PERSONNEL PSYCHOLOGY

THE STUDY OF PEOPLE AT WORK

PERSONNEL PSYCHOLOGY
2014, 67, 531–566

ON THE DISTRIBUTION OF JOB PERFORMANCE: THE ROLE OF MEASUREMENT CHARACTERISTICS IN OBSERVED DEPARTURES FROM NORMALITY

JAMES W. BECK
University of Waterloo

ADAM S. BEATTY
University of Minnesota

PAUL R. SACKETT
University of Minnesota

In a recent article, O'Boyle and Aguinis (2012) argued that job performance is not distributed normally but instead is nonnormal and highly skewed. However, we believe the extreme departures from normality observed by these authors may have been due to characteristics of performance measures used. To address this issue, we identify 7 measurement criteria that we argue must be present for inferences to be made about the distribution of job performance. Specifically, performance measures must: (a) reflect behavior, (b) include an aggregation of multiple behaviors, (c) include the full range of performers, (d) include the full range of performance, (e) be time bounded, (f) focus on comparable jobs, and (g) not be distorted by motivational forces. Next, we present data from a wide range of sources—including the workplace, laboratory, athletics, and computer simulations—that illustrate settings in which failing to meet one or more of these criteria led to a highly skewed distribution providing a better fit to the data than a normal distribution. However, measurement approaches that better align with the 7 criteria listed above resulted in a normal distribution providing a better fit. We conclude that large departures from normality are in many cases an artifact of measurement.

Organization-level performance is improved via job performance at the individual level (e.g., Combs, Liu, Hall, & Ketchen, 2006; Podsakoff, Whiting, Podsakoff, & Blume, 2009). Thus, the field of personnel psychology provides value to organizations through the ability to make sound

We would like to thank Steffanie Wilk for providing the call center data, Maria Rotundo for providing the NBA data, and Nathan Bowling and Gary Burns for providing the I-O psychologist publication data. We would also like to thank Fred Oswald for his comments on an earlier draft of the manuscript.

Adam Beatty is now at the Human Resources Research Organization (HumRRO).

Correspondence and requests for reprints should be addressed to James W. Beck, University of Waterloo, Psychology Department, 200 University Ave. W., Waterloo, ON N2L 3G1, Canada; James.Beck@uwaterloo.ca.

 doi: 10.1111/peps.12060

inferences about the job performance of their members. For instance, organizations use personnel selection to choose employees who will be most likely to perform well and benefit the organization based on information available at the time of hire (e.g., knowledge, skills, and abilities [KSAs] observed in an interview). Likewise, organizational interventions (e.g., training) are evaluated based on their ability to improve the job performance of individual workers. Finally, during the annual performance appraisal, organizations assess employees' job performance to determine the value they have brought to the organization over the past several months. Thus, organizational success is contingent on the accuracy of inferences that are made about employee performance.

Because many inferences made about job performance are derived from the results of statistical procedures, the degree to which the distribution of job performance departs from normality can have important implications for the accuracy of the inferences. Specifically, underlying many common statistical procedures is the assumption that the variables involved are distributed normally (Tabachnick & Fidell, 2007). However, in a recent paper, O'Boyle and Aguinis (2012) argued that job performance is generally not normally distributed but instead follows a nonnormal and highly skewed distribution. This distribution is characterized by the highest number of performers falling at the low end rather than in the center, and by a consistently decreasing number of performers at each subsequent higher level of performance. Furthermore, O'Boyle and Aguinis argued that *ratings* of job performance are often "forced" into normality (e.g., via rating instructions), yet *objective* measures of performance often fail to meet the normality assumption. O'Boyle and Aguinis presented a large amount of evidence from a variety of domains showing just this: highly skewed distributions of objective performance indicators.

If the distributions of job performance reported in O'Boyle and Aguinis' (2012) article are representative of the job performance construct, then their data could be seen as evidence for a need for wholesale retooling of personnel psychology theory, research methods, and data analysis techniques. That is, if the distribution of job performance was typically characterized by most employees clustering around the low end of the distribution, and very few employees achieving very high "superstar" levels of performance, then inferences drawn from procedures assuming a normal distribution could be uninformative or even misleading. Fortunately, there is some indication that this is likely not the case. Specifically, the data reported by O'Boyle and Aguinis were in some ways idiosyncratic and not particularly representative of job performance as it is typically defined. They reported data on rates of publishing in academic journals, on sports performance (e.g., number of golf tournaments won), on literary and artistic performance (e.g., number of Academy

Award nominations received), and on political election results (e.g., number of terms served by a politician). We believe that the characteristics of the job performance construct must be clearly articulated before the question "Is job performance normally distributed?" can be adequately addressed.

That is, before broad inference about the *construct* can be made, we argue that several specific characteristics of job performance *measures* must be present. If one or more of these characteristics are absent in a measure of job performance and that measure is found to be nonnormally distributed, the lack of normality cannot be unambiguously attributed to the underlying distribution of the job performance construct. Rather, we will show that in some cases the absence of these characteristics can account for the vast departure from normality.

In the next section, we articulate what we see as critical characteristics of job performance measures. Next, we provide several empirical demonstrations of how the presence versus absence of these characteristics can affect the observed distributions.

Prototypical Characteristics of Job Performance Measures

A wide range of measures are used as indicators of the job performance construct. These include objective and subjective measures, short- and long-term measures, measures of individual behaviors as well as aggregations of multiple behaviors, and measures of results. These measures are also collected on various samples of performers, ranging from full populations of performers to highly select samples (e.g., studies restricted to top performers). Each of these choices is justifiable for particular settings and particular purposes. Our argument here is that when particular measures and particular samples are used to draw broad inferences about the underlying distribution of job performance, the measures and samples used need to reflect a set of features that make them consistent with conceptual definitions of the *construct* of job performance.

We take as our starting point a very useful definition put forth by Motowidlo and Kell (2013): "job performance is defined as the total expected value of the discrete behavioral episodes an individual carries out over a standard period of time" (p. 82). Note three important aspects of this definition. First, the focus is on behavior, explicitly behavior that is valued by the organization. Second, the notion of "total expected value" indicates the need to include the full range of valued behaviors; thus individual indicators of portions of the domain are deficient. Third, the notion of a standard period of time indicates that when one wishes to compare individuals, the comparison must cover a common time interval

(i.e., the opportunity to perform must be comparable across individuals whose performance is being compared).

Building on this definition, we offer seven conditions that a performance measure must meet if one wishes to draw inferences about the underlying performance construct. It is critical to point out that we are not asserting that performance indicators failing to meet these conditions should not be used. They may indeed serve many useful organizational and research purposes. We are simply asserting that measures not meeting these conditions are not suitable for drawing inferences about the underlying distribution of the job performance *construct*. We return to this point after outlining the seven conditions. Later in the article, we illustrate how the characteristics of job performance measures affect the observed distributions using a variety of empirical examples.

Performance Reflects Behavior, Not Situational Factors

Arguably the most central characteristic of job performance is that it reflects behavior (e.g., Campbell, Dunnette, Arvey, & Hellervik, 1973; Campbell, McCloy, Oppler, & Sager, 1993). Many performance measures are designed to directly sample behavior, as in the case of supervisors monitoring the work behaviors of subordinates. At the same time, many operational measures reflect work outcomes rather than behavior per se. Some common examples include counts of units produced and measures of sales volume. When outcome measures are studied, it is crucial that behavior and outcome are closely linked. In other words, the outcome should reflect individual behavior rather than some external factor. For one vivid example, consider a study by Jones and Terris (1981) of the performance of Salvation Army bell ringers during the Christmas holiday season. Behavior was not observed or evaluated; rather, the outcome measure was dollar volume collected per shift. Clearly, being assigned to the busiest street corner in downtown Chicago versus being assigned to a low-traffic strip mall makes a difference; location is a situational constraint that needs to be considered when attempting to compare individuals. Fortunately, the Salvation Army maintains historical data on yield per location, and Jones and Terris were able to control for location in their study.

Performance Reflects an Aggregation of Multiple Behaviors

Individual indicators of performance are typically deficient (Austin & Villanova, 1992; Dunnette, 1963). Stated differently, a count of any single behavior or outcome may be informative about some aspects of

a person's performance but will almost certainly miss other aspects of performance. For instance, research productivity is one aspect of academic faculty performance, yet so are teaching, advising, and service. Thus, overall performance measures, which may reflect an aggregation of measures of performance on individual tasks, are commonly reported in the industrial-organizational (I-O) psychology literature. In recent years, it has become common to conceptualize overall performance as reflecting three subdomains, namely, task performance, organizational citizenship behavior, and avoidance of counterproductive work behavior (CWB; Rotundo & Sackett, 2002), with each of these subdomains also reflecting an aggregate across multiple behaviors. It may be reasonable to expect that the distribution of performance on individual tasks or activities would be highly skewed. For example, many CWBs may be rare or extreme events, such as stealing from work, falsifying records, or destroying company property. The same is true of many organizational citizenship behaviors (e.g., staying late to help a coworker meet a deadline). However, when multiple indicators are combined, these departures from normality may be expected to "come out in the wash," with the aggregate measure approximating a normal distribution. That is, although one particular aspect of behavior may be extreme or rare, the more rare or extreme events that are considered, the more opportunity individuals have to engage in at least *some* of them.

The Full Range of Performers Is Included

In order to reach conclusions about the distribution of job performance in a population of performers, the sample on which performance measures are obtained and examined should represent the *full* population of interest. However, some measures are gathered on restricted samples, thus limiting our ability to draw inferences about the distribution of performance in the population of interest. For example, a measure of average golf score among professional golfers winning at least one tournament is based on a restricted sample, as most professional golfers have never won a single tournament. Limiting a sample to those exceeding a performance threshold prevents drawing inferences about the distribution of performance in the broader population of interest.

Note that calling for inclusion of the full range of performers is not the same thing as including all performers. For example, in many studies of performance, individuals are excluded because they have not been on the job long enough for a reliable measure of their performance to be obtained (e.g., excluding probationary employees from a test validation study). The concern we are raising here is of excluding individuals *because* of their

performance level, not for excluding individuals for whom performance cannot yet be meaningfully assessed.

The Full Range of Performance Is Included

It has long been understood that one's maximal level of performance can be different from his or her typical level of performance (Sackett, Zedeck, & Fogli, 1988). That is, there tends to be within-person variance in job performance (e.g., Barnes, Reb, & Ang, 2012; Dalal, Lam, Weiss, Welch, & Hulin, 2009), such that sometimes individuals perform at levels above their personal mean and sometimes they perform at levels below their mean. Thus, there is typically a *range* of levels of job performance across performance episodes for each person, and performance measures usually attempt to aggregate across these episodes. Yet, some job performance measures may focus solely on performance episodes at either the high or the low end of the performance distribution. At the high end, consider recognition for exemplary performance episodes, such as rewards and recognitions. Measures that focus on exemplary performance episodes fail to recognize or differentiate between levels of performance that fall *below* the exemplary threshold. Rather, performance is dichotomized into "above threshold" and "below threshold." At the low end, consider an organization that issues formal disciplinary reports for serious violation of work rules. A researcher interested in CWB may go through archives and count the number of disciplinary reports for each employee. The result may be a highly skewed distribution, with the vast majority of employees having a count of zero, a sizeable number with one violation, and very small numbers with multiple violations. However, because a violation needed to be serious in order to be written up, the measure does not reflect the full range of counterproductive behavior, as less severe behaviors are not included.

Performance Is Time-Bounded, Reflecting a Common Opportunity to Perform

The annual performance appraisal is a fixture in organizations, meaning employees are typically evaluated after having performed over the same length of time. When intervals of performance vary across individuals, performance measures are commonly adjusted for the time difference to make a comparison across individuals meaningful. A common example in our field is consideration of research productivity when hiring faculty members. Consider a setting where a department is authorized to make a hire at the assistant professor level. One candidate has 2 years'

experience and has five publications since completing the doctorate; a second candidate has 5 years' experience and has six postdegree publications (assume publication quality is comparable for the two candidates). We suspect that the typical evaluation process in such a setting does not say "6 is greater than 5, therefore candidate 2 is preferred," but rather "in terms of publications *per year*, candidate 1 shows more promise." That is, selection committees are likely to understand the influence of the differences in opportunity to perform on this metric of performance and to take these differences into consideration when making their decisions. Thus, we view comparability of the time period in which performance is evaluated as a prototypic feature of the job performance construct. This is typically achieved either by focusing on a common fixed period of time for all individuals being compared or by computing a measure of performance per unit of time, as in the above example.

Performance Comparisons Focus on Individuals Performing a Comparable Job

Related to the issue of opportunity to perform, substantive comparisons of individual performers require that they be performing a common role. Consider the job of professor of psychology at a research university. One component of performance is publication in scholarly journals, and one might create a count of publications, perhaps weighting by journal prestige or by number of authors. Note that the starting point here is the position of professor of psychology at a research university, where the importance of publishing articles (in terms of promotion, compensation, etc.) is approximately equal across individuals. But in the realm of scholarly publication one might take a different starting point, focusing on the act of publication rather than the job role of the individual publishing the article. Thus, a count of the number of times an author has published in a specific set of journals might be obtained. Yet, by conditioning on publication, rather than job role, such a publication count includes professors for whom publication is a key component of the job, students entering the publication process for the first time, practicing psychologists whose job role does not include publication but who devote discretionary time to the occasional publication, and managers in organizations who receive author credit for facilitating access to organizational data, among others. These individuals likely devote vastly different amounts of time to scholarly publication based on the importance of publication for their particular job. Such a count may be interesting for various purposes, but as it pools information from individuals in vastly different job roles it does not shed light on the distribution of job performance.

The Measure Is Not Distorted by Motivational Forces in the Performance Evaluation Process

Ratings of performance are widely used in organizational settings. However, as O'Boyle and Aguinis (2012) noted, these are not well suited to shedding light on questions about the underlying distribution of job performance due to a variety of possible constraints. In some settings, forced distributions are used, such as those imposing normality regardless of the true distribution. In other settings, operational ratings are subject to organizational norms and customs. For example, we have encountered settings where virtually all employees receive the highest rating, thus permitting raters to avoid the interpersonal difficulties that may result from giving a low evaluation. Ratings gathered for research purposes are likely subject to fewer motivational pressures to distort than operational ratings, but we are still less than confident that research ratings are free of such pressures. Some raters may not trust the assurance that the ratings are only for research purposes. Other raters may retain rating habits from the operational rating context. We suggest that anyone attempting to draw inferences about underlying performance distribution from ratings data has the burden of making a persuasive case that motivation to distort is not a significant factor in the rating context under examination.

Summary

We have presented and discussed seven characteristics of performance measures that need to be present in order to draw sound conclusions about the distribution of job performance. Specifically, these measures are not highly affected by situational factors, reflect aggregates of key tasks, reflect the full intended population of performers, reflect performance differences throughout the entire distribution, are time bounded, focus on individuals performing a common job role, and are not affected by motivational pressures within the rater. Broad conclusions about the distribution of job performance need to include data that meet these criteria.

Empirical Demonstrations of the Effects of Operationalization Choices on the Distribution of Job Performance Measures

In the remainder of this article we offer evidence that the features identified above can affect the distribution of performance measures. We have organized the following section around specific issues we would like to address, and thus have not used the traditional "introduction → method → results → discussion" format. Instead, we have organized our article

around broad issues, drawing on various data sets where appropriate. Some data sources are relevant for multiple issues. In this case, data sets will be described in greater detail when they are first introduced, and less detail will be provided in subsequent instances in which the data are presented.

We do not claim that all of the measures we examine meet all of our seven criteria; rather; we aim to offer concrete demonstrations of settings in which varying one of our criteria alters the distribution. Stated differently, we aim to illustrate how failing to meet any of our seven criteria can lead to large observed departures from normality. This is not to say that variables that do not meet all seven criteria should not be studied but, rather, that when the seven criteria are not met, large departures from normality cannot be unambiguously attributed to a departure from normality in the job performance construct. Instead, when the seven criteria are not met, large departures from normality may be driven by measurement characteristics. Below we compare distributions by plotting the histograms of these performance indicators and by testing the degree to which a normal distribution fit the data, relative to a number of highly skewed distributions. To do so, we used the Decision Tools Suite program @Risk which is an add-on to Microsoft Excel (Palisades Corporation, 2009). This is the same software used by O'Boyle and Aguinis (2012). However, whereas O'Boyle and Aguinis reported how well the Paretian distribution fit their data compared to the normal distribution, in this manuscript when we report the fit of the skewed distribution we are referring to the results for the exponential distribution. This is because using the exponential distribution the @Risk program was able to converge for nearly all data sets, whereas the Paretian distributions failed to converge in several cases. However, in instances where more than one skewed distribution converged (e.g., exponential and Paretian), the results regarding the skewed distributions provided the same interpretation. Thus, we report the results from the exponential distribution.

Before presenting our empirical results, it is important to note that "fit" is a continuum rather than a dichotomy. Thus, we illustrate the degree to which a normal distribution provides a *better* (or worse) fit than a highly skewed distribution. Even if a normal distribution provides a better fit to a data set than a highly skewed distribution, there may still be departures from normality. However, when a normal distribution provides a better fit to a data set than a highly skewed distribution, this is an indication that assumptions of normality are *more tenable* than assumptions that job performance will be highly skewed (as described by O'Boyle & Aguinis, 2012). In other words, evidence that a normal distribution fits better than a highly skewed distribution is evidence that in general job performance is distributed *approximately* normally, even if not *perfectly* normally.

Performance Must Reflect Behavior

We have argued that to draw broad conclusions about the job performance construct, the measure of job performance used must reflect individual behavior rather than situational factors. The measure itself may be a behavioral one, or an outcome measure if a persuasive case can be made that the outcome reflects individual behavior. Below we present data from three quite different settings.

Work task in laboratory. Perhaps the simplest way to examine performance in a setting in which there is a close match between performance and behavior is via a laboratory study. Lab studies can be designed such that performance is a function of behavior, and importantly, studies conducted in the lab can avoid other issues identified above. Specifically, all individuals perform the same task, the full range of performers is included (e.g., standing on the performance measure does not determine whether one's performance is included in the analysis), the full range of performance is included because performance can be defined by the researcher, all participants perform the same number of trials so there are no differences in opportunity to perform, performance can be defined as an aggregation of behaviors, and the motivation to distort is not an issue because behavior can be directly observed. Thus, we examined the distribution of performance using a simulated work task. These data were originally published by Beck and Schmidt (2012).

Beck and Schmidt (2012) conducted two studies ($N = 85$ and $N = 86$) to examine the relationships among self-efficacy—one's confidence in his or her abilities to perform a task (e.g., Bandura, 1997)—and performance. In both studies participants performed a multiple-cue probability learning task (MCPLT) involving the stock market in which they had to choose a stock based on several features (e.g., the stock's long-term performance rating). Participants were awarded points based on their degree of accuracy, and these points were used to determine whether participants would be eligible for a cash prize. Thus, there was an incentive for participants to take the task seriously and perform as well as possible. In both studies, participants performed six blocks of the MCPLT, yielding 510 and 516 observations of performance for Study 1 and 2, respectively. Each block was comprised of 10 trials, where a trial consisted of choosing a stock and submitting an answer. Thus, performance on a *block* was the sum of performance on 10 *trials*. To put this in the terms of our review of the job performance construct, performance on each block of the Beck and Schmidt studies was the *aggregation* of 10 behaviors (stock choices). In Beck and Schmidt's studies, performance was predicted by self-efficacy, and this effect was mediated by resource allocation.

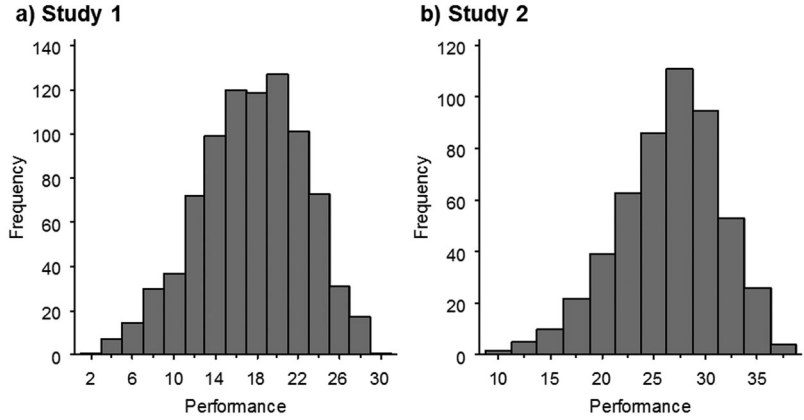

Figure 1: **Distributions of Beck and Schmidt's (2012) Multiple-Cue Probability Learning Task Performance Data.**

As shown in Figure 1, a normal distribution provided a better fit than the skewed distribution in both Study 1 ($\chi^2_{\text{normal}} = 476.30$, $\chi^2_{\text{skewed}} = 1395.13$, *ratio* $= 2.93$) and Study 2 ($\chi^2_{\text{normal}} = 21.98$, $\chi^2_{\text{skewed}} = 916.73$, *ratio* $= 41.71$). The ratio is computed as the quotient of the skewed distribution chi-squared value divided by the normal distribution chi-squared value. Because lower chi-squared statistics indicate better fit, a ratio value greater than 1 indicates that the normal distribution provided a better fit than the skewed distribution.

Inbound call center. We also assessed the distribution of job performance in a field setting. These data are from customer service representatives ($N = 324$) at an inbound call center in an insurance company and were originally described by Rothbard and Wilk (2011). In this company, the call center employees receive phone calls from customers with various needs or problems (e.g., filing a claim following a loss). The employee's task is to help the customer with his or her needs and to do so in as little time as possible so that other customers can be helped. Therefore, a common criterion for evaluating the performance of inbound call center employees is the average amount of time he or she spends with each customer, or average handle-time (AHT). Although the employee may need to be on the phone with any one customer for a long time for reasons that are out of the employee's control, such as trying to solve a particularly difficult technical problem, these extreme episodes should be equally distributed across employees. Thus, *average* handle-time should be largely a function of the employee's behavior, such as following procedures and scripts, rather than a function of external influences. Thus, we would

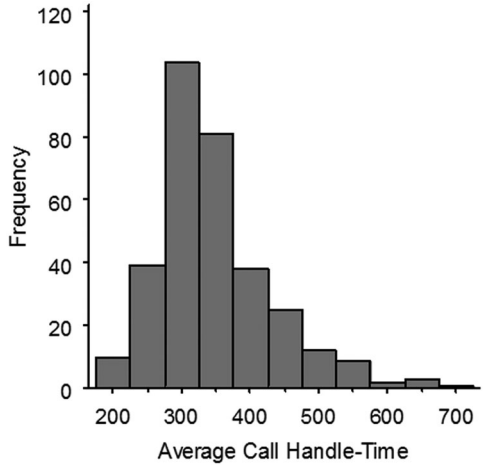

Figure 2: **Distribution of Rothbard and Wilk's (2011) Inbound Call Center Average Call Handle-Time Data.**

expect a normal distribution to provide a better fit to the AHT data than a highly skewed distribution. As shown in Figure 2, this was indeed the case ($\chi^2_{\text{normal}} = 69.33$, $\chi^2_{\text{skewed}} = 210.33$, $ratio = 3.03$).

National Hockey League. We also compiled data from the 2005/2006–2011/2012 seasons of the National Hockey League (NHL; www.nhl.com). These data included 4,046 observations from 1,127 forwards and 2,077 observations from 582 defensemen. Because forwards and defensemen perform different job roles, we have separated these two groups of players in the analyses below. One commonly used indicator of a player's performance in the NHL is his "plus/minus," which is defined as follows: a player receives +1 if he is on the ice when a goal is scored *for* his team and –1 if he is on the ice when a goal is scored *against* his team. Plus/minus is a very important indicator of performance because it may capture aspects of behavior that contribute to organizational effectiveness (in this case, the team's record) that are not reflected in other indicators (e.g., goals, which are discussed in a subsequent section). For instance, a player may be very adept at forcing turnovers in the opponent's offensive zone, an event that can end up in a goal being scored for the player's team. However, to force a turnover this player would not necessarily touch the puck and thus would not record a goal, assist, or point on the scoresheet. Nonetheless, this player's *behavior* would have been integral in scoring the goal, and because he was on the ice when it was scored, he would receive a +1 on his plus/minus rating. As would be expected of

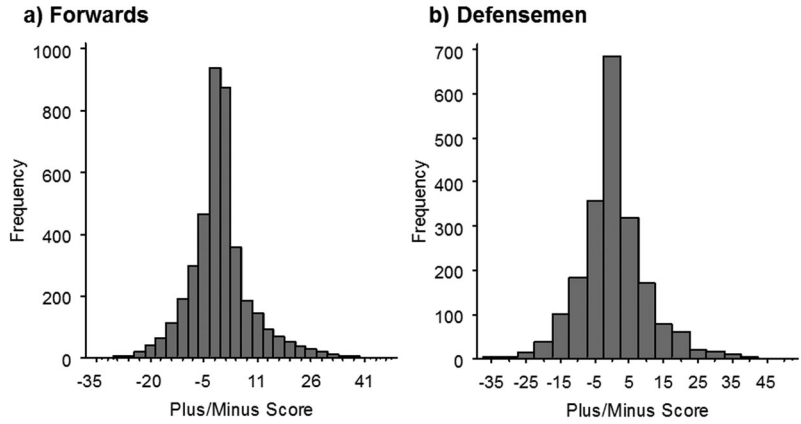

Figure 3: **Distributions of Plus/Minus Data in the National Hockey League for Forwards (a) and Defensemen (b).**

performance indicators that are closely tied to behavior, a normal distribution provided a better fit than a highly skewed distribution for both forwards (Figure 3a; $\chi^2_{\text{normal}} = 6{,}501.72$, $\chi^2_{\text{skewed}} = 13{,}584.49$, *ratio* $= 2.09$) and defensemen (Figure 3b; $\chi^2_{\text{normal}} = 1{,}359.22$, $\chi^2_{\text{skewed}} = 5{,}779.84$, *ratio* $= 4.25$).

Performance Is an Aggregation of Behaviors

We have also argued that performance is most accurately conceptualized as an aggregation of behaviors, rather than one specific behavior. Furthermore, although individual behaviors may follow a highly skewed distribution, we expect a distribution that is much closer to normality when multiple behaviors are aggregated, as is typically done when conceptualizing job performance. We first demonstrate this principle with empirical data collected from a work setting. Second, we illustrate how aggregation of skewed variables can result in a normally distributed composite variable using a computer simulation.

Counterproductive work behaviors. We first illustrate how aggregation of skewed indicators can result in a normally distributed composite variable using data originally published by Sackett, Berry, Weimann, and Laczo (2006). Specifically, Sackett et al. collected self-reported CWB data from 900 nonacademic university employees. Sackett et al. administered 35 items that measured the frequency with which employees engaged in specific CWBs (e.g., "Taken property from work without permission," "Had your performance affected due to a hangover from alcohol"). Participants responded on a four-point scale (1 = *never*; 2 = *rarely*;

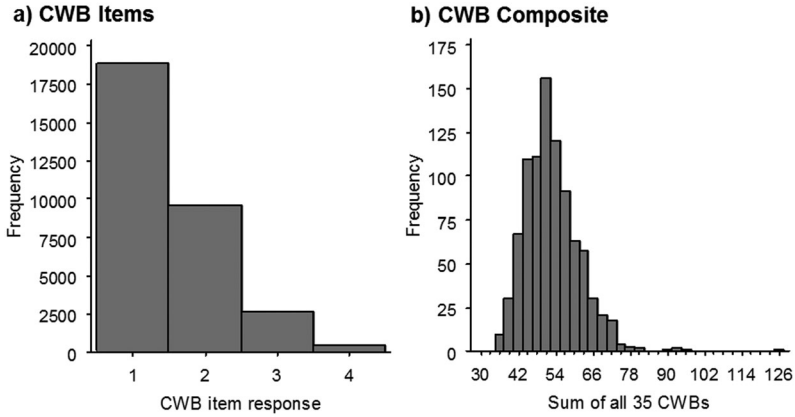

Figure 4: **Distributions of Sackett et al.'s (2006) Individual CWB Items (a) and CWB Composite Score (b).**

3 = *occasionally*; 4 = *frequently*). Responses to all 35 items were summed to create a CWB composite variable.

The frequency of the responses to all 35 CWB items for each of the 900 participants is plotted in Figure 4a. In other words, Figure 4a contains all 31,500 (35 items × 900 participants) responses reported in the Sackett et al. (2006) study. Although we show one graph pooling responses across all 35 items for ease of presentation, the distributions of responses to individual items follow a comparable pattern. Specifically, when asked how often they engaged in specific CWBs, participants responded "never" approximately 60% of the time and "rarely" approximately 30% of the time. "Occasionally" and "frequently" were only given as responses approximately 8% and 1% of the time, respectively. Thus, it appears that the CWB indicators follow a skewed distribution, with the majority of employees falling at the low end of the distribution and a decreasing number of employees at each subsequent level of engagement in CWBs. Given the discrete nature of the responses to these items (there were only four levels), it was not possible to empirically test how well these data fit normal and skewed distributions because these distributions assume continuous level data. Nonetheless, the data plotted in Figure 4a give a fairly clear indication that the CWB indicators were not distributed normally.

Next we assessed the degree to which a normal distribution fit the composite CWB variable data, relative to a skewed distribution. In line with our reasoning, a normal distribution provided a better fit to the composite CWB data than a skewed distribution ($\chi^2_{\text{normal}} = 79.76$, χ^2_{skewed}

= 211.52, *ratio* = 2.65). The distribution of composite CWB scores is plotted in Figure 4b. Thus, aggregating skewed performance indicators can result in a normally distributed composite performance variable.

Simulation. As stated earlier, it may be reasonable to expect individual behaviors to be skewed. Yet, when multiple indicators are combined, these departures from normality are expected to "come out in the wash," as the more behaviors that are considered, the greater the chances an individual will have engaged in at least *some* of them. Furthermore, the greater number of behaviors that are considered, the less the expected departure from normality of the aggregate job performance variable. However, adding additional indicators to the aggregate is only expected to reduce the departure from normality to the extent that these tasks or indicators are not redundant. That is, if the indicators are highly correlated, then the *same* individuals are the ones with opportunities to perform on *multiple* indicators, and thus a highly skewed distribution is likely to remain. This is consistent with established findings from the psychometrics literature (e.g., Cronbach & Warrington, 1952).

We assessed these predictions with a simulation. Specifically, we simulated highly skewed data with a Paretian distribution for a number of "indicators" of performance. We varied the number of indicators (2 to 40) and average intercorrelation (.00 to .80 in steps of .05) among the indicators. Thus, we created 663 simulation conditions corresponding to each combination of number of indicators (2 to 40 = 39 total) and average intercorrelation (.00 to .80 by .05 = 17 total). For each simulation condition, data for 1,000 "individuals" were simulated. The indicators (between 2 and 40, depending on the simulation condition) were then summed to compute a *composite score*. The skewness statistic of the composite score was then recorded. This procedure was repeated 1,000 times for each of the 663 simulation conditions. The average skewness of the composite score across all 1,000 trials was used to determine the relationships among the number of indicators, average intercorrelation, and skew of the composite. Thus, our final simulated data set contained 663 observations and three variables: number of indicators, average intercorrelation of indicators, and average skewness.

As shown in Figure 5, the more indicators included, the lower the skew of the composite variable ($\beta = -.44$). However, there was a curvilinear effect, indicating that the effect of adding indicators decreased as more indicators were added ($\beta = .23$). Using the same simulation, we found that average item intercorrelation had a main effect on the skew of the composite variable such that the higher the average intercorrelation among the indicators, the higher the skew of the composite variable ($\beta = .93$). More importantly, our simulation also showed that the curvilinear effects of the number of indicators was moderated by the average intercorrelation

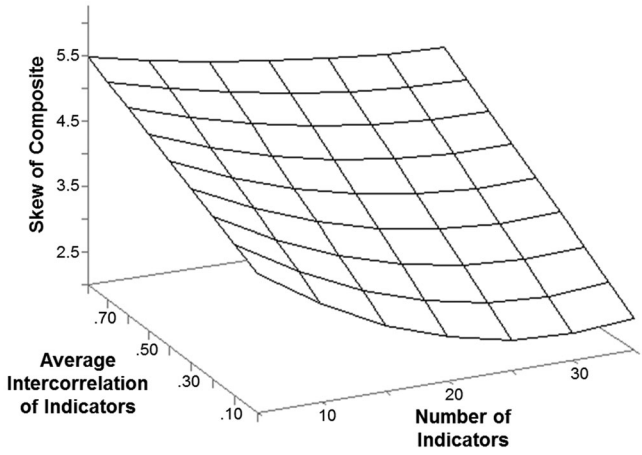

Figure 5: **Interactive Effects of Number of Indicators and Average Intercorrelation Among Indicators on the Skewness of Composite Variables.**

among the indicators ($\beta = -.13$), such that the effect of number of indicators on composite skewness is accentuated when the average intercorrelation is low. The model presented in Figure 5 accounted for 96% of the variance in composite skewness.

Finally, to better illustrate the effects of number of indicators and average intercorrelation among indicators on *composite* skewness, we reran the simulation described above for each of the 663 simulation conditions, yet we generated a sample of 10,000 individuals for each condition (rather than 1,000 individuals 1,000 times). The histograms from several conditions are plotted in Figure 6. Taken together, these results demonstrate that performance can be normally distributed even if individual *indicators* of performance are not. More specifically, Figures 5 and 6 show that a normally distributed composite variable can be recovered as long as the indicators are not highly redundant and that adding more indicators (up to a point) can increase the likelihood of departures from normality "coming out in the wash." Thus, we interpret the results of this simulation to mean that observing a highly skewed distribution for one indicator of job performance is not evidence that the distribution of job performance as a whole vastly departs from normality. Rather, even if an indicator (or several indicators) of job performance is highly skewed, combining multiple skewed indicators into a multidimensional composite can result in an operationalization of job performance that is distributed in a manner closer to normality. Our simulation suggests that this is likely to be true

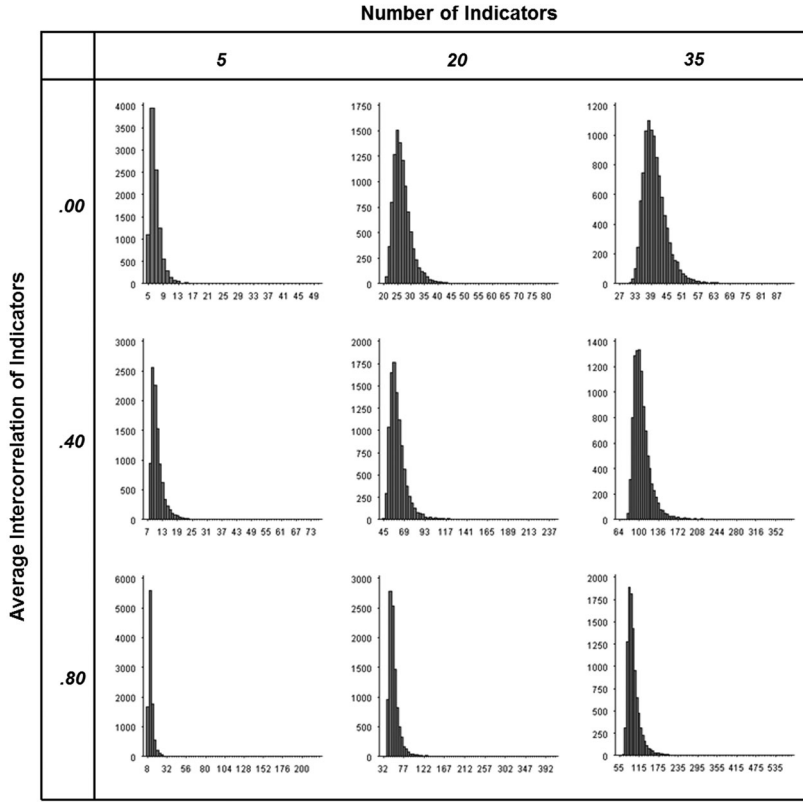

Figure 6: **Histograms Illustrating the Effects of Number of Indicators and Average Intercorrelation Among Indicators on the Skewness of Composite Variables.**

in all but the most extreme cases, such as when very few and/or highly redundant indicators are used to form the composite.

The Full Range of Performers

Simulation. The importance of including the full range of performers when attempting to draw inferences about the distribution of the job performance construct can be illustrated with a very simple simulation. Specifically, first we simulated job performance data with a normal distribution for 100,000 "individuals" (Figure 7a). Next, we created a subsample of simulated job performance scores including only the scores that were greater than or equal to one standard deviation above the mean. This is

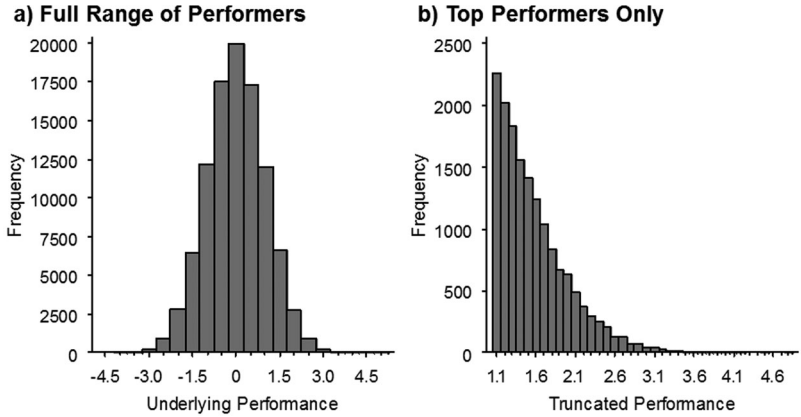

Figure 7: **Simulation Results Illustrating the Effect That Inclusion of Only Top Performers Has on the Distribution of Performance Scores.**

equivalent to sampling only elite performers, such as Professional Golfers Association tournament winners, Academy Award winning actors, or U.S. senators (cf. O'Boyle & Aguinis, 2012). As shown in Figure 7b, the resulting distribution was highly skewed ($\chi^2_{\text{normal}} = 6{,}948.51$, $\chi^2_{\text{skewed}} = 538.98$, $ratio = .08$), even though the underlying distribution was normally distributed ($\chi^2_{\text{normal}} = 186.82$, $\chi^2_{\text{skewed}} = 223{,}084.24$, $ratio = 1{,}194.82$).

Marathon runners. In addition to our simulation results, we are also able to illustrate the importance of including the full range of performers when attempting to draw inferences about the distribution of job performance using observational data. Specifically, we obtained performance data from marathon runners. The first data set included the finish times of all 505,238 runners who participated in and finished a marathon in the United States in 2010. As shown in Figure 8a, when the full range of marathon performers is included, the performance distribution is approximately normal ($\chi^2_{\text{normal}} = 269{,}091.94$, $\chi^2_{\text{skewed}} = 310{,}177.79$, $ratio = 1.15$). Conversely, when only elite marathon runners are considered, a highly skewed distribution emerges. Specifically, we compiled performance data for a group of elite marathon runners—those who qualified for the U.S. Olympic marathon trials in 2012 (USA Track and Field, 2012). To qualify for the U.S. trials, a male runner needed to finish a marathon in 2:19:00 or less. This would only include runners in "Finish Time Group = 15" in Figure 8a. The number of times this standard was met by the 81 men who qualified in 2012 is plotted in Figure 8b. Similar to the CWB item-level data presented above, the discrete nature of this performance indicator (there are only five levels) precludes empirically

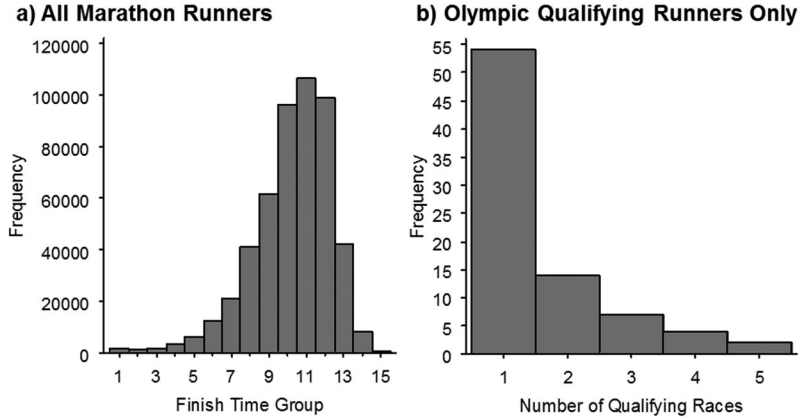

Figure 8: **Distributions of Marathon Race Performance for All Marathon Runners (a) and Runners Qualifying for the U.S. Olympic Marathon Team (b).**

Note. "Finish time group" in panel (a) refers to a range of finishing times. Higher numbers indicate better marathon performance (i.e., shorter finish time). 1 = 9:00:00–14:10:58; 2 = 8:30:00–8:59:59; 3 = 8:00:00–8:29:59; 4 = 7:30:00–7:59:59; 5 = 7:00:00–7:29:59; 6 = 6:30:00–6:59:59; 7 = 6:00:00–6:29:59; 8 = 5:30:00–5:59:59; 9 = 5:00:00–5:29:59; 10 = 4:30:00–4:59:59; 11 = 4:00:00–4:29:59; 12 = 3:30:00–3:59:59; 13 = 3:00:00–3:29:59; 14 = 2:30:00–2:59:59; 15 = 2:05:52–2:29:59.

testing how well these data fit normal and skewed distributions. Yet, similar to the CWB item-level data, Figure 8b indicates a large departure from normality.[1]

The Full Range of Performance

Similar to the point made above, the full range of performance must be included in order to make inferences about the underlying distribution of job performance. That is, along with varying between people, job performance also varies within individuals over time (e.g., Barnes et al., 2012; Dalal et al., 2009; Sackett et al., 1988). Given this variance, some performance episodes are extreme. Performance measures may capture only extreme positive performance episodes, such as winning an award, and performance measures may capture only extreme negative performance episodes, such as being written up for an infraction. If only the most extreme performance episodes are included in a measure of job

[1]The same pattern emerges in examining the 101 women meeting the female qualifying standard of 2:46.

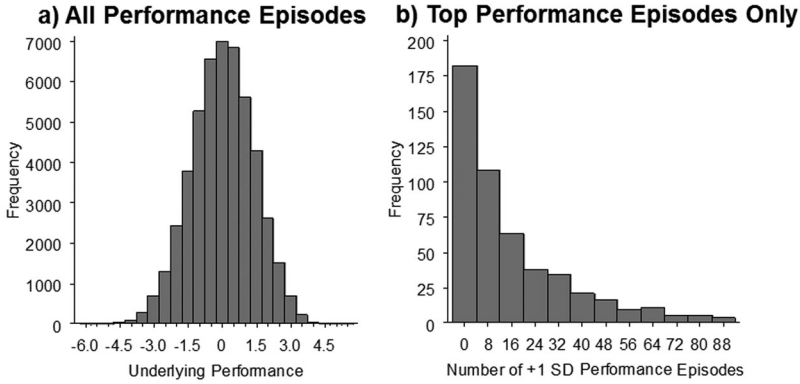

Figure 9: **Simulation Results Illustrating the Effect That Inclusion of Only Top Performance Episodes Has on the Distribution of Performance Scores.**

performance, a highly skewed distribution can result, even if the underlying performance distribution is normal. As with the full range of performers, the influence of failure to include the full range of performance on the distribution of job performance can be demonstrated with a simulation as well as with observational data.

Simulation. Data were simulated for 500 individuals, with multiple "performance episodes" for each individual (*Mean* = 100, *SD* = 10), totaling 49,501 observations. Thus, the simulation is analogous to measuring the job performance of 500 workers every day for approximately 100 days (on average). Given that the data were simulated, we were able to specify that *underlying performance* was normally distributed. In addition, in line with actual job performance data, our simulated data varied both between and within individuals ($ICC_{(1)} = .48$). Next, we computed the *extreme performance* variable. If performance at a given episode was greater than one standard deviation above the mean of the entire sample, this variable was coded 1; if performance was at or below 1 *SD* this variable was coded 0. This procedure is analogous to giving an award or recognition for each performance episode in which an individual does exceptionally well. For example, an actor may perform in 20 films (i.e., performance episodes) yet only win an Academy Award for one of the films (i.e., extreme performance). Finally, we summed the number of times each individual recorded a 1 to compute each individual's total *extreme performance* score.

Not surprisingly, a normal distribution provided a better fit than a skewed distribution to the underlying performance distribution (Figure 9a; $\chi^2_{\text{normal}} = 200.86$, $\chi^2_{\text{skewed}} = 117{,}555.17$, *ratio* = 585.27). Given that the data were simulated, and that we specified that the data should be generated

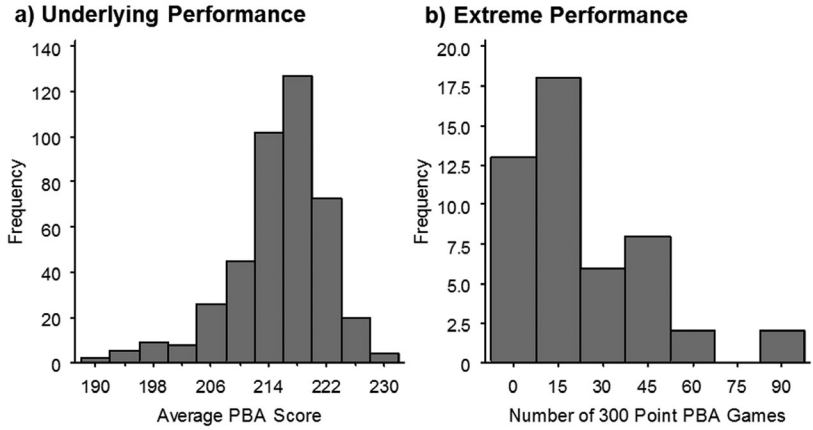

Figure 10: **Distribution of Average Score (a) and "Perfect Game" (b) Data in the Professional Bowlers Association.**

Note. PBA = Professional Bowlers Association.

using a normal distribution, this finding in and of itself is not theoretically meaningful. However, when only extreme performance episodes (i.e., those at or above +1 *SD* of mean level of performance) were "counted" as performance, a skewed distribution provided a better fit to the data (Figure 9b; $\chi^2_{normal} = 753.38$, $\chi^2_{skewed} = 242.72$, *ratio* = .32). Thus, when only extreme events are counted as performance, a highly skewed distribution can emerge, even if the underlying performance distribution is normal.

Professional Bowlers Association. Data were also obtained from 49 members of the Professional Bowlers Association (PBA; www.pba.com). Data from multiple seasons were available for each bowler, and the average bowler had been in the PBA for about 9 seasons (*Mean* = 8.57, *SD* = 3.74, *Min* = 2, *Max* = 14). *Underlying performance* was operationalized as each bowler's average score earned for each season. Thus, an individual bowler could have between 2 and 14 observations. *Extreme performance* was operationalized as the number of 300-point games (i.e., perfect games) a bowler had achieved during his time in the PBA.

As expected, we found that a normal distribution provided a better fit than a highly skewed distribution to the bowlers' average scores (Figure 10a; $\chi^2_{normal} = 45.04$, $\chi^2_{skewed} = 1{,}139.94$, *ratio* = 25.26). Conversely, the distribution of the number of 300-point games each bowler had accumulated over the course of his career, an *extreme performance* event, was better described by a skewed distribution (Figure 10b; $\chi^2_{normal} = 25.61$, $\chi^2_{skewed} = 3.41$, *ratio* = .13). Thus, when only extreme performance events were "counted" as performance, a skewed distribution

emerged, even though the underlying distribution was approximately normal.

Equal Opportunity to Perform

A critical criterion that must be met before making broad conclusions about the distribution of job performance is that each performer in the sample must have an equal opportunity to perform. That is, for comparisons of individual job performance to be meaningful, the amount of time over which each individual has performed should be equivalent. For instance, a meaningful comparison of two door-to-door salespeople in terms of sales volume must take into consideration the amount of time each salesperson was given to sell his or her products (among other things, such as location). That is, for performance indices that accumulate over time, such as sales volume, time must be held constant before comparisons among people are made. Using several data sets, we empirically demonstrate how failure to consider opportunity to perform can lead to misleading results with regard to the underlying distribution of performance.

We do so by presenting the quotient of performance divided by opportunity to perform. Thus, we present performance *per* unit of opportunity. Although this approach is relatively straightforward, it can be problematic for individuals with very low values for opportunity to perform and nonzero values for performance. For instance, consider an NHL player who only plays one game, yet happens to score a goal during that game. Because this player would have a very low value for time on ice, his goals per time on ice value would be exceptionally high. Yet this very high value is likely to be a function of dividing by a very small number rather than any true "superstar" qualities of the player. Thus, when presenting our results we have excluded individuals with very few opportunities to perform. Note that this is the conceptual analogue to the common practice of excluding individuals with very short tenure in research using performance measures in jobs settings. Low tenure employees are commonly viewed as probationary and have not had sufficient performance opportunities for a reliable assessment of their performance to be made.

The distributions of opportunity to perform for each data set (which are described in more detail below) are shown in Figure 11. As shown in Figure 11, in each data set there is a large increase in frequency at very low values of opportunity to perform. This may be due to the nature of the setting; specifically, in professional sports it is common for athletes from lower leagues to appear in a small number of games in order to replace injured or otherwise unavailable starting players. We argue that such players are not part of the population of interest (players from the National

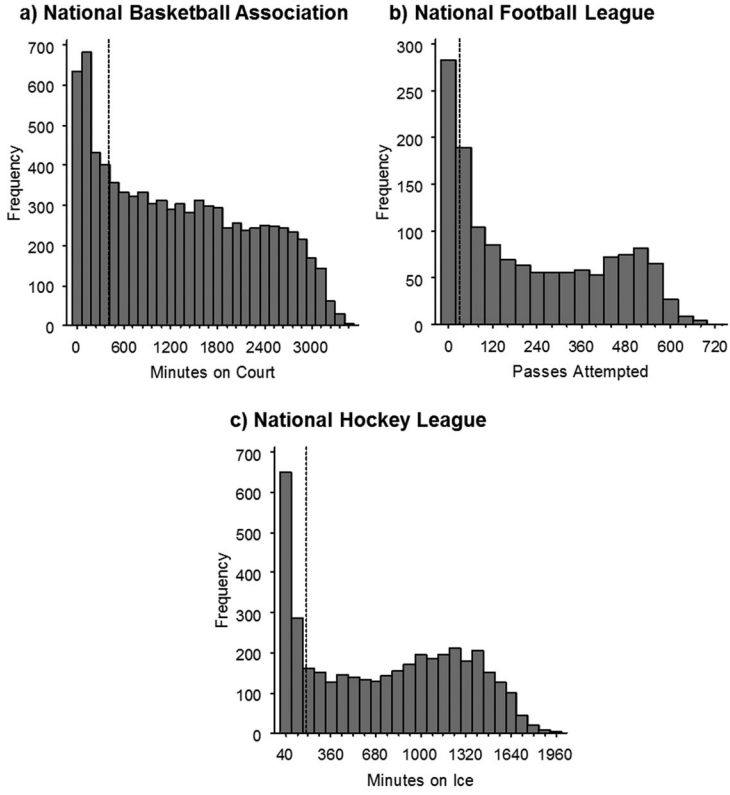

Figure 11: **Distributions of Opportunity to Perform in the (a) National Basketball Association, (b) National Football League, and (c) National Hockey League.**

Note. The dashed lines represent the first quartile for opportunity to perform.

Basketball Association, National Football League, and National Hockey League) and thus should not be included in the data set. Unfortunately, the distinction between "major league player" and "minor league player" is often not straight forward, as players from the lower leagues often move up to the major leagues on a relatively permanent basis as they gain experience, and players from the major leagues often move down to lower leagues when they fail to perform to the standards of the higher league. Thus, we are forced to make a relatively arbitrary cutoff when excluding players with too few opportunities to perform. For the purposes of this manuscript, we have made this cutoff at the first quartile for opportunity to perform, which is shown by the dashed vertical lines in Figure 11. Yet, we

have assessed a wide range of different inclusion cutoffs, and the specific value of the cutoff does not affect the interpretation of our results.[2]

National Basketball Association. Data were obtained from the 1990–2004 National Basketball Association (NBA) seasons, which were originally published by Rotundo, Sackett, Enns, and Mann (2012). These data included 8,473 observations from 1,673 players. *Performance* was operationalized as the number of points each player scored during the season, and *opportunity to perform* was operationalized as the number of minutes the player was on the court during the season. As shown in Figure 12a, when opportunity to perform was not taken into account, the distribution of points scored in the NBA was highly skewed (χ^2_{normal} = 6,460.10, χ^2_{skewed} = 1,093.74, *ratio* = .17). Yet, as shown in Figure 12b, taking opportunity to perform into account yielded a distribution that was much closer to normality (χ^2_{normal} = 194.18, χ^2_{skewed} = 7,962.38, *ratio* = 41.01).

National Football League. Next, we compiled data from the 1994–2012 National Football League (NFL; www.nfl.com) seasons. These data included 1,409 observations from 304 players. Only quarterbacks were included in this data set. *Performance* was operationalized as the number of passes resulting in a touchdown thrown by each quarterback over the course of the season, and *opportunity to perform* was operationalized as the number of passes each quarterback attempted. As shown in Figure 12c, when opportunity to perform was not taken into account, the distribution of touchdown passes in the NFL was highly skewed (χ^2_{normal} = 3,220.88, χ^2_{skewed} = 573.34, *ratio* = .18). Conversely, as shown in Figure 12d, a much more normal distribution emerged when opportunity to perform was taken into account (χ^2_{normal} = 71.79, χ^2_{skewed} = 952.04, *ratio* = 13.26).

National Hockey League. Finally, we used the data from the 2005/2006–2011/2012 NHL seasons, which were introduced earlier. These data included 4,046 observations from 1,127 players. Only forwards were included in this data set. *Performance* was operationalized as the number of goals each player scored during the season, and *opportunity to perform* was operationalized as the number of minutes the player was on the ice during the season. As with previous data sets, failing to account for opportunity to perform resulted in a highly skewed distribution (Figure 12e; χ^2_{normal} = 12,844.40, χ^2_{skewed} = 1,632.29, *ratio* = .13), yet accounting for opportunity to perform resulted in a more normal distribution (Figure 12f; χ^2_{normal} = 159.10, χ^2_{skewed} = 1,824.01, *ratio* = 11.46).

[2]These results are available from the first author upon request.

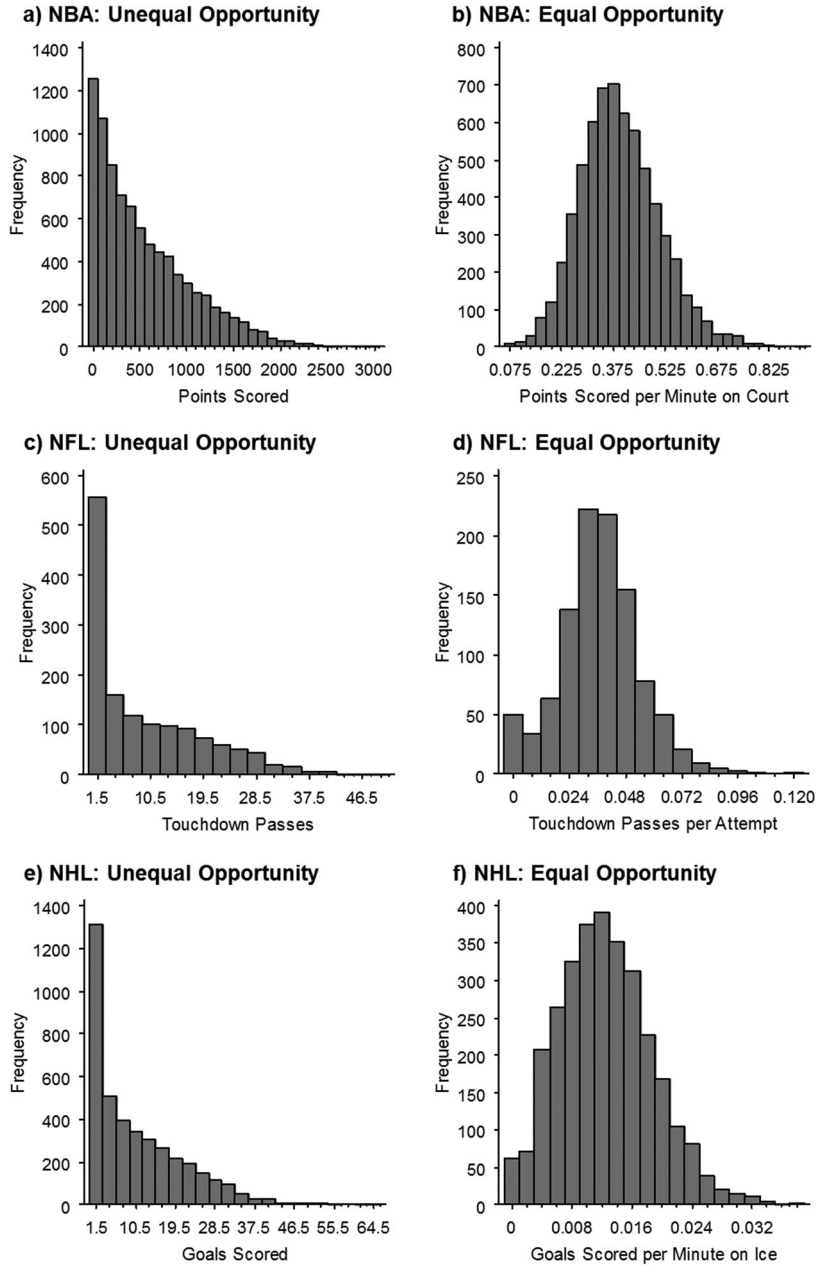

Figure 12: **Distributions Demonstrating the Importance of Accounting for Opportunity to Perform for Determining the Distribution of Performance.**

Note. NBA = National Basketball Association; NFL = National Football League; NHL = National Hockey League.

Comparable Jobs

Failure to match individuals on jobs impedes comparisons between those individuals on job performance, along with examinations of the distribution of job performance. At the extreme this is because many work behaviors are job specific. For example, it may be reasonable to use the number of tickets written to evaluate a police officer's job performance, but this same criterion does not make sense for evaluating a fire fighter's job performance. Yet, even behaviors that are exhibited on multiple jobs are not always directly comparable across jobs.

National Hockey League. For example, consider the data from the NHL that were presented above. In hockey there are three primary positions: goaltenders, defensemen, and forwards. Both forwards and defensemen score goals,[3] yet for forwards scoring goals is a primary function. On the other hand, for defensemen the primary function is defending one's own net, and scoring goals is a secondary function. Indeed, on average NHL forwards score over four times as many goals per minute as NHL defensemen ($M_{\text{forward}} = .013$, $SD_{\text{forward}} = .006$, $M_{\text{defense}} = .003$, $SD_{\text{defense}} = .003$, $t_{(4590)} = 59.51$, $p < .001$, $d = 1.82$).[4] As shown in Figure 13a, when job role is not taken into account, a highly skewed performance distribution is observed ($\chi^2_{\text{normal}} = 1{,}382.76$; $\chi^2_{\text{skewed}} = 1{,}117.58$; *ratio* $= .81$). However, for individuals for whom the performance indicator in question is a core component of job performance (in this case, forwards and scoring goals), the distribution is much closer to being normally distributed (Figure 13b, $\chi^2_{\text{normal}} = 131.96$, $\chi^2_{\text{skewed}} = 1{,}824.59$, *ratio* $=$ 13.83). Furthermore, the departure from normality that was observed when job role was not taken into account may be driven by those individuals for whom the performance indicator in question is *not* a core component of job performance (in this case, defensemen and scoring goals).

Publication records of I-O psychologists. We were also able to examine the effects that failing to account for job role can have on the distribution of performance in a more traditional work setting. Specifically, we used publication record data from a sample of I-O psychologists in academic positions. These data were originally published by Bowling and Burns (2010). The original sample included 300 I-O psychologists who were either working in an academic department with a research-oriented doctoral program ($n = 181$) or an academic department without such a doctoral

[3]It is also possible for goaltenders to score goals, yet this is an extremely infrequent event, having occurred only 13 times in the history of the NHL.

[4]Only players with time on ice values greater than the first quartile for both forwards and defensemen combined (240.9 minutes on ice) are included. The interpretation of the results does not change when other inclusion cutoffs are used.

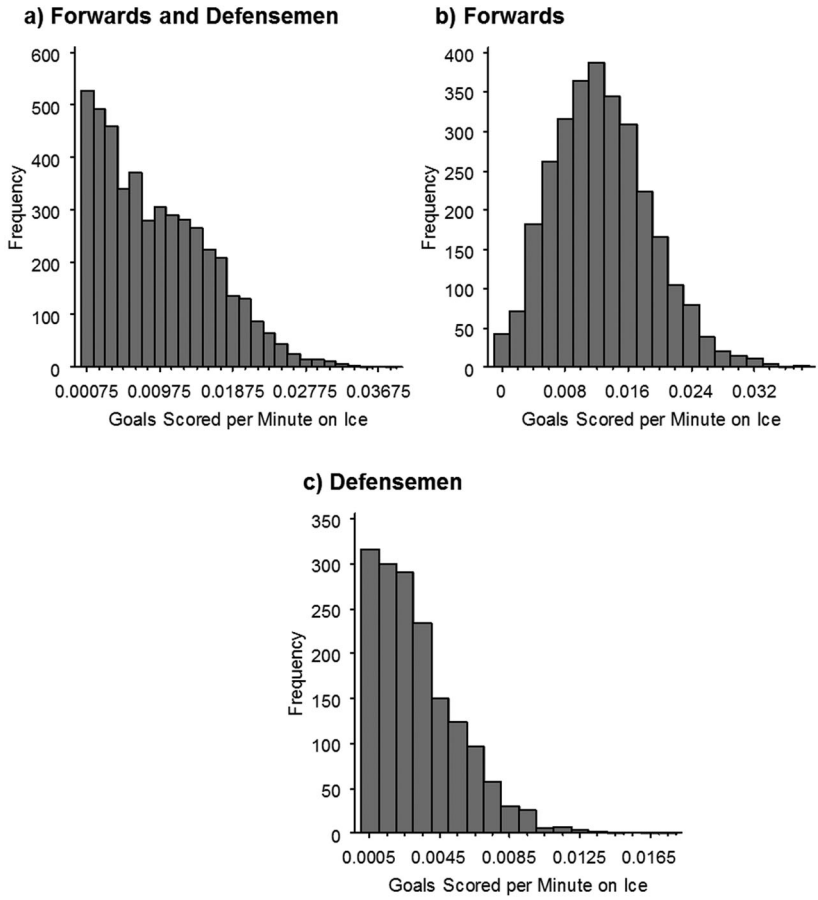

Figure 13: **Performance Distributions From the National Hockey League Demonstrating How Comparing Individuals With Different Job Roles Can Result in an Observed Departure From Normality.**

program ($n = 118$). At the time of the survey the average time elapsed since the respondent received his or her PhD was 19.67 years ($SD = 11.13$, $min = 3$, $max = 45$). For this manuscript we focused on individuals who were likely to be "pretenure," which we operationalized as 4 to 6 years post-PhD ($n = 36$).

We argue that I-O psychology professors in doctoral programs hold a different job role than I-O psychology professors in nondoctoral programs, with publishing being a more central job component in doctoral programs than nondoctoral programs. Indeed, pretenure professors in doctoral

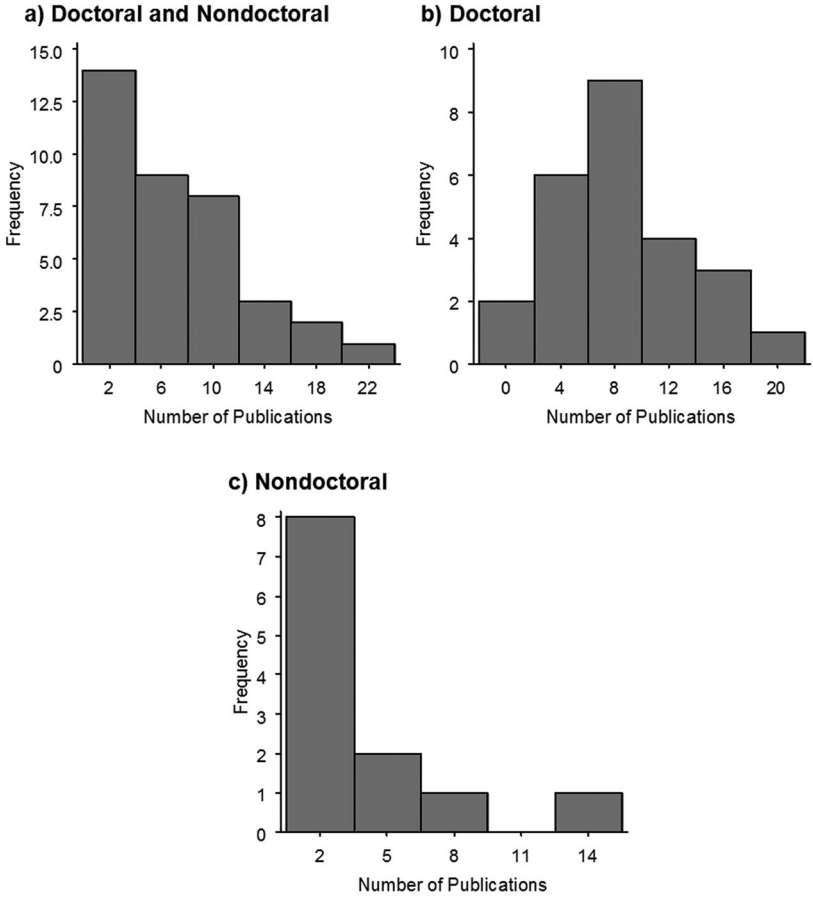

Figure 14: **Distributions of Bowling and Burns' (2010) I-O Psychologist Publication Data Demonstrating How Comparing Individuals With Different Job Roles Can Result in an Observed Departure From Normality.**

programs published more articles on average than pretenure professors in nondoctoral programs ($M_{\text{doctoral}} = 8.16$, $SD_{\text{doctoral}} = 5.15$, $M_{\text{nondoctoral}} = 3.75$, $SD_{\text{nondoctoral}} = 3.77$, $t_{(35)} = 2.64$, $p < .05$, $d = .95$). Furthermore, when job role (doctoral program vs. nondoctoral program) is not taken into account, the total number of publications produced by pretenure I-O faculty follows the highly skewed distribution depicted in Figure 14a. However, among those in departments with a doctoral program, a quite different distribution emerges as depicted in Figure 14b, whereas in nondoctoral programs the distributions was highly skewed, as shown in Figure 14c.

For these data, we rely on visual interpretation of the figures, as the @Risk program failed to converge in some cases and generally did not produce consistently interpretable findings due to the relatively small number of pretenure individuals. Finally, during the "posttenure" years, publication distributions take on a more skewed distribution in both doctoral ($\chi^2_{\text{normal}} = 80.92$, $\chi^2_{\text{skewed}} = 27.26$, *ratio* = .34) and nondoctoral programs ($\chi^2_{\text{normal}} = 77.71$, $\chi^2_{\text{skewed}} = 11.09$, *ratio* = .14). Similar to our explanation of goal scoring among NHL forwards and defensemen, we expect this finding may be explained by the degree to which publishing journal articles is a nonfungible core job activity. Specifically, during the pretenure period, publications are one of the key indicators of job performance in doctoral programs and as such are one of the primary determinants of pay and promotion. Conversely, in nondoctoral programs less emphasis is placed on publication when making pay and promotion decisions, making publication less of a core job activity. Similarly, during the posttenure phase, faculty members in both doctoral and nondoctoral programs can begin to make different choices about how their time is spent (e.g., continuing to publish in journals, writing text books, consulting, taking on administrative roles). We suggest that performance indicators that reflect core activities common to all performers are more likely to approximate a normal distribution, and indicators reflecting settings where performers have discretion as to which of a number of behaviors in which to engage are more likely to be substantially skewed.

No Motivation to Distort

Finally, we have argued that broad conclusions about the distribution of job performance cannot be drawn from subjective job performance ratings if there is motivation to distort on the part of the rater. In fact, O'Boyle and Aguinis (2012) found this to be such an important issue that they did not include any job performance measures that were based on performance appraisal ratings in their manuscript. Specifically, O'Boyle and Aguinis stated that raters who give very few high ratings and many low ratings are probably considered "severe raters" (and raters with the opposite pattern would be "lenient") and that this severity is likely considered to be an error. These raters may be screened out or removed before any analyses are conducted. Furthermore, O'Boyle and Aguinis cited several sources stating that raters should "force" ratings to follow the normal distribution. Given these potential problems, O'Boyle and Aguinis reached the conclusion that normally distributed performance appraisal ratings cannot be used to draw the inference that the underlying distribution of the job performance construct is in fact normal. We agree that motivation to

distort job performance ratings (e.g., not wanting to give one's subordinates low performance ratings) can be an impediment to drawing conclusions about the underlying distribution of job performance. However, we do not believe that the use of performance appraisal ratings is *necessarily* problematic. Specifically, even though raters can sometimes be motivated to distort performance ratings, this does not mean raters are *always* motivated to distort. Furthermore, when raters are not motivated to distort performance ratings, a normal distribution of performance ratings may emerge. To illustrate this point, we obtained the job performance ratings of 21,945 individuals nested within 117 jobs ($Mean_n = 188$, $SD_n = 116$) from research on the General Aptitude Test Battery (GATB). These data were originally collected over multiple decades from the 1940s to the 1980s as criterion measures in GATB validation studies (see Hartigan & Wigdor, 1989 for details). A total of 745 validity studies were conducted, using a wide range of criteria. We identified 117 that used a common six-dimension performance rating scale. No explicit instructions were given regarding how performance ratings should be distributed, and it was made clear to all raters that the ratings would only be used for research purposes. In line with the objective data reported above (where motivation to distort was not an issue), we found that a normal distribution provided a better fit to the data than a highly skewed distribution in all 117 samples (*ratios*: $Mean = 6.45$, $SD = 6.57$, $min = 1.39$, $max = 49.00$).

Although we cannot definitively rule out the possibility that raters may have "forced" a normal distribution even without being explicitly told to do so (e.g., perhaps they had been told to do so in the past), we believe these data are still useful for understanding the distribution of job performance. For one, a normal distribution provided a better fit to the performance ratings than a skewed distribution in all 117 samples. If job performance was truly highly skewed, we believe at least *some* raters would have captured this distribution given the circumstances under which the ratings were collected (no instructions about distribution, data collected for research).

Discussion

Summary of Results

The nature of the distribution of job performance, arguably one of the most important variables in I-O psychology, has been called into question (O'Boyle & Aguinis, 2012). Specifically, O'Boyle and Aguinis argued that job performance is not distributed normally but instead follows a nonnormal and highly skewed distribution, characterized by the highest number of performers falling at the low end of the distribution and by a

consistently decreasing number of performers at each subsequent level of performance. In support of this argument O'Boyle and Aguinis presented a large amount of data from a variety of sources in which job performance data were highly skewed. They reached the conclusion that job performance is not normally distributed and that theory, methods, and statistical procedures used in I-O psychology should be revised accordingly.

However, we do not agree with this conclusion. Rather, we have argued that in order for broad conclusions to be drawn about job performance at the *construct* level, several *measurement* criteria must be met. Specifically, performance measures should reflect behavior rather than results, performance measures should be an aggregation of multiple behaviors rather than a single indicator, performance measures should cover the full range of *performers* and *performance*, only individuals doing similar jobs should be compared, and motivation to distort performance ratings on the part of the rater should be minimized. To draw broad inferences about the job performance construct, measures of performance must reflect these characteristics. Using data from a broad range of sources, including the workplace, the laboratory, athletics, and computer simulations, we have demonstrated that when these criteria are *not met*, a nonnormal and highly-skewed distribution often emerges.

On the other hand, on the whole our data indicate that when these criteria are *met*, a normal distribution tends to provide a better fit to job performance data than highly skewed distributions (e.g., exponential, Pareto). This is not to say that every data set we have presented meets all seven of our criteria. Yet, when looking across all of our data sets, which include laboratory studies that were designed so that the seven performance criteria were met and simulations in which the underlying distribution of job performance was perfectly known (and normal), we do not see evidence that job performance, as it is typically defined, can be expected to be highly skewed. Thus, our central conclusion is that observations of vast departures from normality in job performance measures are likely due to features of the measures themselves, and we do not see evidence for a consistent finding of vast departures from normality in the distribution of job performance.

Contrast With O'Boyle and Aguinis' (2012) Conclusions

The conclusions we have drawn about the distribution of job performance are different than the conclusions drawn by O'Boyle and Aguinis (2012). Specifically, O'Boyle and Aguinis concluded that job performance likely does not follow a normal distribution and instead is characterized by a highly skewed Pareto distribution. On the other hand, we have

concluded that it is not the case that the distribution of job performance can generally be expected to depart so dramatically from normality. We believe this difference in conclusions stems primarily from a difference in the definition of job performance. We have argued, along with others (e.g., Motowidlo & Kell, 2013), that job performance has several defining features. Variables that do not possess these features cannot be used to make inferences about the distribution of job performance.

However, the variables included in O'Boyle and Aguinis' (2012) study lacked many of these features. For example, several of the variables were heavily dependent on others' actions rather than the individual's behavior (e.g., election wins, award nominations). In addition, many of the variables used by O'Boyle and Aguinis were unidimensional rather than aggregations of behavior (e.g., publishing represents only a portion of academic job performance). Furthermore, many of the variables examined by O'Boyle and Aguinis were restricted to top performers and, more so, only exemplary performance episodes (e.g., winning an Academy Award). Many of O'Boyle and Aguinis' variables represented performance over one's entire career and thus did not hold the opportunity to perform constant. Finally, some of O'Boyle and Aguinis' variables included individuals from different job roles (e.g., publishing by professors vs. publishing by consultants).

To be clear, we do not disagree that the variables studied by O'Boyle and Aguinis (2012) had distributions with vast departures from normality. We do, however, disagree with the conclusion that job performance is highly skewed. Rather, we contend that the variables in O'Boyle and Aguinis' study lacked critical features of job performance measures. Furthermore, our data demonstrate that failure to meet one or more of these criteria can result in vast departures from normality, yet when these features are present, a normal distribution is often more likely to emerge.

Theoretical Implications

The results presented in this article have several theoretical implications. For one, our results have implications for the study of what Aguinis, Gottfredson, and Joo (2013) called "interesting outliers," which are defined as "accurate data points that lie at a distance from other data points and may contain valuable or unexpected knowledge" (pp. 281–282). Aguinis et al. strongly cautioned against simply assuming all outliers are "error" or "noise" and subsequently applying transformations to the data or deleting the outliers altogether. Rather, because these outliers may lead to valuable theoretical insights, Aguinis and colleagues recommended these outliers be given special attention via quantitative and qualitative empirical research. We agree that understanding extremely high

performers can be very valuable for the field of personnel psychology. For example, organizations would benefit from being able to attract, select, and train extremely high performing individuals. However, given our results presented above, we believe that extreme departure from normality may in many settings be an artifact of measurement decisions. Thus, we urge caution when trying to study individuals with extremely high performance scores. That is, outliers may not be as extreme as they initially appear given the influence of measurement characteristics on the degree of departure from normality.

This extreme departure from normality could lead to overinterpretation of *apparently* very high performance scores and subsequently lead to misattributions about the nature of "superstar" performers. For instance, extremely high scores could be due to external influences (e.g., a salesperson being assigned to a high-traffic area) rather than behaviors and thus would not be likely to be replicated when these external influences were removed. Similarly, a person may achieve an extremely high level of performance on one criterion, yet because performance is typically multidimensional, this person may not be a "superstar" *overall* (i.e., across all relevant criteria). Therefore, it is important to understand the influences that measurement characteristics can have on the distribution of performance scores before attempting to identify "superstar" performers.

Another theoretical implication of this work is the importance of distinguishing between "performance" and "eminence." Specifically, eminence can be seen as maintaining some high level of performance over an extended period of time. For indicators of performance that accumulate (e.g., publications, election victories), the longer an individual is on the job, the more opportunities he or she has to perform, and subsequently, the more performance accumulates. To this point, in the samples included in this manuscript the average correlation between cumulative performance measures and opportunities to perform was $r = .91$. However, we are unable to draw firm conclusions about the causal nature of this correlation. Although it is necessarily true that opportunities to perform are causally related to cumulative performance (if opportunities to perform are removed, cumulative performance drops to zero), the opposite causal ordering is also probable. In other words, better performers are likely to be given more opportunities to perform than their lower performing counterparts. Yet in order for "performance" to become "eminence," the individual must maintain some level of performance over time. This opens the door to a variety of theoretically important questions. For instance, which performers are likely to maintain their skills over time? Which performers are likely to "burn out?" How can organizations select and/or develop employees who are likely to maintain a high level of performance over a sustained period of time? Our data indicate that unlike performance, eminence likely

follows a highly skewed distribution. Thus, when eminence is under investigation, the call to study "interesting outliers" (Aguinis et al., 2013) and "superstars" (O'Boyle & Aguinis, 2012) becomes highly relevant.

Practical Implications

Our results also have a number of practical implications, both for researchers interested in the study of job performance as well as organizations seeking to understand the performance of their employees. For one, our results have implications for the development and use of performance appraisal instruments. We argue against a general expectation that a highly skewed distribution will be obtained and caution against forced distribution approaches or rater training programs that force or encourage a skewed distribution of ratings. We also argue against an inference that a normal distribution of ratings signals rating error based on an expectation that a proper distribution will be skewed.

However, this is not to say that observed departures from normality are *necessarily* errors, as some groups may be comprised of truly exemplary performers. Conversely, an organization may only be interested in separating the "best" (or "worst") performers from the "rest." For example, an organization may decide that it is only interested in serious incidents of counterproductivity and that minor incidents will be overlooked. Similarly, a university may decide it will only reward top journal publication and that publication in lower tier journals will not be counted toward tenure and promotion. In this case, normally distributed performance ratings would *not* be expected.

With regard to research with job performance as the criterion of interest, we recommend that researchers carefully consider the characteristics of their job performance measures. For researchers seeking to generalize their findings to job performance in a broad range of settings, ideally the measures of job performance used in such research will meet our seven criteria, and such measures can be constructed *before* data are collected. However, we fully understand that practical constraints may make this impossible. For instance, researchers are often forced to work with the data collected by organizations. When this is the case, and when the data do not meet the criteria outlined above, we recommend caution in the interpretations researchers draw from their data. Specifically, researchers must be careful not to overinterpret highly skewed performance data as necessarily representative of the underlying construct, as the distribution is likely being driven at least to some extent by measurement characteristics. This is not to say that variables that do not meet our seven criteria are uninteresting, unimportant, or unworthy of study. For instance, it is

certainly legitimate for a researcher to focus on an individual behavior (e.g., absenteeism). Likewise, researchers may be interested in studying exemplary performance episodes (e.g., election victories) or exemplary performers (e.g., Academy Award winners). Furthermore, as stated above we believe "cumulative career performance" or "eminence" to be a highly valuable avenue for future research. However, our concern is generalizing from the distribution of variables that do not meet the seven criteria we have identified to the distribution of *typical* job performance.

Conclusion

In order to draw sound inferences about constructs from empirical data, it is first important to consider the definition and theoretical basis of the construct in question. Furthermore, special attention should be paid to the extent to which operational measures and sample characteristics are adequate for drawing inferences about the construct. In this manuscript we have argued that the vast departures from normality observed in O'Boyle and Aguinis' (2012) study likely say more about the measures included than they do about the nature of job performance.

REFERENCES

Aguinis H, Gottfredson RK, Joo H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*, 270–301.

Austin JT, Villanova P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology*, *77*, 836–874.

Bandura A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.

Barnes CM, Reb J, Ang D. (2012). More than just the mean: Moving to a dynamic view of performance-based compensation. *Journal of Applied Psychology*, *97*, 711–718.

Beck JW, Schmidt AM. (2012). Taken out of context? Cross-level effects of between-person self-efficacy and difficulty on the within-person relationship of self-efficacy with resource allocation and performance. *Organizational Behavior and Human Decision Processes*, *119*, 195–208.

Bowling NA, Burns GN. (2010). Scholarly productivity of academic SIOP members: What is typical and what is outstanding? *The Industrial-Organizational Psychologist*, *47*(4), 11–20.

Campbell JP, Dunnette MD, Arvey RD, Hellervik LV. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, *57*, 15–22.

Campbell JP, McCloy RA, Oppler SH, Sager CE. (1993). A theory of performance. In Schmitt N, Borman WC, & Associates (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco, CA: Jossey-Bass.

Combs J, Liu Y, Hall A, Ketchen D. (2006). How much do high-performance work practices matter? A meta-analysis of their effects on organizational performance. PERSONNEL PSYCHOLOGY, *59*, 501–528.

Cronbach LJ, Warrington WG. (1952). Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika*, *17*, 127–147.

Dalal RS, Lam H, Weiss HM, Welch ER, Hulin CL. (2009). A within-person approach to work behavior and performance: Concurrent and lagged citizenship-counterproductivity associations, and dynamic relationships with affect and overall job performance. *Academy of Management Journal*, *52*, 1051–1066.

Dunnette MD. (1963). A note on the criterion. *Journal of Applied Psychology*, *47*, 251–254.

Hartigan JA, Wigdor AK. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academies Press.

Jones J, Terris W. (1981, June). *Predictive validity of a dishonesty test that measures theft proneness*. Paper presented at the XVIII International Congress of Psychology, Santa Domingo, The Dominican Republic.

Motowidlo SJ, Kell HJ. (2013). Job performance. In Schmitt NW, Highhouse S, Weiner I (Eds.), *Handbook of psychology, volume 12: Industrial and organizational psychology* (2nd ed., pp. 82–103). Hoboken, NJ: Wiley.

O'Boyle E, Aguinis H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. PERSONNEL PSYCHOLOGY, *65*, 79–119.

Palisades Corporation. (2009). *@RISK 5.5: Risk analysis and simulation*. Ithaca, NY.

Podsakoff NP, Whiting SW, Podsakoff PM, Blume BD. (2009). Individual- and organizational-level consequences of organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology*, *94*, 122–141.

Rothbard NP, Wilk SL. (2011). Waking up on the right or wrong side of the bed: Start-of-workday mood, work events, employee affect, and performance. *Academy of Management Journal*, *54*, 959–980.

Rotundo M, Sackett PR. (2002). The relative importance of task, citizenship, and counterproductive performance for supervisor ratings of overall performance: A policy capturing study. *Journal of Applied Psychology*, *87*, 66–80.

Rotundo M, Sackett PR, Enns JR, Mann SL. (2012). Temporal changes in individual job performance: The role of reallocation of effort across performance dimensions. *Human Performance*, *25*, 1–14.

Sackett PR, Berry CM, Wiemann S, Laczo RM. (2006). Citizenship and counterproductive work behavior: Clarifying relationship between the two domains. *Human Performance*, *19*, 441–464.

Sackett PR, Zedeck S, Fogli L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, *73*, 482–486.

Tabachnick BG, Fidell LS. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn and Bacon.

USA Track and Field. (2012). *Recent & upcoming championships.* Retrieved from http://www.usatf.org/events/2012/OlympicTrials-Marathon/entry/entryMen/eligible.asp