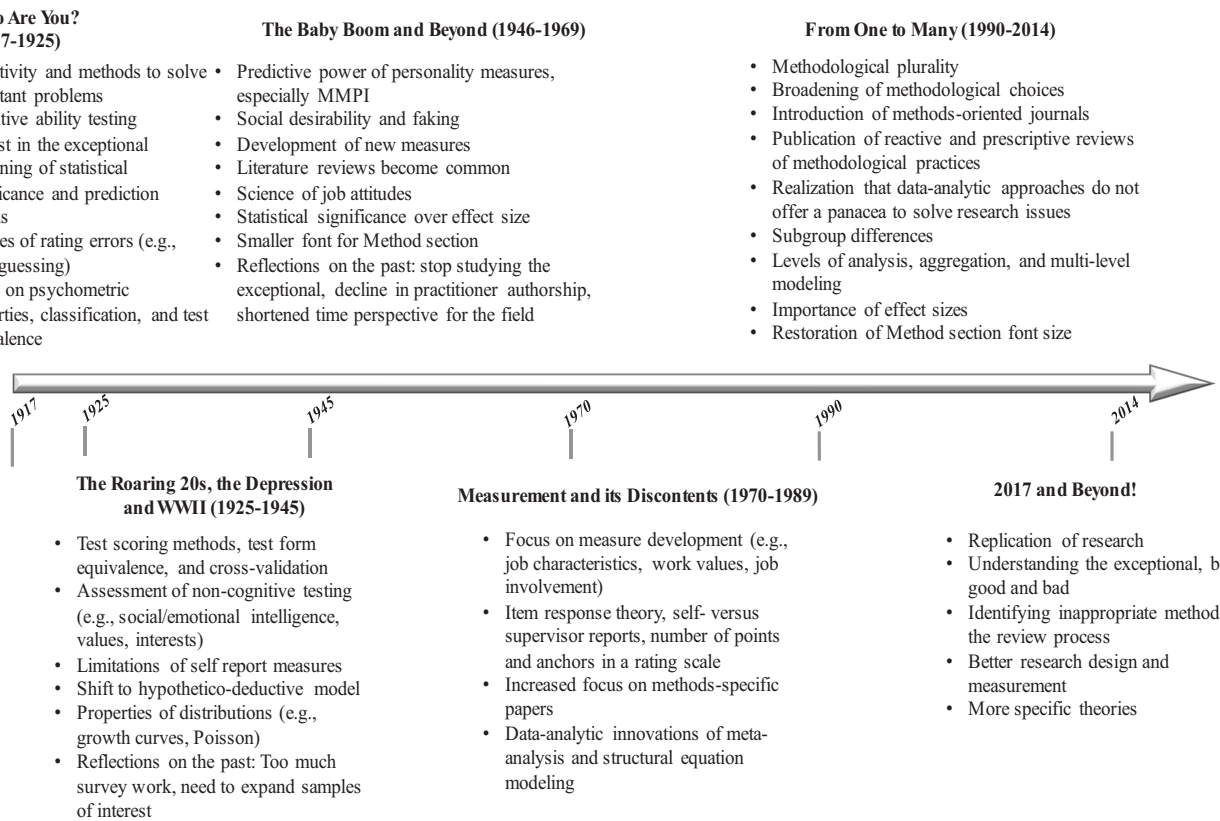

Figure 1. A chronology of methodological topics in the *Journal of Applied Psychology* (1917–present).

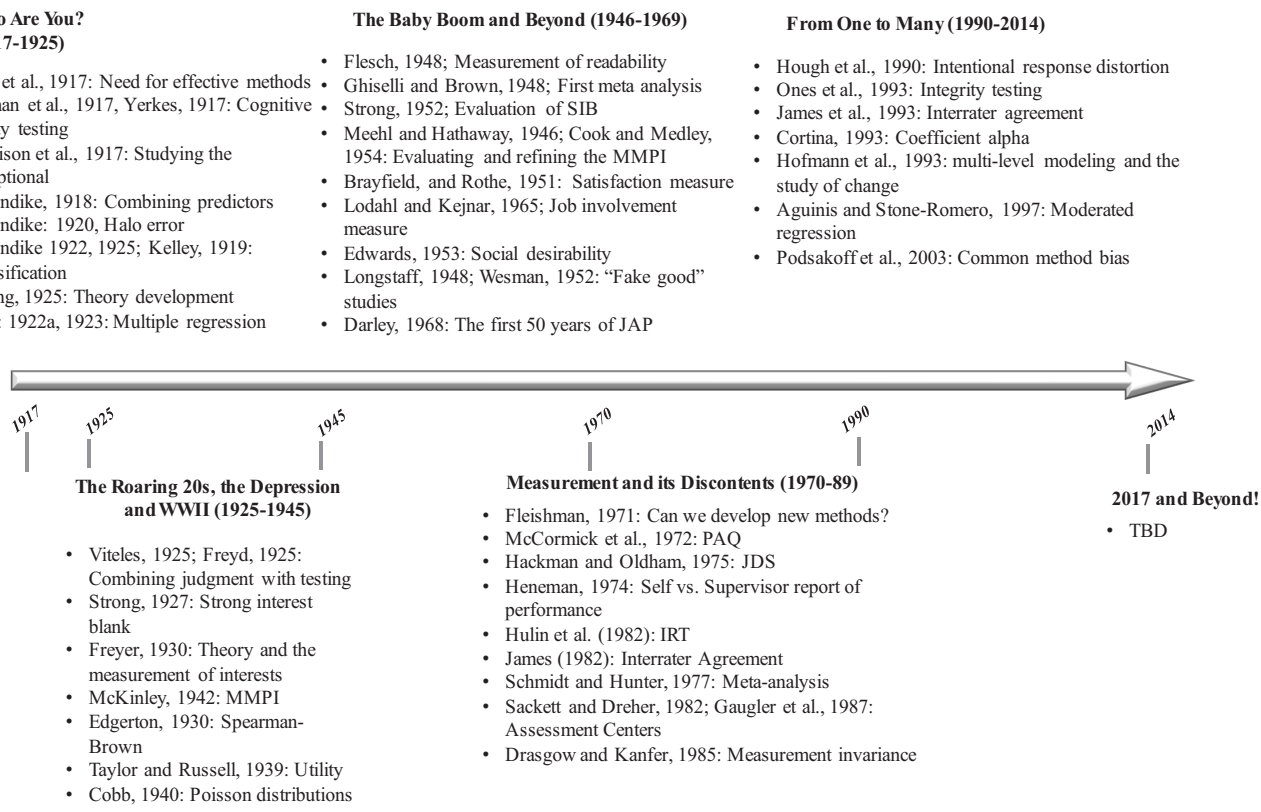

Figure 2. A selective chronology of influential methods papers in the *Journal of Applied Psychology* (1917–present).

comparing methods of measuring intelligence. Such studies formed the basis for the research to come regarding subgroup differences, norming, scoring, and criterion-related validity, all of which are inextricably related to methodological issues.

A second emphasis contained in the first few issues is noteworthy because it has since disappeared almost entirely. There was a great deal of interest in the exceptional, that is, the notably high and low scorers on a given measure. As Garrison, Burke, and Hollingworth (1917) put it,

> If we had a scientific record of the mental status and development of J. Stuart Mill, of Thomas Edison, of Madame Curie, of Abraham Lincoln and of George Eliot up to the age of fifteen years, or if we knew the intelligence quotients of all the Nobel prize winners at the age of eight years, what guidance for educational practice might be contained therein? (p. 102)

Coy (1918) and Terman (1918) conducted case studies of exceptional children. Gates (1918) conducted a case study of an exceptional marksman, a Mr. William Blasse. Like the others, Gates conducted a detailed analysis of a single individual in order to determine the attributes that set him apart in his domain of expertise. At the other end of the spectrum, in a paper titled "The Moron as a War Problem," Mateer (1917) called for research on the unique costs and opportunities presented by low-intelligence soldiers. The case study method was a frequent approach for increasing our understanding of the exceptional.

This sort of case study approach has been seen very rarely in *JAP* over the last 70 years or so. Is it because we are only interested in those near the middle of distributions? We think not (see Aguinis, Gottfredson, & Joo, 2013, for a detailed discussion of this issue). Instead, the emphasis of our field on quantitative analysis generally and statistical significance particularly precludes the research methods that allow for the study of the exceptional. We have detailed guidelines for conducting case research (e.g., Bitektine, 2007; Piekkari, Welch, & Paavilainen, 2009), and we might gain important insights about organizations by applying rigorous case study and other qualitative methods in order to understand the exceptional.

In any case, the launch of *JAP* in 1917 was highly successful as evidenced by the fact that, excluding book reviews, the journal published an average of 34.12 research papers per volume in its first several years. Applied psychological research was blossoming, and *JAP* was the natural home for this sort of work.

During this time, scholars grappled with the tension of solidifying the paradigm for applied psychological research, while remaining open to new methods, new problems, and new ways of thinking about applied psychological problems. When considering the research methods used in this period, it is important to recognize that the field of inferential statistics was new. Fisher had not yet published his revolutionary texts on statistical inference (Fisher, 1925) and experimental methods (Fisher, 1935). Neyman and Pearson had not published their hypothesis testing framework (i.e., Neyman & Pearson, 1933). For the most part, scientific inference in this period was based on the eyeball test, and data collections referred to as "experiments" were actually observations

taken under controlled settings but lacking control groups and random assignment. Toward the end of the period, the phrase "statistically significant" began to be used (e.g., Paterson & Langlie, 1925), but details on how significance was determined were not provided clearly or systematically.

The dominant research paradigm during these early years was atheoretical and focused on the measurement of practically relevant individual and group differences such as intelligence (e.g., Pintner, 1919; Thorndike, 1919), aptitudes (e.g., aptitude for flying; Henmon, 1919), traits such as aggressiveness (e.g., Moore & Gilliland, 1921), and vocational interests (e.g., Freyd, 1922). Group norms and comparisons between age, race, and gender groups on these measures were commonplace. Pearson's product-moment correlation served as the other dominant descriptive statistic to examine relations between tests (e.g., Stevenson, 1918; Terman & Chamberlain, 1918) and to link those tests to outcomes such as school grades (e.g., Burtt & Arps, 1920) and supervisor ratings (Flanders, 1918). This early measurement and testing literature is disturbingly familiar in the sense that, at present, we seem to be struggling with many of the same issues with which scholars were wrestling 100 years ago (Aguinis, Culpepper, & Pierce, 2010, 2016).

Given the intense focus on measurement in this period, it is not surprising to find a burgeoning interest in the psychometric properties of measures. Thorndike's (1920) paper on halo error in ratings is the most cited paper in *JAP* during this period. The reliability and validity of tests and assessments received a great deal of attention (e.g., Gates, 1923; Root, 1921; Slawson, 1922; Ruch & Del Manzo, 1923). Factors that contributed to unreliability and invalidity, such as guessing (Chapman, 1922) and item stems containing double negatives (Wembridge, 1918), began to receive attention. Finally, Thorndike (1920, 1922, 1925) provided the foundations of test-equating methods used to make scores from different forms of the same test interchangeable.

The beginnings of more complex prediction models and the use of multiple regression are also found in this early period. Thorndike (1918) provided a remarkably sophisticated treatment of methods for combining predictors that exhibit linear, monotonic, nonmonotonic, and nonlinear relations with job performance. Thurstone (1919) examined the zero-order and multiple correlation of eight mental tests with telegraphy speed and found results virtually indistinguishable from those found today. Hull (1923) provided a formal introduction to the use of linear multiple regression and multiple correlation for predicting a criterion from a set of predictors, and also introduced the concept of rescaling scores using affine transformations to make both the scores and the regression results more interpretable (Hull, 1922b, 1923).

We conclude the discussion of this period with descriptions of a number of important innovations and insights that provide strong hints of what was to come in subsequent periods. Strong (1918) and Kelley (1919) provided detailed treatments of the problem of classification. Oddly, classification became highly important in subsequent years, but is almost completely absent from modern selection research and practice. Pressey (1921) and Sturges (1924) provided initial forays into the concepts underlying modern-day sampling theory. Kohs and Irle (1920) and Bedaux (1921) discussed problems with the conceptualization and assessment of the performance criterion presaging the rise of the criterion problem (a history of the criterion problem can be found in Austin & Villanova, 1992). In a highly cited paper entitled "Theories of Selling," Strong (1925) provided one of the first treatments of applied theory and theory development. Finally, it is remarkable to note that the examination of average growth curves was relatively common in this time period (e.g., Burtt & Dobell, 1925; Chapman, 1919; Kuhlmann, 1921; Thorndike, 1917), but then this longitudinal method largely disappeared from the journal for the next 70 years.

In sum, many of the most influential papers published in the journal during this time were methodological in nature. We see the early treatment of various methodological topics that would become staples later on (e.g., correlational analysis, halo, test norming, experience sampling). In many ways, these first few years set a tone for methods that would define research and training in the field for decades to come.

The Roaring '20s, the Depression, and World War II

It is in this period that we see a topical diaspora of sorts. Whereas a great deal of early emphasis was placed on ability testing, this period sees a dramatic expansion of topics within the purview of our field. Interest in ability testing continued, however. Paterson and Langlie (1925); Arthur (1925), and Symonds (1925) compared different scoring methods. It is also in this period that we see the first mention of cross-validation as a method of test evaluation (Kornhauser, 1927).

We see an expansion of topics related to testing of attributes other than ability. For example, although interest in interest, as it were, began just after the end of World War I (e.g., Freyd, 1922; Watts, 1921), it was in the late 1920s that E. K. Strong's interest blanks were developed (Strong, 1926, 1927). Many of the norming and evaluation studies for these tests were published in *JAP* (e.g., Freyd, 1925; Strong, 1927; Strong & Green, 1932). We see early mention of measurement of social intelligence and emotional insight (Hunt, 1928; Tendler, 1930), values (Sarbin & Berdie, 1940), and student skills (Locke, 1940), to name a very few illustrations.

It was also in this period that personality and its measurement began to take hold. And with this interest came a greater appreciation of the limitations of self-reports, even as one of the most influential self-report measures in history, the Minnesota Multiphasic Personality Inventory (MMPI), was being developed (e.g., McKinley & Hathaway, 1942). Projective paper-and-pencil techniques were developed to measure not only personality (e.g., Manson, 1925) but also attitudes (Vernon, 1930) and interests (Freyer, 1930). Indeed, we can see in the Manson (1925) paper, with its surreptitious use of ability tests to coax personality information from respondents, the roots of Larry James's work on conditional reasoning seven decades later (L. R. James, 1998).

We also see the seedlings of research methods topics that would develop over the next half century. Cureton and Dunlap (1930) demonstrated an early appreciation of Fisher's work on distributions. In Edgerton's (1930) work on Spearman Brown prophecy values, we see a recognition of sampling error. Jordan (1930) examined growth curves in ability scores. Anderson (1935) culled items on the basis of a discrimination parameter. Taylor and Russell (1939) developed their utility tables. Cobb (1940) explained the importance of the Poisson distribution for accident data.

Another characteristic of this period is that we begin to see the transition from a discovery model to a hypotheticodeductive model. The vast majority of papers in this period are still "reporters" (cf. Colquitt & Zapata-Phelan, 2007). This is to say that most papers asked a question without speculating about the answer, collected data, and reported results. We also see, however, the first "testers" and "builders" of theories. Manson's (1925) work, mentioned earlier, was based on a theory of guessing behavior. Freyer's (1930) study of interest measurement was based on an "acceptance-rejection theory." In fact, Schiller (1935) may represent the first example of a comprehensive theory (of handedness) to appear in the journal. In any case, it is in this period that we see a much needed pendulum swing from pure empiricism toward theory testing.

Finally, it seems that the field had been in existence long enough for there to be an evaluation of where we were and where we should be headed. Viteles wrestled with Freyd and others regarding the role of "clinical judgment" as opposed or in addition to standardized test scores (e.g., Freyd, 1925; Viteles, 1925). Ruckmick (1930) claimed that there was too much survey research conducted by those who did not understand its limitations. He referred to survey research as "prescientific." One wonders how Ruckmick would evaluate our progress since then.

In summary, we see in this period a shift from discovery to hypothesis testing, a shift that continued to the present day. We also see an expansion from cognitive ability testing to the measurement of many other workplace-relevant attributes, most notably interests. Finally, we see some introspection as a field, with early attempts to compare the field as it was with the field as it perhaps should have been.

The Baby Boom and Beyond: 1946–1969

It is in the period of time from the end of World War II (WWII) to the middle of the Vietnam conflict that we see the field begin to take the methodological shape into which it has since solidified. To be sure, many of the topics that had generated interest in the previous time period continued to do so. For example, there was a continued emphasis on the measurement of interests and their role in prediction, spurred on perhaps by Strong's (1952) 19-year follow up study showing remarkable predictive power of his measure, the Strong Interest Blank, over time.

Similarly, there was a great deal of interest in the characteristics of different measures of personality and the strengths and weaknesses of different measurement approaches. H. O. Schmidt (1945) described efforts to norm MMPI scales, for example. In two of the most cited papers of the period, Meehl and Hathaway (1946) and Cook and Medley (1954) added and evaluated refinements to the MMPI. A. L. Edwards's (1953) highly influential paper on social desirability also appeared in this period. We also see some new approaches to personality measurement. Krathwohl (1952) inferred personality from IQ–GPA discrepancies. Longstaff (1948) and Wesman (1952) conducted some of the first "fake good" studies of intentional distortion. What we learned from this work clearly was not comforting, as it was toward the end of this period that we as a field lost faith in personality as a predictor of performance.

It was in this period that the science of job attitudes began to develop. The cornerstone of this science was measurement. The paper by Brayfield and Rothe (1951), in which they developed their measure of job satisfaction, is the second most cited paper of the period. The third most cited was Lodahl and Kejner (1965), in which they defined and measured job involvement.

Many of our modern-day research practices were formed during this period. Paterson and Jenkins (1948) contains one of the first extensive literature reviews in an empirical paper. Georgopoulos, Mahoney, and Jones (1957) and Ziller, Behringer, and Goodchilds (1962) represent some of the first tests of formal theory. We also see early work on interrater agreement. Balinsky, Blum, and Dutka (1951) extended existing agreement formulas to judgments regarding product preference. We also see statistical significance testing begin to take hold. Richardson (1948) is an early example, but perhaps the most intriguing is Ghiselli and Brown (1948), which was also the first example of a meta-analysis that we were able to find. It is interesting that the words in results and discussion sections in the era immediately following WWII are still driven by magnitude of effect, albeit subjective evaluation thereof. By the end of the period in question (i.e., late 1960s), one is hard pressed to find any mention of effect size at all.

The end of this period marked the 50th anniversary of the journal. In his critical examination of the first half century of *JAP*, Darley (1968) made many important points. He lamented, as do we, that our field no longer studied the exceptional. He noted that the proportion of authors who were practitioners had fallen, and this proportion has continued to fall since then (Cascio & Aguinis, 2008). Of particular interest was Darley's description of the observation by Xhignesse and Osgood (1967) that "psychology as a science is in danger of 'forgetting where it has been,' of repeatedly rediscovering facts and theories that have been well worked in the past [p. 790]." Given the proliferation of constructs and theories since Darley wrote those words (see Leavitt, Mitchell, & Peterson, 2010), and our unwillingness to prune those theories, it may be that the same criticism could be leveled today.

Perhaps the most worrisome characteristic of this period, however, is a seemingly cosmetic change that we consider to be important, and that has, to our knowledge, escaped notice entirely. In 1954, the format of the journal was changed such that the font of the Method section was decreased. This was part of a growing trend. *Journal of Abnormal Psychology* made this shift in 1951, for example. Not all journals made this change (e.g., *Personnel Psychology*), but many did. The advantage of decreasing font size for any section is simple: Smaller font means lower printing and distribution costs. And if one were going to reduce the font size of a certain section, which section would one choose? The font size of footnotes is smaller because footnotes contain information that is either peripheral or relatively trivial and would detract from the larger message were it placed in the heart of the text. We postulate that the font size of the Method section could be reduced because it was deemed less important for the reader to digest methodological details than for the reader to digest the increasing detail of Introduction sections.

In 2007, *JAP* returned the size of the font in Method sections to that of other sections, a change of which we heartily approve. We wonder, however, what the effects were of relegating the information in Method sections to a level between main text and footnote. Might this change have led readers to gloss over Method sections in the same way that they often gloss over footnotes? One thing is

certain. This change would not have increased the scrutiny applied to methods.

We have now reached the halfway point in the history of *JAP*. This period saw tremendous strides forward regarding the development, evaluation, and refinement of measures. It also includes a growing appreciation of the limitations of self-report measurement. This would prompt new approaches to measurement in the years to come. Two of the cornerstones of today's research methods, theory development and significance testing, came into their own in this period. This period ended with Darley's look back, and many of the problems that he identified can also be seen today. Finally, we saw a relegation of Method sections to the fine print.

1970–1989: Measurement and Its Discontents

The early 1970s coincided with the appointment of Edwin A. Fleishman as *JAP*'s sixth editor. This was a time of enormous social and political changes and, in his inaugural editorial, Fleishman (1971) issued the following challenges:

> Can we apply the research knowledge and methods already developed toward the solution of pressing problems? Can we be sufficiently innovative to develop new methods for dealing with these research problems, often in an action-oriented setting, with sufficient scientific rigor to allow dependable generalizations? . . . Should not the *Journal of Applied Psychology* be a primary outlet for research addressed to these questions? (p. 1)

From a methodological perspective, the first step toward addressing Fleishman's (1971) call was to develop good measures of critical constructs. The early 1970s produced highly influential measurement instruments, many of which are still in use today. For example, Hackman and Oldham (1975) developed the Job Diagnostic Survey (JDS) and found that scores on the JDS were related to absenteeism, performance, general satisfaction, and work motivation. In a *JAP* monograph, McCormick, Jeanneret, and Mecham (1972) described the Position Analysis Questionnaire (PAQ) as an instrument to understand dimensions of human behavior for specific jobs. Moreover, the resulting data could be used to understand the extent to which seemingly different jobs share common behavioral requirements (i.e., "job elements"). Several additional measurement instruments were developed during this time. For example, Wollack, Goodale, Wijting, and Smith (1971) developed the survey of work values, and Wanous and Lawler (1972) compared nine different measures of job satisfaction based on data collected from 13 different jobs at a telephone company. Additional measures that were developed included organizational communication (Roberts & O'Reilly, 1974) and perceived organizational support for innovation (Siegel & Kaemmerer, 1978). The measure development trend extended into the 1980s, when instruments were developed to assess commitment to the union (Gordon, Philpot, Burt, Thompson, & Spiller, 1980—a *JAP* monograph), job involvement (Kanungo, 1982), and perceived supervisory power (Hinkin & Schriesheim, 1989), among others.

Not surprisingly, the increased attention devoted to the development of new measures led to improvements in the evaluation of measures. Hulin, Drasgow, and Komocar (1982) and Parsons and Hulin (1982) were the first papers on IRT published in the journal (note that these articles also involve applications of factor analysis, but not the development of new approaches or techniques directly addressing factor analysis per se). Within a few years, IRT techniques were the preeminent techniques for evaluating and comparing measures (e.g., Ironson, Smith, Brannick, Gibson, & Paul, 1989; Roznowski, 1989).

It was also during this period that Larry R. James began his seminal work on interrater reliability and agreement. James (1982) described aggregation bias, the distinction between ICC(1) and ICC(2), and their implications for climate research. James, Demaree, and Wolf (1984) developed $r_{wg(j)}$ as an index of agreement for a single group of judges on a single variable for a single target. As we describe later, this index was criticized by F. L. Schmidt and Hunter (1989), a criticism that was later challenged by Kozlowski and Hattrup (1992) and James et al. (1993). In any case, James (1982) and James et al. (1984) spurred a great deal of research on agreement (e.g., Lindell & Brandt, 1999; Burke & Dunlap, 2002) and composition (e.g., Chan, 1998), and were among the most influential papers of this period. This early work was relevant not only for climate research (e.g., Zohar, 2000, 2002), but for all domains that had transitioned from a sole emphasis on the individual level of analysis to examining phenomena at the within-person level (e.g., Totterdell, 2000), the team level (e.g., Simons & Peterson, 2000), and the organization level (e.g., Takeuchi, Lepak, Wang, & Takeuchi, 2007).

The barrage of new measures also led to the recognition that there were critical and thorny issues that compromised the validity of those measures, most of which were based on self-reports. Bass, Cascio, and O'Connor (1974) provided evidence that increasing the number of scale points also increases the potential overlap between a respondent's choices, leading to lack of precision in the resulting scores. Heneman (1974) found that self-reported performance scores differed substantially from scores provided by supervisors, and he suggested that self-reports of performance seemed to be particularly useful when resulting scores are used for research as opposed to evaluative purposes. In what seems to be the first Monte Carlo study published in *JAP*, Lissitz and Green (1975) simulated the effect of the number of scale points on reliability. They concluded that there is little improvement in reliability if a scale includes more than five anchors, challenging a fairly standard contemporary practice of using 7-point scales.

These and other studies addressing questions about the accuracy of data collected using available measures continued to be published in the 1980s, culminating with the publication of seminal articles that opened up new lines of research over the following decade and beyond: Feldman and Lynch (1988) guided subsequent work on common method bias, and Drasgow and Kanfer (1985) guided subsequent work on measurement invariance. In fact, in an early sign of the future importance of common method bias, John P. Campbell, who followed Edwin A. Fleishman in the editor role, reflected in his outgoing editorial that "there were few degrees of freedom within which to be a gatekeeper," with perhaps one exception: the "use of a self-report questionnaire to measure *all* the variables in a study . . . I am biased against the study and believe that it contributes very little" (Campbell, 1982, p. 692).

The development and improvement of measurement instruments was quickly followed by two immensely influential data-analytic innovations: meta-analysis and structural equation modeling (SEM). Particularly within the domain of preemployment testing, work published in the 1960s and earlier (e.g., Guion, 1965) had concluded that validity coefficients change from organization to

organization and, therefore, are situation-specific. In a seminal article challenging the situational specificity hypothesis, F. L. Schmidt and Hunter (1977) described a new analytic approach, a type of meta-analysis that could be used to establish validity generalization (VG). Meta-analysis involved first assessing the degree of variability of validity coefficients across studies, and then calculating the extent to which such variability may be substantive or, instead, accounted for by measurement error and other methodological and statistical artifacts such as sampling error and range restriction.

Methods for synthesizing a body of literature quantitatively had been available for some decades (e.g., Ghiselli & Brown, 1948), but the F. L. Schmidt and Hunter (1977) approach set itself apart through its reliance on "corrections" for methodological and statistical artifacts. The establishment of VG through meta-analysis represented an important turning point in the field and led to a general belief that it is possible to gather a group of studies conducted using unreliable measures and then draw conclusions from mean relations across studies (Aguinis, Pierce, Bosco, Dalton, & Dalton, 2011). Within a few years of the publication of Schmidt and Hunter's article, there was a veritable explosion of VG studies, such as the *JAP* monograph on the validity of assessment centers (Gaugler, Rosenthal, Thornton, & Bentson, 1987), examination of the relation between age and job performance (McEvoy & Cascio, 1989; Waldman & Avolio, 1986), the effects of goal setting (Tubbs, 1986) and realistic job previews (Premack & Wanous, 1985), and of ratee race on performance ratings (Kraiger & Ford, 1985).

The second data-analytic innovation revolved around SEM. Particularly within the broader context of measurement development and improvement—and the assessment of overall measure quality—SEM was seen as a fundamental tool for understanding dimensionality (Sackett & Dreher, 1982), hierarchical structures (L. A. James & James, 1989), and relations between underlying constructs (L. J. Williams & Hazer, 1986). There were many examples of SEM being used to develop and refine measures (Ironson et al., 1989).

At the end of the 1980s, the zeitgeist was that the use of data-analytic solutions such as meta-analysis and SEM would mitigate challenges regarding measurement. Also, the introduction of powerful computers allowed researchers to conduct analyses at lightning speed compared with capabilities available just a few years earlier. Thus, a data-analytic as opposed to a research design solution was the practical and seemingly logical choice. From a methodological perspective, Fleishman got more than he bargained for. *JAP* published not only important and influential articles describing new measures that could be used to address research on socially relevant problems but also articles that identified problems, and offered some solutions, for many persistent measurement challenges (e.g., choosing the number of scale points, establishing measurement invariance, aggregating to higher levels) that are still relevant today.

In sum, this period began with the development of two of the most influential measures in the history of our field, the JDS and the PAQ. Shortly thereafter, VG/meta-analysis methods showed us that relations between some variables were less situation-specific than we had imagined, and SEM gave us a method for comprehensive model testing. By the time the next period in our chronology began, virtually every issue of *JAP* contained at least one application of meta-analysis and of SEM.

1990–2014: From One to Many

Neal Schmitt became the *JAP* editor in 1989 and drew attention to methodological issues that he considered important. Specifically, after processing manuscripts for a full year, he mentioned a trend that we identified in the previous section, and wrote that he was

> very surprised at the large number of authors who use LISREL [a software program to conduct SEM] as their data-analytic technique. As one who has published work using LISREL, I certainly support its use when appropriate . . . [but] it is *not* a requirement that authors use LISREL to publish in *JAP*. (Schmitt, 1989, p. 844).

Schmitt's (1989) observation reflected the need to expand our methodological repertoire. Novel methodological approaches such as meta-analysis and SEM, which were being used frequently, were certainly welcome and became popular across substantive domains ranging from integrity testing (Ones, Viswesvaran, & Schmidt, 1993) to job burnout (Lee & Ashforth, 1996) and leadership (Gerstner & Day, 1997). However, there was the realization that using any single methodological approach, no matter how potent, would not offer a silver-bullet answer to important theoretical and practical questions. Thus, the period beginning in the 1990s was marked by what we label a movement "from one to many." This movement toward increased methodological plurality involved conceptual, design, measurement, and analysis topics.

As examples of conceptual issues, there was a movement from assessing one type of relation between two variables (e.g., direct effects) to many (e.g., moderating and mediating effects; Tett & Burnett, 2003), and investigating one possible shape of the relation between two variables (e.g., linear) to many (i.e., curvilinear; Baer & Oldham, 2006).

Regarding design, there was a transition from implementing one type of research design (e.g., cross-sectional study based on self-reports) to others, including policy capturing (Kristof-Brown, Jansen, & Colbert, 2002) and extreme-group designs (Cortina & DeShon, 1998); from collecting data at one point in time to two or more time periods (e.g., Liden, Wayne, & Stilwell, 1993); from collecting data in one context to many contexts such as from inside and outside the organization (e.g., Ahearne, Bhattacharya, & Gruen, 2005); from a focus on one hierarchical level in the organization such as employees to other hierarchical levels as well (i.e., top management teams; e.g., West & Anderson, 1996); from one type of data source (i.e., self-reports) to many (e.g., peers, supervisors; e.g., Chiaburu & Harrison, 2008); and from between-person designs to experience sampling (e.g., K. J. Williams, Suls, Alliger, Learner, & Wan, 1991).

Regarding measurement, there was a movement from consideration of one source of measurement error (e.g., the use of different items) to consideration of many (e.g., the passage of time, the use of multiple raters; e.g., Cortina, 1993); from using one type of scale (i.e., Likert-type scales) to different scale formats (Maurer & Pierce, 1998); and from outcomes measured at one level of analysis only (i.e., employee) to many (e.g., team level; Kearney & Gebert, 2009). This period also saw the gradual extinction of measure development and validation papers, a strange occurrence

given the fact that so many of the most cited papers in the history of the journal described the development and validation of new measures.

Regarding data analysis, many articles were published that addressed refinements and improvements in procedures and the estimation of parameters within the context of multiple regression (e.g., Aguinis & Stone-Romero, 1997), meta-analysis (e.g., Aguinis & Whitehead, 1997), measurement equivalence (e.g., Raju, Laffitte, & Byrne, 2002), and multilevel modeling (e.g., Mathieu, Aguinis, Culpepper, & Chen, 2012), among many others. The period also saw the first uses in our field of multilevel modeling (also referred to as hierarchical linear, mixed-effect, random coefficient, and covariance components modeling). Hofmann, Jacobs, and Baratta (1993) provided a roadmap for the use of multilevel modeling in the study of change. Vancouver, Millsap, and Peters (1994) used it to study goal congruence and Zohar (2000) to study safety climate. Understanding multilevel modeling coincided nicely with an increased appreciation of the importance of levels issues more broadly.

There was also a good deal of work in this period that had important applied implications regarding subgroup differences. For example, Oswald and his colleagues investigated strategies for the development of biodata measures that reduce race differences (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Schmitt et al., 2003). De Corte, Lievens, and Sackett (2007) imported optimization techniques from the operations literature to determine optimal weighting schemes for maximizing validity while minimizing subgroup differences.

It is clear that the past quarter century involved a remarkable and even overwhelming broadening of the methodological choices available. Naturally, this methodological expansion resulted in an increased interest in methodology as a substantive topic itself. As we mentioned in previous sections, *JAP* had had a long-lasting interest in publishing articles addressing methodological topics. But the increased interest in methodology compelled editor Philip Bobko (1995) to write that "the primary focus of each submission should be the application and extension of psychological knowledge. Submissions that do not meet this criterion will continue to be returned to authors unreviewed" (p. 3). Bobko's 3-year editorial term was followed by Kevin R. Murphy's, who acknowledged the new reality of a broader methodological landscape explicitly and wrote that "we do not favor one set of methods over any others . . . we look forward to receiving . . . studies with a wide range of research methods" (Murphy, 1997, p. 4). Similar to Murphy, editor Sheldon Zedeck (2003) wrote that "we want to expand the approach to the topics that fit the JAP mission . . . First, we are broadening the methodological orientation" (p. 3). And, similar to Bobko, editor Steve W. J. Kozlowski (2009) mentioned,

> We are open to methodological articles, as long as they provide a clear conceptual contribution to research in applied psychology . . . . This has been the tradition of the *Journal of Applied Psychology* and it will continue to be our foundation. (p. 1)

The movement from one to many created important challenges. First, researchers were now faced with many more choices than in the past in terms of theory, design, measurement, and analysis. In many cases, it was not clear which would be the right choice and why. Hence, there was an urgent need for methodological guidance. Second, there was an increased level of sophistication in the analytic repertoire. Whereas multiple regression and ANOVA had been the norm in past decades, newer techniques including not only meta-analysis and SEM, but also multilevel and longitudinal modeling, began to be used frequently (Aguinis, Pierce, Bosco, & Muslin, 2009). Moreover, a movement from one to many meant that some of the methodological choices were not mutually exclusive and could be combined within the same study (e.g., collecting data using both field and laboratory designs, testing both moderated and mediated relations, combining meta-analysis and SEM). But, again, there was little guidance on how to go about implementing these integrative approaches.

The aforementioned challenges opened up new opportunities. First, *Psychological Methods* and *Organizational Research Methods* were launched as new journals specifically devoted to methodology in the mid- to late 1990s. Second, specifically regarding *JAP*, it published articles reviewing methodological practices and offering specific guidelines and best-practice recommendations. Most of these review articles addressed methodological debates that had lasted years and even decades. For example, in a *JAP* monograph, Hough, Eaton, Dunnette, Kamp, and McCloy (1990) reviewed the literature on intentional response distortion in preemployment testing—particularly in the domain of personality testing. Their review led to a set of recommendations on scale construction and appropriate use of scores to minimize the potential biasing impact of distortion on the resulting validity coefficients. As was mentioned previously, L. R. James et al. (1993) recast the derivation of r_{wg} within an interrater agreement framework, which addressed an ongoing debate on the specification of constructs measured at a lower level but analyzed and interpreted at a higher level of analysis. F. L. Schmidt and Hunter (1989) had criticized L. R. James et al. (1984) as a measure of interrater reliability. Kozlowski and Hattrup (1992) showed that if, as James et al. (1984) intended, their index was conceptualized as an index of consensus rather than consistency, then it functioned as James et al. (1984) claimed. James et al. (1993) clarified the intent of the 1984 paper as well as the difference between this intent and the F. L. Schmidt and Hunter (1989) critique. As a third illustration, Podsakoff, MacKenzie, Lee, and Podsakoff (2003) reviewed sources of common method bias and offered recommendations for selecting appropriate procedural and statistical remedies for different types of research settings. As yet another example, Aguinis, Beaty, Boik, and Pierce (2005) reviewed the literature on moderating effects and offered recommendations on how to maximize statistical power and minimize research design and measurement threats that may lead to the incorrect conclusion that such effects do not exist.

From a methodological perspective, a few themes emerged over the past quarter century of articles published in *JAP*. First, it became obvious that solutions based exclusively on data-analytic approaches would not suffice to address methodological challenges. For example, addressing such problems as intentional distortion in self-reports, lack of interrater agreement, low internal consistency, bias caused by common method variance, and insufficient statistical power to detect moderating effects require the implementation of solutions that combine theory, design, measurement, and data analysis (Aguinis & Vandenberg, 2014). Second, the publication of articles reviewing methodological issues was often reactive: It took several years of substantive researchers

mentioning a particular challenge until such a review was published. This is not surprising given that, as mentioned earlier, many editors have stated that methodological articles need to be placed within a particular substantive context. Third, the adoption of novel data analytic approaches tended to happen rather quickly—the wider availability of statistical software packages accelerated the speed of the adoption process. For example, meta-analysis and SEM, and, more recently, multilevel modeling, are data-analytic approaches that were adopted fairly quickly. However, innovations regarding research design were slow and often are not implemented at all.

In addition to the aforementioned methodological issues, we highlight a change in how methodological practices were reported. The trend toward longer Introduction sections seemed to shoot upward, placing greater length constraints on Method sections. As noted in Guion's (1988) parting editorial, "In 1961, most articles in JAP went directly to the method section after a brief introductory paragraph or two. . . . In 1986, most articles had a couple of pages of theoretical foundation" (p. 693). Since the early 1990s, Introduction sections have become much longer in relation to the method sections, reflecting an increased interest in the study's theoretical foundation. Murphy (2002) referred to this issue explicitly in his parting editorial as follows:

> The idea that theory is unimportant is absolutely wrong. In the last 11 years, I have written several thousand decision letters, and the issues that most frequently lead to the decision to reject a paper fall under the heading of "conceptual development." The most critical step in getting a paper published in JAP is to make a strong and direct link to relevant theory and research. (Murphy, 2002, p. 1019)

We agree entirely that this is the most critical step in getting a paper published in the journal. We wonder if this is desirable and if it comes at the cost of methodological rigor. It would be impossible to prove a (negative) causal relationship between theoretical and methodological emphasis, but given finite attentional resources, it does not seem unreasonable to suggest that such a relationship might exist. Strong theory is crucial to our field, but we should perhaps give more thought to the meaning of theory, especially if it might come at the expense of methods (Cortina, 2016).

A second issue regarding reporting practices concerns null hypothesis significance testing. Beginning in the 1990s, there was an increased awareness regarding the need to report not only test statistics and *p* values but also effect size estimates and their meaning in particular contexts (Aguinis, Werner, et al., 2010). The need to report effect sizes and discuss the practical importance of research results is related to the documented gap between science and practice (Cascio & Aguinis, 2008) and the concern that "the *Journal of Applied Psychology* may become almost exclusively a journal of articles written by academics for other academics" (Guion, 1988, p. 693).

Overall, the time period including 1990 through 2014 involved the introduction of many methodological innovations and a staggering broadening of the methodological landscape, to the point that the usual doctoral-level training regarding methodology may have fallen behind. In fact, a study by Aiken, West, and Millsap (2008) involving graduate training in statistics, research design, and measurement in 222 psychology departments concluded that "statistical and methodological curriculum has advanced little [since the 1960s]" (p. 721) and that "new developments in statistics, measurement, and methodology are not being incorporated into most graduate training programs" (p. 730). Accordingly, it is not surprising that the most recent *JAP* editors have scrambled to find reviewers sufficiently knowledgeable to evaluate manuscripts using more novel methods. Given the proliferation of methodological techniques, our field may be forced to revisit methods training and the infusion of methods expertise into the review process.

2017 and Beyond!

When we were graduate students in the late 1980s and early 1990s, we remember conversations with students in other areas of psychology remarking upon the amount of methods training that industrial-organizational students receive. The fact that we apply psychological science to the real world does not make us unique. School psychology, clinical psychology, developmental psychology, and other fields do this as well. The fact that we look for ways to improve workplace functioning does not set us apart either. Human Factors psychology and others do this as well. What sets us apart is our commitment to the appropriate use of first-rate research methods. From the very beginning of our field generally, and *JAP* particularly, controversies, debates, and discoveries were anchored in state-of-the-science methodology.

Hall was right about the obscurities of a century ago. They were the twilight of dawn. One of the main reasons that we were able to shed more light on those obscurities was that we were constantly searching for and finding better ways to collect and analyze information. We wanted better answers to bigger questions, and we searched for the methods that allowed us to get them.

In our opinion (and we wish to make clear that the editors of this issue should not be held accountable for these opinions), our field has different obscurities now, and many of these obscurities are self-inflicted. We very seldom publish constructive replications of the Eden and Shani (1982) sort because they do not make a theoretical contribution. We rarely study the exceptional or indeed anything that does not lend itself to quantitative methods. We seek to confirm rather than test our hypotheses because manuscripts with unsupported hypotheses do not survive the review process. We often apply analytic techniques without understanding them. In many respects, our field responded well to the calls for change issued by Darley and others 50 years ago. He called for new/interesting applications, and we now have optimization techniques for optimal test weighting, multilevel modeling, and IRT, among many others. We are also beginning to embrace Big Data in the study of job attitudes (Hernandez, Newman, & Jeon, 2016), assessment (Illingworth, Lippstreu, & Deprez-Sims, 2016), and turnover (Hausknecht & Li, 2016), among other phenomena.

In other ways, such as studying the exceptional and involving practitioners in research, Darley would probably be disappointed. In the present article, we have attempted to make note of the advances that have been made over the course of the history of *JAP* without glossing over the stumbles. We end our article with a discussion of things that we hope as-yet-unborn authors will write in a possible special issue of *JAP* to be published in the year 2067 to celebrate the journal's sesquicentennial anniversary.

Constructive Replication Stopped Being a Second-Class Activity

In the first issue of *JAP*, Bingham (1917) called for a cooperative system in which applied psychology problems were essentially assigned to the labs best suited to study them. The idea would be that a given lab would design studies, form hypotheses, test them, refine hypotheses, and retest until the lab had triangulated onto a solution. Although this and similar processes were reported in the early years of the journal, they do not happen anymore.

Kacmar and Whitfield (2000) and Colquitt and Zapata-Phelan (2007) found that the models offered in empirical papers are rarely tested again. J. R. Edwards, Berry, and Kay (2015) found that the models offered in *Academy of Management Review* articles are rarely tested at all. The reason is simple. In order to be published, an empirical paper must make a "theoretical contribution." In other words, it cannot test someone else's theory. If top journals are reluctant to publish constructive replications, then few researchers will conduct them. The solution here is simple. Our top journals must encourage and publish high-quality constructive replications.

It bears mentioning that constructive replication and repetition are not the same thing. We have heard many times that criticism of our field regarding lack of replication is unwarranted, because if it were, there would be no meta-analyses. We are not claiming that bivariate relationships from broad, observational variables are not reported by multiple authors. But this is not constructive replication. A good replication involves an improved, or at least different, attempt to test either an entire theory or a self-contained portion of a theory. Such papers do exist in *JAP* (e.g., Tan & Aryee, 2002), but they are rare. Within-paper replications (e.g., Hochwarter, Ferris, Zinko, Arnell, & James, 2007; O'Boyle & Aguinis, 2012) are becoming more common, but even these suffer from the bias toward significant results, among other things. Our field must embrace a model, common in other fields, that involves independent verification and constructive replication.

We Embraced Methods That Allowed Us to Study the Exceptionally Good and Bad

The exceptional is, by definition, rare. As we moved toward a Neyman-Pearson epistemological model, the need for larger samples made the study of the exceptional all but impossible. But surely there is value in knowing why some of the most influential and prolific scholars in our field continue to work full time, mostly pro bono, after retirement. Surely there is value in knowing why the employee with the perfect attendance record never missed a day. Those of us who have read the book *Moneyball* know that this case study of the Oakland A's was the only way to understand how they had one of the best records in baseball over several years with only one third the payroll of other teams. At the other end of the spectrum, a few months before the writing of our manuscript, National Transportation Safety Board officials were in Philadelphia combing through the May 2015 wreckage of AMTRAK train 188 in order to understand the exceptionally bad. Surely we could learn something about organizational functioning by studying people, units, and organizations that fail badly (see Sheldon, Dunning, & Ames, 2014, for a recent example). In order to do this, however, we may need to embrace qualitative methods such as case studies and grounded theory. If we do not, then we might consign ourselves to being a science of the mean (Aguinis & O'Boyle, 2014).

Big Data may also offer mechanisms for studying the exceptional. Rare cases out of 200 are too rare to study quantitatively, but rare cases out of 200,000 are not. Big Data contain many unique challenges, but it might also contain unique solutions to old problems.

Misguided or Incomplete Analyses Do Not Survive the Review Process

When the present authors were in graduate school, one could not conduct SEM analyses without knowing exactly what one was doing. The disadvantage of this was that very few people could conduct SEM analyses. The advantage was only those who knew what they were doing could engage in SEM.

Today, SEM analyses are semiautomatic with a variety of software choices. The result is that many such analyses are done incorrectly. For example, Cortina, Green, Keeler, and Vandenberg (in press) found that nearly 40% of papers in *JAP* and *Academy of Management Journal* (*AMJ*) reported incorrect degrees of freedom for their SEMs, which means the authors of these papers were not testing the models that they claimed to be testing. Similarly, models that integrate mediation and moderation are quite common, and authors have access to user friendly macros that allow for the testing of such models. Yet Holland, Shore, and Cortina (2016) found among other things that authors of papers in *JAP*, *AMJ*, and other top journals who hypothesize full mediation rarely defend full mediation and almost never test for it properly.

These are not merely the quibbles of stats geeks. These are problems that result in the wrong words going into Results and Discussion sections. The only solution to this problem is to ensure that every paper that survives the review process has been evaluated carefully by someone with expertise in the methods described in the paper. Given the strain on the reviewer pool, this will be very, very difficult to do. *JAP* received 928 new submissions in 2015, and as of April 18, 2016, it has received 344, which puts it on pace for about 1,000 for the year. This is double the number of submissions in the year 2000. Whether the solution lies in graduate training, continuing education, reviewer credentialing, reviewer compensation, or some combination, it will not be easy. Our hope is that the sesquicentennial authors will be able to write that we, as a field, tackled this problem in an effective manner.

We Shifted Emphasis Toward Research Design and Measurement

This may seem an odd wish, given the previous few paragraphs, but we hope that the fascination with abstruse data analysis techniques gets replaced by fascination with appropriate research design, including top-notch measurement. As data analysis software became easier to use, more researchers were able to perform advanced analyses. It seems to us that, as a field, our focus regarding methodological rigor shifted from design and measurement, which was and is hard, to analysis, which has become much easier. But whereas the data from a rigorous design can be analyzed any number of ways, including simple ones, there is no analysis that can fix data from a bad design.

One possible means to achieving this end would be a publishing model in which Introduction and Method sections alone are sub-

jected to a review process, sometimes called a Registered Reports model. Once these sections are approved, the author need only execute the design in order for the paper to be published. The data need not behave themselves vis-à-vis the Introduction, they need only be collected in a manner consistent with the proposed method. An alternative would be to have reviewers and editors review only the Introduction and Method sections of completed papers before seeing the results. Either approach would go a long way toward eliminating HARKing (hypothesizing after results are known), a problem that has become pervasive in our field (Bosco, Aguinis, Field, Pierce, & Dalton, 2016). Of course, ensuring that experts on a given design get a look at every viable submission to journals would help, but given the strain on the review system, this would be difficult.

Our hope is that future historians of our field will look back upon the next 50 years and observe that the review process rewarded researchers who made the difficult and time-consuming but appropriate design choices, even if that meant tolerating limitations of the study.

We Got More Specific With Our Theories

As M. Edwards (2010), Leavitt et al. (2010), and others have pointed out, our theories tend to be vague. They contain hypotheses that are, at best, directional. Over time, rather than refining them, we add boxes and arrows. A model expanded in this way "effectively closes it off from rebuttal or disconfirmation by anything in the world" (Healy, in press, p. 4). Other scientific fields move in the direction of parameterization of models. Prospect theory is a good example. But we do not do this, and we should, particularly given the availability of information on the current state of our knowledge, in the form of bivariate relations and their distributions, in the most popular domains in applied psychology and related fields (Bosco, Aguinis, Singh, Field, & Pierce, 2015).

One way to shift emphasis to research design and measurement would be to embrace computational modeling, which involves very detailed descriptions of processes complete with point estimates of parameters that can then be cross validated and adjusted. Another way to move in this direction would be to embrace categorical shift models of human behavior. Approaches such as catastrophe modeling and spline regression involve not only the identification of slope parameters but also the identification of points along an axis of predictor values at which a dependent variable value and/or its relationship to the predictor changes suddenly (Pierce & Aguinis, 2013). There are many organizational phenomena that are likely to be described by such models, but our field seems to shy away from them. Finally, a more Bayesian mindset might help us here (Kruschke, Aguinis, & Joo, 2012). If we were to evaluate study k + 1 not in isolation but as a mechanism for adjusting beliefs that had been driven only by study k, then theory refinement is more likely to move forward.

These are only a few examples of approaches that would help us to refine our theories. Whether through these or other mechanisms, our field would benefit from theoretical specificity.

Conclusion

The *Journal of Applied Psychology* began as an outlet for scientific psychologists who were interested in applying psychological principles to real-world problems. Ours was a problem- or phenomenon-driven field, but because the problems being studied had not been studied before (at least not by those trained in psychology), we knew little about how to solve them. Thus, because we needed rigorous and novel tools to study these issues, we devoted a great deal of attention to research methods, the emphasis and training of which set us apart from other social sciences.

Indeed, it is evident that developments in research methods had more impact than any other single category of paper published in *JAP* for at least the first half century of its existence. By far, the most widely cited *JAP* paper, from the inception of the journal through WWII, was Thorndike's (1920) paper on halo error. The most cited paper of the first half-century of the journal was Flesch's (1948) article on the measurement of readability. The second, third, and fourth most cited papers were also measurement papers. Only three of the top 10 most cited papers in this period were non-research-methods papers.

Various editorials have suggested a fear that we might become too focused on research methods rather than the questions at which those methods might be directed. There have been repeated reminders that research methods papers will only be considered at *JAP* if those papers examine methods that are specific to applied psychology questions. However, the halo error described by Thorndike in 1920 was not specific to applied psychology. Neither was Flesch's (1948) on readability. On the other hand, Brayfield and Rothe's (1951) measure of job satisfaction was specific to our field, as was Lodahl and Kejner's (1965) measure of job involvement. But measure development and validation articles have been rare exceptions for many years. So, it seems that many of the *JAP* papers that had the greatest impact would not survive the review process, or perhaps even the desk-reject process, now. Is this a reflection of the advancement of our field? Or does it suggest instead that the rules for publishing in *JAP* reflect a waning appreciation of the importance of research methods?

We do not know. And perhaps the answers to these questions do not matter. It all depends on what we as a field do next. Jose Cortina and Rick DeShon have served as Associate Editors of *JAP*, Herman Aguinis has served as consulting editor for many years, and we can offer first-hand corroboration of Kevin Murphy's observation that the most common deal breakers for *JAP* submissions revolve around the Introduction section. But this is not, in our opinion, because there is less to criticize in Method sections. This is instead because, overall, the field still tends to think of research methods as the details of the conscientious bookkeeper: They need to be in the report somewhere, but they should be relegated to the footnote, or other repositories of small-font finer points, so that they do not distract a reader from the heart of a paper. One can even imagine a drift toward placing methods information in online supplemental materials, rather than in the manuscript proper, as is currently the case regarding certain aspects of meta-analyses.

So, what should be done next? It may be that the authors of the sesquicentennial article will look not at 2017 as a turning point, but at 2007, when the scarlet letter of smaller font was removed from Method sections. Or, perhaps we as a field will respond en masse to the various calls for more rigorous methods. The present authors will not be around to see what is written in 2067. Our hope is that today will be marked as the beginning of a sea change for the

journal and the field, one in which we ensure that we do not sacrifice research methods at the altar of "selling a good story" and we return to the methodological preeminence that is the hallmark of an applied science (see Mathieu, in press, and Cortina, 2016, for recent discussion of this issue). The overarching conclusion that a reader draws from Darley's (1968) review of the first half century of *JAP* was that we, as a field, had not really learned from our mistakes and were therefore doomed to repeat them. It is our hope that the same will not be said 100 years thence, that just as the obscurities of 1917 marked the twilight of dawn, the obscurities of 2017 marked the twilight of a new day.

References

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90,* 94–107. http://dx.doi.org/10.1037/0021-9010.90.1.94

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95,* 648–680. http://dx.doi.org/10.1037/a0018714

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology, 108,* 1045–1059. http://dx.doi.org/10.1037/edu0000104

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16,* 270–301. http://dx.doi.org/10.1177/1094428112470848

Aguinis, H., & O'Boyle, E., Jr. (2014). Star performers in twenty-first-century organizations. *Personnel Psychology, 67,* 313–350. http://dx.doi.org/10.1111/peps.12054

Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., & Dalton, C. M. (2011). Debunking myths and urban legends about meta-analysis. *Organizational Research Methods, 14,* 306–331. http://dx.doi.org/10.1177/1094428110375720

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of Organizational Research Methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods, 12,* 69–112. http://dx.doi.org/10.1177/1094428108322641

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82,* 192–206. http://dx.doi.org/10.1037/0021-9010.82.1.192

Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior, 1,* 569–595. http://dx.doi.org/10.1146/annurev-orgpsych-031413-091231

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13,* 515–539. http://dx.doi.org/10.1177/1094428109333339

Aguinis, H., & Whitehead, R. (1997). Sampling variance in the correlation coefficient under indirect range restriction: Implications for validity generalization. *Journal of Applied Psychology, 82,* 528–538. http://dx.doi.org/10.1037/0021-9010.82.4.528

Ahearne, M., Bhattacharya, C. B., & Gruen, T. (2005). Antecedents and consequences of customer-company identification: Expanding the role of relationship marketing. *Journal of Applied Psychology, 90,* 574–585. http://dx.doi.org/10.1037/0021-9010.90.3.574

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63,* 32–50. http://dx.doi.org/10.1037/0003-066X.63.1.32

Anderson, J. E. (1935). The effect of item analysis upon the discriminative power of an examination. *Journal of Applied Psychology, 19,* 237–244. http://dx.doi.org/10.1037/h0057233

Arthur, G. (1925). A new point performance scale. *Journal of Applied Psychology, 9,* 390–416. http://dx.doi.org/10.1037/h0069996

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77,* 836–874. http://dx.doi.org/10.1037/0021-9010.77.6.836

Baer, M., & Oldham, G. R. (2006). The curvilinear relation between experienced creative time pressure and creativity: Moderating effects of openness to experience and support for creativity. *Journal of Applied Psychology, 91,* 963–970. http://dx.doi.org/10.1037/0021-9010.91.4.963

Balinsky, B., Blum, M. L., & Dutka, S. (1951). The coefficient of agreement in determining product preferences. *Journal of Applied Psychology, 35,* 348–351. http://dx.doi.org/10.1037/h0054500

Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology, 59,* 313–320. http://dx.doi.org/10.1037/h0036653

Bedaux, C. E. (1921). The Bedaux unit principle of industrial measurement. *Journal of Applied Psychology, 5,* 119–126. http://dx.doi.org/10.1037/h0075950

Bingham, W. V. (1917). Mentality testing of college students. *Journal of Applied Psychology, 1,* 38–45. http://dx.doi.org/10.1037/h0073261

Bitektine, A. (2007). Prospective case study design: Qualitative method for deductive theory testing. *Organizational Research Methods, 11,* 160–180. http://dx.doi.org/10.1177/1094428106292900

Bobko, P. (1995). Editorial. *Journal of Applied Psychology, 80,* 3–5. http://dx.doi.org/10.1037/h0092450

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69,* 709–750. http://dx.doi.org/10.1111/peps.12111

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100,* 431–449. http://dx.doi.org/10.1037/a0038047

Brayfield, A. H., & Rothe, H. F. (1951). An index of job satisfaction. *Journal of Applied Psychology, 35,* 307–311. http://dx.doi.org/10.1037/h0055617

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5,* 159–172. http://dx.doi.org/10.1177/1094428102005002002

Burtt, H. E. (1918). The perception of slight changes of equilibrium, with especial reference to problems of aviation. *Journal of Applied Psychology, 2,* 101–115. http://dx.doi.org/10.1037/h0069778

Burtt, H. E., & Arps, G. F. (1920). Correlation of Army Alpha Intelligence Test with academic grades in high schools and military academies. *Journal of Applied Psychology, 4,* 289–293. http://dx.doi.org/10.1037/h0070893

Burtt, H. E., & Dobell, E. M. (1925). The curve of forgetting for advertising material. *Journal of Applied Psychology, 9,* 5–21. http://dx.doi.org/10.1037/h0073966

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology, 67,* 691–700. http://dx.doi.org/10.1037/h0077946

Cascio, W. F., & Aguinis, H. (2008). Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology, 93,* 1062–1081. http://dx.doi.org/10.1037/0021-9010.93.5.1062

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83,* 234–246. http://dx.doi.org/10.1037/0021-9010.83.2.234

Chapman, C. J. (1919). The learning curve in type writing. *Journal of Applied Psychology, 3,* 252–268. http://dx.doi.org/10.1037/h0072933

Chapman, J. C. (1922). Individual injustice and guessing in the true-false examination. *Journal of Applied Psychology, 6,* 342–348. http://dx.doi.org/10.1037/h0076011

Chiaburu, D. S., & Harrison, D. A. (2008). Do peers make the place? Conceptual synthesis and meta-analysis of coworker effects on perceptions, attitudes, OCBs, and performance. *Journal of Applied Psychology, 93,* 1082–1103. http://dx.doi.org/10.1037/0021-9010.93.5.1082

Cobb, P. W. (1940). The limit of usefulness of accident rate as a measure of accident proneness. *Journal of Applied Psychology, 24,* 154–159. http://dx.doi.org/10.1037/h0055475

Colquitt, J. A., & Zapata-Phelan, C. P. (2007). Trends in theory building and theory testing: A five-decade study of the Academy of Management Journal. *Academy of Management Journal, 50,* 1281–1303. http://dx.doi.org/10.5465/AMJ.2007.28165855

Cook, W. W., & Medley, D. M. (1954). Proposed hostility and pharisaic-virtue scales for the MMPI. *Journal of Applied Psychology, 38,* 414–418. http://dx.doi.org/10.1037/h0060667

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104. http://dx.doi.org/10.1037/0021-9010.78.1.98

Cortina, J. M. (2016). Defining and operationalizing theory in the organizational sciences. *Journal of Organizational Behavior, 37,* 1142–1149. http://dx.doi.org/10.1002/job.2121

Cortina, J. M., & DeShon, R. P. (1998). Determining relative importance of predictors with the observational design. *Journal of Applied Psychology, 83,* 798–804. http://dx.doi.org/10.1037/0021-9010.83.5.798

Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (in press). Degrees of freedom in SEM: Are we testing the models that we claim to test? *Organizational Research Methods*.

Coy, G. L. (1918). The mentality of a gifted child. *Journal of Applied Psychology, 2,* 299–307. http://dx.doi.org/10.1037/h0075175

Cureton, E. E., & Dunlap, J. W. (1930). Note on the testing of departure from normality. *Journal of Applied Psychology, 14,* 91–94. http://dx.doi.org/10.1037/h0072832

Darley, J. G. (1968). 1917: A journal is born. *Journal of Applied Psychology, 52,* 1–9. http://dx.doi.org/10.1037/h0025256

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92,* 1380–1393. http://dx.doi.org/10.1037/0021-9010.92.5.1380

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662–680. http://dx.doi.org/10.1037/0021-9010.70.4.662

Eden, D., & Shani, A. B. (1982). Pygmalion goes to boot camp: Expectancy, leadership, and trainee performance. *Journal of Applied Psychology, 67,* 194–199.

Edgerton, H. A. (1930). A table for finding the probable error of R obtained by use of the Spearman-Brown formula (n=2). *Journal of Applied Psychology, 14,* 296–302. http://dx.doi.org/10.1037/h0075759

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology, 37,* 90–93. http://dx.doi.org/10.1037/h0058073

Edwards, J. R., Berry, J. W., & Kay, V. S. (2015). *Bridging the great divide between theoretical and empirical management research. Working paper, Kenan-Flagler Business School*. Chapel Hill, NC: University of North Carolina.

Edwards, M. (2010). *Organizational transformation for sustainability: An integral metatheory*. New York, NY: Routledge.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73,* 421–435. http://dx.doi.org/10.1037/0021-9010.73.3.421

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd.

Flanders, J. K. (1918). Mental tests of a group of employed men showing correlations with estimates furnished by employer. *Journal of Applied Psychology, 2,* 197–206. http://dx.doi.org/10.1037/h0072524

Fleishman, E. A. (1971). Editorial. *Journal of Applied Psychology, 55,* 1–2. http://dx.doi.org/10.1037/h0020017

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32,* 221–233. http://dx.doi.org/10.1037/h0057532

Freyd, M. (1922). A method for the study of vocational interests. *Journal of Applied Psychology, 6,* 243–254. http://dx.doi.org/10.1037/h0072563

Freyd, M. (1925). The statistical viewpoint in vocational selection. *Journal of Applied Psychology, 9,* 349–356. http://dx.doi.org/10.1037/h0074663

Freyer, D. (1930). The objective and subjective measurement of interests—An acceptance-rejection theory. *Journal of Applied Psychology, 14,* 549–556. http://dx.doi.org/10.1037/h0073774

Garrison, C. G., Burke, A., & Hollingworth, L. S. (1917). The psychology of a prodigious child. *Journal of Applied Psychology, 1,* 101–110. http://dx.doi.org/10.1037/h0070864

Gates, A. I. (1918). The abilities of an expert marksman tested in the psychological laboratories. *Journal of Applied Psychology, 2,* 1–14. http://dx.doi.org/10.1037/h0074646

Gates, A. I. (1923). The unreliability of M.A. and I.Q. based on group tests of general ability. *Journal of Applied Psychology, 7,* 93–100. http://dx.doi.org/10.1037/h0071792

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511. http://dx.doi.org/10.1037/0021-9010.72.3.493

Georgopoulos, B. S., Mahoney, G. M., & Jones, N. W., Jr. (1957). A path-goal approach to productivity. *Journal of Applied Psychology, 41,* 345–353. http://dx.doi.org/10.1037/h0048473

Gerstner, C. R., & Day, D. V. (1997). Meta-analytic review of leader–member exchange theory: Correlates and construct issues. *Journal of Applied Psychology, 82,* 827–844. http://dx.doi.org/10.1037/0021-9010.82.6.827

Ghiselli, E. E., & Brown, C. W. (1948). The effectiveness of intelligence tests in the selection of workers. *Journal of Applied Psychology, 32,* 575–580. http://dx.doi.org/10.1037/h0060336

Gordon, M. E., Philpot, J. W., Burt, R. E., Thompson, C. A., & Spiller, W. E. (1980). Commitment to the union: Development of a measure and an examination of its correlates. *Journal of Applied Psychology, 65,* 479–499. http://dx.doi.org/10.1037/0021-9010.65.4.479

Guion, R. M. (1965). *Personnel testing*. New York, NY: McGraw-Hill.

Guion, R. M. (1988). Editorial: Special section: From psychologists in organizations. *Journal of Applied Psychology, 73,* 693–694. http://dx.doi.org/10.1037/h0092447

Hackman, J. R., & Oldham, G. R. (1975). Development of the job diagnostic survey. *Journal of Applied Psychology, 60,* 159–170. http://dx.doi.org/10.1037/h0076546

Hall, G. S. (1917). Practical relations between psychology and the war. *Journal of Applied Psychology, 1,* 9–16. http://dx.doi.org/10.1037/h0070238

Hall, G., Baird, J. W. E., & Geissler, L. R. (1918). Communications regarding "A plan for the technical training of consulting psychologists." *Journal of Applied Psychology, 2,* 174–178. http://dx.doi.org/10.1037/h0071739

Hausknecht, J. P., & Li, H. (2016). Big data in turnover and retention. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 250–271). New York, NY: Routledge.

Healy, K. (in press). Fuck nuance. *Sociological Theory*.

Heneman, H. G. (1974). Comparisons of self-and superior ratings of managerial performance. *Journal of Applied Psychology, 59,* 638–642. http://dx.doi.org/10.1037/h0037341

Henmon, V. A. C. (1919). Air service tests of aptitude for flying. *Journal of Applied Psychology, 3,* 103–109. http://dx.doi.org/10.1037/h0070342

Hernandez, I., Newman, D. A., & Jeon, G. (2016). Twitter analysis: Methods for data management and a word count dictionary to measure city-level job satisfaction. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 64–114). New York, NY: Routledge.

Hinkin, T. R., & Schriesheim, C. A. (1989). Development and application of new scales to measure the French and Raven (1959) bases of social power. *Journal of Applied Psychology, 74,* 561–567. http://dx.doi.org/10.1037/0021-9010.74.4.561

Hochwarter, W. A., Ferris, G. R., Zinko, R., Arnell, B., & James, M. (2007). Reputation as a moderator of political behavior-work outcomes relationships: A two-study investigation with convergent results. *Journal of Applied Psychology, 92,* 567–576. http://dx.doi.org/10.1037/0021-9010.92.2.567

Hofmann, D. A., Jacobs, R., & Baratta, J. E. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology, 78,* 194–204. http://dx.doi.org/10.1037/0021-9010.78.2.194

Holland, S. J., Shore, D. B., & Cortina, J. M. (2016). Review and recommendations for integrated mediation and moderation. *Organizational Research Methods*. Advance online publication. http://dx.doi.org/10.1177/1094428116658958

Hothersall, D. (1990). *History of psychology* (2nd ed.). New York, NY: McGraw-Hill.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595. http://dx.doi.org/10.1037/0021-9010.75.5.581

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67,* 818–825. http://dx.doi.org/10.1037/0021-9010.67.6.818

Hull, C. L. (1922a). The computation of Pearson's r from ranked data. *Journal of Applied Psychology, 6,* 385–390. http://dx.doi.org/10.1037/h0071512

Hull, C. L. (1922b). The conversion of test scores into series which shall have any assigned mean and degree of dispersion. *Journal of Applied Psychology, 6,* 298–300. http://dx.doi.org/10.1037/h0071530

Hull, C. L. (1923). Prediction formulae for teams of aptitude tests. *Journal of Applied Psychology, 7,* 277–284. http://dx.doi.org/10.1037/h0073241

Hunt, T. (1928). The measurement of social intelligence. *Journal of Applied Psychology, 12,* 317–334. http://dx.doi.org/10.1037/h0075832

Illingworth, A. J., Lippstreu, M., & Deprez-Sims, A. (2016). Big data in talent selection and assessment. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 213–249). New York, NY: Routledge.

Ironson, G. H., Smith, P. C., Brannick, M. T., Gibson, W. M., & Paul, K. B. (1989). Construction of a job in general scale: A comparison of global, composite, and specific measures. *Journal of Applied Psychology, 74,* 193–200. http://dx.doi.org/10.1037/0021-9010.74.2.193

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219–229. http://dx.doi.org/10.1037/0021-9010.67.2.219

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1,* 131–163. http://dx.doi.org/10.1177/109442819812001

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 85–98. http://dx.doi.org/10.1037/0021-9010.69.1.85

James, L. R., Demaree, R. G., & Wolf, G. (1993). An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78,* 306–309. http://dx.doi.org/10.1037/0021-9010.78.2.306

James, L. A., & James, L. R. (1989). Integrating work environment perceptions: Explorations into the measurement of meaning. *Journal of Applied Psychology, 74,* 739–751. http://dx.doi.org/10.1037/0021-9010.74.5.739

Jordan, A. M. (1930). Mental growth. *Journal of Applied Psychology, 14,* 517–531. http://dx.doi.org/10.1037/h0074829

Kacmar, K. M., & Whitfield, J. M. (2000). An additional rating method for journal articles in the field of management. *Organizational Research Methods, 3,* 392–406. http://dx.doi.org/10.1177/109442810034005

Kanungo, R. N. (1982). Measurement of job and work involvement. *Journal of Applied Psychology, 67,* 341–349. http://dx.doi.org/10.1037/0021-9010.67.3.341

Kearney, E., & Gebert, D. (2009). Managing diversity and enhancing team outcomes: The promise of transformational leadership. *Journal of Applied Psychology, 94,* 77–89. http://dx.doi.org/10.1037/a0013077

Kelley, T. L. (1919). Principle underlying the classification of men. *Journal of Applied Psychology, 1,* 50–67. http://dx.doi.org/10.1037/h0075815

Kohs, S. C., & Irle, K. W. (1920). Prophesying army promotion. *Journal of Applied Psychology, 4,* 73–87. http://dx.doi.org/10.1037/h0070002

Kornhauser, A. W. (1927). Results from a quantitative questionnaire on likes and dislikes used with a group of college freshmen. *Journal of Applied Psychology, 11,* 85–94. http://dx.doi.org/10.1037/h0073120

Kozlowski, S. W. J. (2009). Editorial. *Journal of Applied Psychology, 94,* 1–4. http://dx.doi.org/10.1037/a0014990

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77,* 161–167. http://dx.doi.org/10.1037/0021-9010.77.2.161

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology, 70,* 56–65. http://dx.doi.org/10.1037/0021-9010.70.1.56

Krathwohl, W. C. (1952). Specificity of over-and under-achievement in college courses. *Journal of Applied Psychology, 36,* 103–106. http://dx.doi.org/10.1037/h0061633

Kristof-Brown, A. L., Jansen, K. J., & Colbert, A. E. (2002). A policy-capturing study of the simultaneous effects of fit with jobs, groups, and organizations. *Journal of Applied Psychology, 87,* 985–993. http://dx.doi.org/10.1037/0021-9010.87.5.985

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15,* 722–752. http://dx.doi.org/10.1177/1094428112457829

Kuhlmann, F. (1921). The results of repeated mental reexamination of 639 feeble-minded over a period of ten years. *Journal of Applied Psychology, 5,* 195–224. http://dx.doi.org/10.1037/h0071579

Leavitt, K., Mitchell, T. R., & Peterson, J. (2010). Theory pruning: Strategies to reduce our dense theoretical landscape. *Organizational Research Methods, 13,* 644–667. http://dx.doi.org/10.1177/1094428109345156

Lee, R. T., & Ashforth, B. E. (1996). A meta-analytic examination of the correlates of the three dimensions of job burnout. *Journal of Applied Psychology, 81,* 123–133. http://dx.doi.org/10.1037/0021-9010.81.2.123

Liden, R. C., Wayne, S. J., & Stilwell, D. (1993). A longitudinal study on the early development of leader-member exchanges. *Journal of Applied Psychology, 78,* 662–674.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of CVI, T, rWG(J)}, and r*WG(J)} indexes. *Journal of Applied Psychology, 84,* 640–647. http://dx.doi.org/10.1037/0021-9010.84.4.640

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60,* 10–13. http://dx.doi.org/10.1037/h0076268

Locke, N. M. (1940). The student skills inventory: A study habits test. *Journal of Applied Psychology, 24,* 493–504. http://dx.doi.org/10.1037/h0058668

Lodahl, T. M., & Kejner, M. (1965). The definition and measurement of job involvement. *Journal of Applied Psychology, 49,* 24–33. http://dx.doi.org/10.1037/h0021692

Longstaff, H. P. (1948). Fakability of the Strong interest blank and the Kuder preference record. *Journal of Applied Psychology, 32,* 360–369. http://dx.doi.org/10.1037/h0055301

Manson, G. E. (1925). Personality differences in intelligence test performance. *Journal of Applied Psychology, 9,* 230–255. http://dx.doi.org/10.1037/h0069913

Mateer, F. (1917). The moron as a war problem. *Journal of Applied Psychology, 1,* 317–320. http://dx.doi.org/10.1037/h0073157

Mathieu, J. E. (in press). The problem with [in] management theory. *Journal of Organizational Behavior*.

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97,* 951–966. http://dx.doi.org/10.1037/a0028380

Maurer, T. J., & Pierce, H. R. (1998). A comparison of Likert scale and traditional measures of self-efficacy. *Journal of Applied Psychology, 83,* 324–329. http://dx.doi.org/10.1037/0021-9010.83.2.324

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the position analysis questionnaire (PAQ). *Journal of Applied Psychology, 56,* 347–368.

McEvoy, G. M., & Cascio, W. F. (1989). Cumulative evidence of the relationship between employee age and job performance. *Journal of Applied Psychology, 74,* 11–17. http://dx.doi.org/10.1037/0021-9010.74.1.11

McKinley, J. C., & Hathaway, S. R. (1942). A Multiphasic Personality Schedule (Minnesota): IV. Psychasthenia. *Journal of Applied Psychology, 26,* 614–624. http://dx.doi.org/10.1037/h0063530

Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology, 30,* 525–564. http://dx.doi.org/10.1037/h0053634

Moore, H. T., & Gilliland, A. R. (1921). The measurement of aggressiveness. *Journal of Applied Psychology, 5,* 97–118. http://dx.doi.org/10.1037/h0073691

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82,* 3–5. http://dx.doi.org/10.1037/h0092448

Murphy, K. R. (2002). Editorial. *Journal of Applied Psychology, 87,* 1019. http://dx.doi.org/10.1037/0021-9010.87.6.1019

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical and Physical Sciences, 231,* 289–337. http://dx.doi.org/10.1098/rsta.1933.0009

O'Boyle, E., Jr., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65,* 79–119. http://dx.doi.org/10.1111/j.1744-6570.2011.01239.x

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679–703. http://dx.doi.org/10.1037/0021-9010.78.4.679

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187–207. http://dx.doi.org/10.1037/0021-9010.89.2.187

Parsons, C. K., & Hulin, C. L. (1982). An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. *Journal of Applied Psychology, 67,* 826–834. http://dx.doi.org/10.1037/0021-9010.67.6.826

Paterson, D. G., & Jenkins, J. J. (1948). Communication between management and workers. *Journal of Applied Psychology, 32,* 71–80. http://dx.doi.org/10.1037/h0054451

Paterson, D. G., & Langlie, T. A. (1925). Empirical data on the scoring of true-false tests. *Journal of Applied Psychology, 9,* 339–348. http://dx.doi.org/10.1037/h0069813

Piekkari, R., Welch, C., & Paavilainen, E. (2009). The case study as disciplinary convention: Evidence from international business journals. *Organizational Research Methods, 12,* 567–589. http://dx.doi.org/10.1177/1094428108319905

Pierce, J. R., & Aguinis, H. (2013). The too-much-of-a-good-thing effect in management. *Journal of Management, 39,* 313–338. http://dx.doi.org/10.1177/0149206311410060

Pintner, R. (1919). A non-language group intelligence test. *Journal of Applied Psychology, 3,* 199–214. http://dx.doi.org/10.1037/h0072783

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903. http://dx.doi.org/10.1037/0021-9010.88.5.879

Premack, S. L., & Wanous, J. P. (1985). A meta-analysis of realistic job preview experiments. *Journal of Applied Psychology, 70,* 706–719. http://dx.doi.org/10.1037/0021-9010.70.4.706

Pressey, S. L. (1921). The problem of the unselected group in the standardization of tests. *Journal of Applied Psychology, 5,* 64–71. http://dx.doi.org/10.1037/h0075051

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529. http://dx.doi.org/10.1037/0021-9010.87.3.517

Richardson, H. M. (1948). Adult leadership scales based on the Bernreuter personality inventory. *Journal of Applied Psychology, 32,* 292–303. http://dx.doi.org/10.1037/h0056428

Roberts, K. H., & O'Reilly, C. A. (1974). Measuring organizational communication. *Journal of Applied Psychology, 59,* 321–326. http://dx.doi.org/10.1037/h0036660

Root, W. T. (1921). Two cases showing marked change in I.Q. *Journal of Applied Psychology, 5,* 156–158. http://dx.doi.org/10.1037/h0071066

Roznowski, M. (1989). Examination of the measurement properties of the Job Descriptive Index with experimental items. *Journal of Applied Psychology, 74,* 805–814. http://dx.doi.org/10.1037/0021-9010.74.5.805

Ruch, G. M., & Del Manzo, M. C. (1923). The Downey Will-Temperament Group Test: A further analysis of its reliability and validity. *Journal of Applied Psychology, 7,* 65–76. http://dx.doi.org/10.1037/h0074794

Ruckmick, C. A. (1930). The uses and abuses of the questionnaire procedure. *Journal of Applied Psychology, 14,* 32–41. http://dx.doi.org/10.1037/h0074349

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401–410. http://dx.doi.org/10.1037/0021-9010.67.4.401

Sarbin, T. R., & Berdie, R. F. (1940). Relation of measured interests to the Allport-Vernon study of values. *Journal of Applied Psychology, 24,* 287–296. http://dx.doi.org/10.1037/h0061981

Schiller, G. (1935). An experimental study of the appropriateness of color and type in advertising. *Journal of Applied Psychology, 19,* 652–664. http://dx.doi.org/10.1037/h0056090

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540. http://dx.doi.org/10.1037/0021-9010.62.5.529

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology, 74,* 368–370. http://dx.doi.org/10.1037/0021-9010.74.2.368

Schmidt, H. O. (1945). Test profiles as a diagnostic aid: The Minnesota Multiphasic Inventory. *Journal of Applied Psychology, 29,* 115–131. http://dx.doi.org/10.1037/h0060192

Schmitt, N. (1989). Editorial. *Journal of Applied Psychology, 74,* 843–845. http://dx.doi.org/10.1037/h0092216

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., & Yoo, T.-Y. (2003). Impact of elaboration on socially desirable responding and the validity of biodata measures. *Journal of Applied Psychology, 88,* 979–988. http://dx.doi.org/10.1037/0021-9010.88.6.979

Sheldon, O. J., Dunning, D., & Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: Reactions to feedback about deficits in emotional intelligence. *Journal of Applied Psychology, 99,* 125–137. http://dx.doi.org/10.1037/a0034138

Siegel, S. M., & Kaemmerer, W. F. (1978). Measuring the perceived support for innovation in organizations. *Journal of Applied Psychology, 63,* 553–562. http://dx.doi.org/10.1037/0021-9010.63.5.553

Simons, T. L., & Peterson, R. S. (2000). Task conflict and relationship conflict in top management teams: The pivotal role of intragroup trust. *Journal of Applied Psychology, 85,* 102–111. http://dx.doi.org/10.1037/0021-9010.85.1.102

Slawson, J. (1922). The reliability of judgment of personal traits. *Journal of Applied Psychology, 6,* 161–171. http://dx.doi.org/10.1037/h0072975

Stevenson, J. A. (1918). Correlation between different forms of sensory discrimination. *Journal of Applied Psychology, 2,* 26–42. http://dx.doi.org/10.1037/h0072324

Strong, E. K. (1918). Work of the committee on classification of personnel in the Army. *Journal of Applied Psychology, 2,* 130–139. http://dx.doi.org/10.1037/h0074881

Strong, E. K. (1925). Theories of selling. *Journal of Applied Psychology, 9,* 75–86. http://dx.doi.org/10.1037/h0070123

Strong, E. K. (1927). Vocational guidance of executives. *Journal of Applied Psychology, 11,* 331–347. http://dx.doi.org/10.1037/h0075674

Strong, E. K. (1952). Nineteen-year followup of engineer interests. *Journal of Applied Psychology, 36,* 65–74. http://dx.doi.org/10.1037/h0056227

Strong, E. K., & Green, H. J. (1932). Short-cuts to scoring an interest test. *Journal of Applied Psychology, 16,* 1–8. http://dx.doi.org/10.1037/h0071956

Strong, E. K., Jr. (1926). Interest analysis of personnel managers. *Journal of Personnel Research, 5,* 235–242.

Sturges, H. A. (1924). Notes on the theory of sampling and applications in estimates of reliability and causal independence in statistical series. *Journal of Applied Psychology, 8,* 354–356. http://dx.doi.org/10.1037/h0069849

Symonds, P. M. (1925). Notes on rating. *Journal of Applied Psychology, 9,* 188–195. http://dx.doi.org/10.1037/h0069901

Takeuchi, R., Lepak, D. P., Wang, H., & Takeuchi, K. (2007). An empirical examination of the mechanisms mediating between high-performance work systems and the performance of Japanese organizations. *Journal of Applied Psychology, 92,* 1069–1083. http://dx.doi.org/10.1037/0021-9010.92.4.1069

Tan, H. H., & Aryee, S. (2002). Antecedents and outcomes of union loyalty: A constructive replication and an extension. *Journal of Applied Psychology, 87,* 715–722. http://dx.doi.org/10.1037/0021-9010.87.4.715

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578. http://dx.doi.org/10.1037/h0057079

Tendler, A. D. (1930). A preliminary report on a test for emotional insight. *Journal of Applied Psychology, 14,* 122–136. http://dx.doi.org/10.1037/h0076066

Terman, L. M. (1918). A experiment in infant education. *Journal of Applied Psychology, 2,* 219–228. http://dx.doi.org/10.1037/h0071908

Terman, L. M., & Chamberlain, M. B. (1918). Twenty three serial tests of intelligence and their intercorrelations. *Journal of Applied Psychology, 2,* 341–354. http://dx.doi.org/10.1037/h0072077

Terman, L. M., Otis, A. S., Dickson, V., Hubbard, O. S., Norton, J. K., Howard, L., . . . Cassingham, C. C. (1917). A trial of mental and pedagogical tests in a civil service examination for policemen and firemen. *Journal of Applied Psychology, 1,* 17–29. http://dx.doi.org/10.1037/h0073841

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88,* 500–517. http://dx.doi.org/10.1037/0021-9010.88.3.500

Thorndike, E. L. (1917). The curve of work and the curve of satisfyingness. *Journal of Applied Psychology, 1,* 265–267. http://dx.doi.org/10.1037/h0074929

Thorndike, E. L. (1918). Fundamental theorems in judging men. *Journal of Applied Psychology, 2,* 67–76. http://dx.doi.org/10.1037/h0074876

Thorndike, E. L. (1919). A standardized group examination of intelligence independent of language. *Journal of Applied Psychology, 3,* 13–32. http://dx.doi.org/10.1037/h0070037

Thorndike, E. L. (1920). Equality in difficulty of alternative intelligence examinations. *Journal of Applied Psychology, 4,* 283–288. http://dx.doi.org/10.1037/h0070952

Thorndike, E. L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology, 6,* 29–33. http://dx.doi.org/10.1037/h0072880

Thorndike, E. L. (1925). On the provision of alternative forms of examinations equal in difficulty. *Journal of Applied Psychology, 9,* 1–4. http://dx.doi.org/10.1037/h0072503

Thurstone, L. L. (1919). Mental tests for prospective telegraphers: A study of the diagnostic value of mental tests for predicting ability to learn telegraphy. *Journal of Applied Psychology, 3,* 110–117. http://dx.doi.org/10.1037/h0071741

Totterdell, P. (2000). Catching moods and hitting runs: Mood linkage and subjective performance in professional sport teams. *Journal of Applied Psychology, 85,* 848–859. http://dx.doi.org/10.1037/0021-9010.85.6.848

Tubbs, M. E. (1986). Goal setting: A meta-analytic examination of the empirical evidence. *Journal of Applied Psychology, 71,* 474–483. http://dx.doi.org/10.1037/0021-9010.71.3.474

Vancouver, J. B., Millsap, R. E., & Peters, P. A. (1994). Multilevel analysis of organizational goal congruence. *Journal of Applied Psychology, 79,* 666–679. http://dx.doi.org/10.1037/0021-9010.79.5.666

Vernon, P. E. (1930). A method for measuring musical taste. *Journal of Applied Psychology, 14,* 355–362. http://dx.doi.org/10.1037/h0071071

Viteles, M. S. (1925). The clinical viewpoint in vocational selection. *Journal of Applied Psychology, 9,* 131–138. http://dx.doi.org/10.1037/h0071305

Waldman, D. A., & Avolio, B. J. (1986). A meta-analysis of age differences in job performance. *Journal of Applied Psychology, 71,* 33–38. http://dx.doi.org/10.1037/0021-9010.71.1.33

Wanous, J. P., & Lawler, E. E. (1972). Measurement and meaning of job satisfaction. *Journal of Applied Psychology, 56,* 95–105. http://dx.doi.org/10.1037/h0032664

Watts, F. (1921). The construction of tests for the discovery of vocational fitness. *Journal of Applied Psychology, 5,* 240–252. http://dx.doi.org/10.1037/h0071871

Wells, F. L. (1917). Alternative methods for mental examiners. *Journal of Applied Psychology, 1,* 134–143. http://dx.doi.org/10.1037/h0075148

Wembridge, E. R. (1918). Obscurities in voting upon measures due to double-negative. *Journal of Applied Psychology, 2,* 156–163. http://dx.doi.org/10.1037/h0075886

Wesman, A. G. (1952). Faking personality test scores in a simulated employment situation. *Journal of Applied Psychology, 36,* 112–113. http://dx.doi.org/10.1037/h0055134

West, M. A., & Anderson, N. R. (1996). Innovation in top management teams. *Journal of Applied Psychology, 81,* 680–693. http://dx.doi.org/10.1037/0021-9010.81.6.680

Williams, K. J., Suls, J., Alliger, G. M., Learner, S. M., & Wan, C. K. (1991). Multiple role juggling and daily mood states in working mothers: An experience sampling study. *Journal of Applied Psychology, 76,* 664–674. http://dx.doi.org/10.1037/0021-9010.76.5.664

Williams, L. J., & Hazer, J. T. (1986). Antecedents and consequences of satisfaction and commitment in turnover models: A reanalysis using latent variable structural equation methods. *Journal of Applied Psychology, 71,* 219–231. http://dx.doi.org/10.1037/0021-9010.71.2.219

Wollack, S., Goodale, J. G., Wijting, J. P., & Smith, P. C. (1971). Development of the survey of work values. *Journal of Applied Psychology, 55,* 331–338. http://dx.doi.org/10.1037/h0031531

Xhignesse, L. V., & Osgood, C. E. (1967). Bibliographical citation characteristics of the psychological journal network in 1950 and 1960. *American Psychologist, 22,* 778–791. http://dx.doi.org/10.1037/h0024961

Yerkes, R. M. (1917). The Binet versus the point scale method of measuring intelligence. *Journal of Applied Psychology, 1,* 111–122. http://dx.doi.org/10.1037/h0070364

Zedeck, S. (2003). Editorial. *Journal of Applied Psychology, 88,* 3–5. http://dx.doi.org/10.1037/0021-9010.88.1.3

Ziller, R. C., Behringer, R. D., & Goodchilds, J. D. (1962). Group creativity under conditions of success or failure and variations in group stability. *Journal of Applied Psychology, 46,* 43–49. http://dx.doi.org/10.1037/h0045647

Zohar, D. (2000). A group-level model of safety climate: Testing the effect of group climate on microaccidents in manufacturing jobs. *Journal of Applied Psychology, 85,* 587–596. http://dx.doi.org/10.1037/0021-9010.85.4.587

Zohar, D. (2002). Modifying supervisory practices to improve subunit safety: A leadership-based intervention model. *Journal of Applied Psychology, 87,* 156–163. http://dx.doi.org/10.1037/0021-9010.87.1.156

Received June 1, 2015
Revision received August 12, 2016
Accepted August 15, 2016 ■