Article

Organizational Research Methods
2014, Vol. 17(3) 324-345

DOI: 10.1177/1094428114526927
orm.sagepub.com

Selecting Null Distributions When Calculating r_{wg}: A Tutorial and Review

Rustin D. Meyer[1], Troy V. Mumford[2], Carla J. Burrus[1], Michael A. Campion[3], and Lawrence R. James[1]

Abstract

r_{wg} is a common metric used to quantify interrater agreement in the organizational sciences. Finn developed r_{wg} but based it on the assumption that raters' deviations from their true perceptions are influenced by random chance only. James, Demaree, and Wolf extended Finn's work by describing procedures to account for the additional influence of response biases. We demonstrate that organizational scientists have relied largely on Finn's procedures, at least in part because of a lack of specific guidance regarding the conditions under which various response biases might be present. In an effort to address this gap in the literature, we introduce the concept of *target-irrelevant, nonrandom forces* (those aspects of the research context that are likely to lead to response biases), then describe how the familiar "5Ws and an H" framework (i.e., who, what, when, where, why, and how) can be used to identify these biases a priori. It is our hope that this system will permit those who calculate r_{wg} to account for the effects of response biases in a manner that is simultaneously rigorous, consistent, and transparent.

Keywords

agreement, quantitative: multilevel research, aggregation, philosophy of science

Organizational scientists are often interested in determining whether rater observations are similar enough to justify aggregating them to represent a homogenous whole (Klein & Kozlowski, 2000). For example, when making inferences at levels of analysis that are superordinate to the individual (e.g., teams, organizational culture), researchers who collect individual-level data must first build a theoretical case for why a given phenomenon can meaningfully be conceptualized as a higher-order

[1]Georgia Institute of Technology, Atlanta, GA, USA
[2]Colorado State University, Fort Collins, CO, USA
[3]Krannert Graduate School of Management, Purdue University, West Lafayette, IN, USA

Corresponding Author:
Rustin Meyer, School of Psychology, Georgia Institute of Technology, 654 Cherry St. NW, Atlanta, GA 30332-0170, USA.
Email: rustin.meyer@psych.gatech.edu.

construct, then empirically demonstrate that raters' perceptions of it are sufficiently similar. This last step (i.e., demonstrating "interrater agreement" or "within-group agreement") is therefore "more than a statistical hurdle. It is an integral element in the definition of the group-level construct" (Klein, Conn, Smith, & Sorra, 2001, p. 4) because it affects the meaning of researchers' data, the logic of their conclusions, and the validity of subsequent inferences.

Fortunately, a number of statistics have been developed to help researchers assess interrater agreement (e.g., Brown & Hauenstein, 2005; Burke, Finkelstein, & Dusig, 1999; McGraw & Wong, 1996; Schmidt & Hunter, 1989). Among the most popular is r_{wg}, which quantifies the extent to which multiple judges' ratings are interchangeable due to their absolute (as opposed to rank-ordered) similarity. This goal is accomplished using the following equation, wherein the numerator represents the observed variability in judges' actual ratings and the denominator (i.e., the "null distribution") represents the variability in raters' deviations from their true perceptions that would theoretically be expected if no agreement existed (James, Demaree, & Wolf, 1984, 1993; LeBreton, James, & Lindell, 2005):

$$r_{wg} = 1 - (S^2/\sigma 2_E).$$

Because the specific value used in the denominator of this equation is left to the discretion of the researcher, those who calculate r_{wg} must seriously consider the form(s) that null agreement may take. The process of determining the shape of null agreement is therefore contingent upon researchers recognizing that specific types of systematic deviations from raters' true perceptions may be present. This situation creates a potential conflict of interest because those who may have a vested interest in empirically demonstrating agreement are also responsible for identifying relevant biases that, if present, would make statistical agreement less likely to be observed. These issues, coupled with the fact that there is little accumulated guidance regarding how one should go about selecting a null distribution, have led some to argue that "choosing the null distribution is the single greatest factor complicating the use of r_{wg}-based indices" (LeBreton & Senter, 2008, p. 829).

The Present Paper

In an effort to shed both prescriptive and descriptive light on the null distribution selection process, the present paper utilizes a two-study design to achieve several broad aims. Study 1 provides a tutorial of the concepts researchers should understand when selecting a null distribution as well as the specific steps they should take to ensure that this process is completed in a thorough manner. Central to these goals, we introduce the concept of *target-irrelevant, nonrandom forces* as a way to conceptualize those aspects of rating environments that are most likely to influence the shape of rater error distributions, then use the "5Ws and an H" framework (i.e., who, what, when, where, why, and how) to develop a process that researchers can use to identify the specific forces that are most likely to be present in a specific research context. Study 2 provides an empirical review of contemporary practices in order to examine the extent to which organizational scientists have relied on particular classes of null distributions and to assess the practical consequences thereof. It is our hope that this information will help researchers, reviewers, and editors navigate the null distribution selection process in a way that is simultaneously rigorous, consistent, and transparent.

Study 1: A Null Distribution Tutorial

The Meaning and Importance of Null Distributions

In the original article detailing the logic and calculation of r_{wg}, Finn (1970) proposed that a lack of agreement is best conceptualized as mathematically random deviations from raters' true perceptions,

represented by a null distribution wherein all response options (other than the one that best represents the rater's true perception) are equally likely to be endorsed. Such a pattern of error is known as a ''rectangular'' or ''uniform'' null distribution. Those who utilize this null distribution therefore implicitly assume that deviations from raters' true perceptions are caused exclusively by ''brief fluctuations in mood and motivation, momentary inattention, uncontrolled administration conditions (e.g., noise, distraction), illness, fatigue, emotional strain, or chance'' (James et al., 1984, p. 86) or ''momentary variations in attention, mental efficiency, distractions, and so forth within a given occasion'' (Schmidt, Le, & Ilies, 2003, p. 208). James et al. (1984) extended Finn's work by arguing that (a) response biases exist to the extent that raters systematically demonstrate patterns of responses that differ from their actual perceptions and (b) researchers should account for any such biases by utilizing null distributions that reflect the strength and direction of these effects.

Types of Null Distributions. Although there are technically an infinite number of ways in which null agreement can be operationalized when calculating r_{wg} using the procedures outlined by James et al. (1984), realistic options traditionally come in three broad categories (LeBreton et al., 2005; LeBreton & Senter, 2008). First, triangular distributions represent response patterns wherein raters are most likely to endorse the scale's midpoint when deviating from their actual perceptions and each of the more extreme options is progressively less likely to be endorsed. These distributions (of which the normal distribution is an example) represent the response pattern that would result from raters engaging in a central tendency response set (James et al., 1984), with the specific form of the triangle (i.e., ranging from extremely peaked to relatively flat) corresponding to the strength of the effect.

Second, a family of skewed distributions represents the null that would exist in the presence of either a leniency or severity bias, such that raters systematically gravitate toward either the positive or negative end of the response scale (respectively) when deviating from their actual perceptions. In these cases, the specific degree of the skew is a function of the strength of the bias. For example, subordinates who rate their superiors on some organizationally valued characteristic under non-anonymous conditions might deviate from their actual perceptions in a manner that generally favors the supervisor, thereby suggesting that a slightly skewed null distribution should be used, but might greatly deviate in a manner that favors the supervisor if asked to provide these ratings in a face-to-face meeting with him or her, thereby suggesting that a heavily skewed null distribution should be used to account for the presumably strong resulting bias.

As mentioned previously, the rectangular distribution (also known as the uniform distribution) represents the null that would exist if deviations from raters' actual perceptions occurred in a truly random fashion. This assumption is likely to be tenable only if raters are perfectly unbiased, act as random number generators, or their biases completely offset each other (Brown & Hauenstein, 2005). Given the numerous potential sources of systematic error in typical organizational research settings, some have argued that the conditions necessary to justify the use of the rectangular null ''will rarely (if ever) be fully met'' (LeBreton & Senter, 2008, p. 830).

Selecting Null Distributions

It is also important to point out, however, that all null response options require researchers to make certain assumptions about the shape of rater error. As such, those who calculate r_{wg} should view null distribution choice as one that requires a conceptually sound justification, as opposed to viewing the rectangular null as the standard default option. In those cases wherein response biases are present but the researcher utilizes the rectangular null distribution, agreement estimates will be artificially inflated because subsequent calculations conflate true variance (i.e., actual agreement) with variance that is best attributed to bias. As such, James et al. (1984) argued that the first step researchers should

take when calculating r_{wg} is to answer the following question: "If there is no true variance in the judgments and the true IRR [agreement] is zero, then what form of distribution would be *expected* to result from response bias, and, of course, some random measurement error?" (p. 90).

In order to answer this question, James et al. (1984) recommended that researchers "gather as much pertinent information regarding null distributions as possible, including empirical data designed to identify response bias for the judges in the sample at hand" (p. 94). It is important to explicate here, however, that the type of information to be gathered does *not* include the shape of the distribution of the concept upon which targets are being assessed (i.e., the focal construct). Although the shape of the focal construct is sometimes erroneously used to justify one's null distribution choice, the distribution of the focal construct should actually have no bearing on this decision. For example, researchers who are interested in assessing agreement in managers' ratings of CEO charisma may be tempted to conclude that they should use a skewed distribution as their null because individuals who become CEOs are often highly charismatic, but CEO charisma is unlikely to systematically influence raters' *deviations* from their actual perceptions, so any influence CEO charisma has on managers' ratings is best considered true variance, as opposed to error. Absent empirical evidence of actual response biases, however, little guidance exists regarding the type of information one should consider when making this judgment—we argue that a critical first step is appreciating the nature and effects of target-irrelevant, nonrandom forces.

Target-Irrelevant, Nonrandom Forces. Deviations from raters' true perceptions of a particular target (i.e., the object/person being rated) are influenced by random error under most (if not all) realistic conditions. Several social, psychological, political, and/or methodological considerations, however, also have the potential to systematically influence raters' deviations by making specific types of alternative response options more appealing than others—that is, to encourage response biases. We refer to such considerations as target-irrelevant, nonrandom forces because their presence influences raters' responses in a manner that is independent of the target's standing on the construct of interest, yet is also systematic in nature. The presence of target-irrelevant, nonrandom forces is therefore of substantive concern because these considerations lead to variability that is not attributable to either the rater's actual assessment of the target's standing on the focal construct or random error, thereby damaging the construct validity of r_{wg}-based agreement estimates that are calculated using the rectangular null (see Table 1 for a noncomprehensive list of these forces and a description of the ways in which they are likely to influence rater error).

Although target-irrelevant, nonrandom forces operate on the judgments of individuals, they ultimately influence the distribution of the ratings provided by *groups* of raters. This phenomenon occurs because target-irrelevant, nonrandom forces serve as strong situational cues that put "psychological pressure on the individual to engage in and/or refrain from particular courses of action" (Meyer, Dalal, & Hermida, 2010, p. 122). In the case of r_{wg}, raters are encouraged to endorse a specific type of alternative response option, which means that raters who are exposed to the same target-irrelevant, nonrandom forces will likely err in a relatively homogenous manner. Although appreciating the role of target-irrelevant, nonrandom forces is an important first step in selecting null distributions, identifying which forces are most likely to be present in a given research context is a substantial challenge due to the sheer number of potentially relevant considerations. Thus, one of the primary contributions of the present study is that we demonstrate how researchers can use the 5Ws and an H framework (who, what, when, where, why, and how) to address this issue in a simple yet comprehensive way.

The 5Ws and an H Framework

The impetus for drawing from the 5Ws and an H framework comes from Johns's (2006) recommendations for better understanding the role that context plays in the expression of organizational

Table 1. Explanation and Documentation of Potential Target-Irrelevant, NonRandom Forces That May Exist in Realistic Rating Contexts.

Force	Explanation	Relevant Citation(s)	5Ws and H Category	Resultant Null
Familiarity[a]	The "mere exposure effect" suggests raters will generally react more favorably to stimuli to which they have been previously exposed.	Bornstein (1989); Zajonc (1968)	Who, what	Skewed
Organizational stature[a]	Same-level raters are more likely than higher-level raters to allow irrelevant ratee characteristics to influence perceptions, thereby increasing the probability for halo error. Raters making assessments of higher-level targets (e.g., supervisors) are more likely to show a leniency response set. Raters from different levels are likely exposed to fundamentally different information.	Borman, White, and Dorsey (1995); LeBreton, Burgess, Kaiser, Atchley, and James (2003)	Who, what	Skewed
Cultural heterogeneity	Cultural norms regarding acceptable forms of criticism (for example) could inflate the level of agreement observed in ratings by encouraging raters to gravitate toward certain response options over others.	Barron and Sackett (2008)	Who, where	Skewed, triangular
Personal relationships[a]	Raters who have relationships with other raters and/or the ratee are more likely to use a "go along to get along" heuristic than to engage in independent, systematic processing.	Chen, Shechter, and Chaiken (1996)	Who, what	Skewed
Level of judgment[a]	Broad judgments are more likely than narrow judgments to be influenced by moderately relevant information.	Strack, Martin, and Schwarz (1988)	How	Skewed, triangular
Training/motivation[a]	Heuristic processing is most likely to occur when raters are distracted, using unclear instruments, and/or when making seemingly unimportant ratings.	Forgas and George (2001); Kulik and Perry (1994)	How	Skewed, triangular
Nature of constructs assessed	Individuals are more likely to select socially desirable response alternatives when self-reporting on affective variables such as anxiety.	Nunnally (1978); James, Demaree, and Wolf (1984)	What	Skewed, triangular
Time pressure[a]	Time pressure encourages a focus on information with a negative valence and increases anchoring effects.	Edland and Svenson (1993); Pennington and Roese (2003); Liberman, Molden, Idson, and Higgins (2001)	When	Skewed
Lab versus field[a]	Raters are more accurate in natural versus artificial settings, and in lab settings, the shape of error varies as a function of the nature of irrelevant environmental characteristics such as the dirtiness of the room.	Funder (1987); Kruglanski (1990); Swaan (1984); Teven and Comadena (1996)	Where	Skewed

(continued)

Table 1. (continued)

Force	Explanation	Relevant Citation(s)	5Ws and H Category	Resultant Null
Purpose of assessment[a]	Raters attempt to discern the reason why ratings are being collected so as to provide information of appropriate quantity, quality, and scope. For example, data collected to help make administrative decisions has been shown to be systematically more lenient than similar data collected for research purposes.	Bernardin and Orban (1990); Cleveland, Murphy, and Williams (1989); Dunning, Heath, and Suls (2004); Jawahar and Williams, (1997); Murphy, Cleveland, Skattebo, and Kinney (2004); Murphy, Jako, and Anhalt (1993); Schwarz (1999); Norenzayan and Schwarz (1999)	Why	Skewed
Anonymity[a]	Identifiable raters are more likely than anonymous raters to provide ratings that are inflated due to social desirability.	Joinson (1999); London, Smither, and Adsit (1997)	How	Skewed
Response formats	Participants are more likely to self-report high success in life when using –5 to +5 rating scale compared to when using a 0 to 10 scale, even when the wording of all questions and anchors are identical.	Schwarz, Knauper, Hippler, Noelle-Neumann, and Clark (1991)	How	Skewed
Item quality	Raters are more likely to endorse neutral response options when items are ambiguous or poorly written	Guilford (1954); Guion (1965); James et al. (1984)	How	Triangular
Rater personality	Raters who are highly agreeable and/or lack conscientiousness have been shown to be particularly lenient raters of others. Modest raters underevaluate their own standing on agentic traits such as intelligence, health, and sociability, whereas those with an independent self-construal tend to show the opposite effect.	Bernardin, Cooke, and Villanova (2000); Kurman (2001)	Who	Skewed
Emotional valence of stimuli	Ratings of affective stimuli tend to be influenced by automatic and intuitive processes, whereas ratings of non-affective stimuli tend to be influenced by reflective and deliberative processes.	Kahneman and Frederick (2001); Slovic, Finucane, Peters, and MacGregor (2007)	What	Skewed
Order of stimuli	Raters tend to overemphasize the importance of specific factors when making an overall judgment if the specific factor is rated before making the overall judgment	Kahneman, Krueger, Schkade, Schwarz, and Stone (2006); Strack et al. (1988)	How	Skewed
Cultural influences	Raters of East Asian descent (or more generally, those from institutionally collective cultures) tend to underestimate their standing on performance and perhaps other positively valenced characteristics.	Barron and Sackett (2008); Farh, Dobbins, and Cheng (1991); Korsgaard, Meglino, and Lester (2004)	Where	Skewed, triangular

[a]Target-irrelevant, nonrandom forces that were coded for in the published empirical studies reviewed in this paper.

behavior. Specifically, Johns argues that organizational context has the potential to affect the meaning and instantiation of various organizationally relevant phenomena in a variety of theoretically grounded ways. Most relevant to the present study, context has the ability to restrict the range of observable behaviors, affect the base rates of specific behaviors, and threaten validity by incentivizing certain forms of behaviors over others. These issues are pertinent to the shape of rater error because deviating from one's true perceptions in response to the presence of particular aspects of the rating context is a form of organizational behavior that can be better understood by analyzing the context in which it was instantiated. Thus, consistent with Johns's perspective, we posit that the 5Ws and an H framework can be meaningfully used to assess the likelihood that target-irrelevant, nonrandom forces from each of these broad categories are present in a given research context.

In the following subsections, we therefore explain the ways in which several specific forces that can be categorized into each of the 5Ws and an H are likely to affect the shape of rater error. It is important to note here, however, that although we claim that this *framework* is comprehensive, we do *not* claim that the specific examples provided within each category represent an exhaustive list. Indeed, we explicitly recognize that a comprehensive list of target-irrelevant, nonrandom forces and their corresponding response biases would be impossible to create. Our goal here is instead to highlight a few examples of target-irrelevant, nonrandom forces and explain how and why they engender specific biases in rating contexts that are pertinent to the organizational sciences. Table 1 contains additional target-irrelevant, nonrandom forces above and beyond those discussed here, but again, this table should not be viewed as comprehensive.

Who? The first consideration pertains to who is providing the ratings; that is, characteristics that (a) a preponderance of raters share and (b) are likely to systematically encourage particular types of deviations from one's internally held perspective, thereby invalidating Finn's assumption that all forms of deviations are equally likely to occur. For example, individuals who are particularly modest have been shown to underevaluate their own standing on agentic traits such as intelligence, health, and sociability, whereas those with an independent self-construal (i.e., one that emphasizes distinctness and separateness from others) tend to show the opposite effect (Kurman, 2001). Further, those who are highly agreeable and/or lack conscientiousness have been shown to be particularly lenient raters of others (Bernardin, Cooke, & Villanova, 2000).

With each of the previous examples, the end result is that researchers should use a skewed null distribution because either a leniency (for independent self-construal) or severity (for modesty) bias is likely to exist. Because it is impractical to thoroughly assess every rater's individual differences profile before calculating r_{wg}, however, we advocate that researchers simply ask themselves the following question: "Do a preponderance of the raters likely possess any traits that might systematically influence their judgments in a particular direction?" If this question is answered in the affirmative, researchers should then attempt to estimate the likely strength and direction of these effects (e.g., based on the nature of the bias and the proportion of raters who possess the trait in question) in order to help them select an appropriately skewed null distribution.

When applicable, a second category of *who*-based, target-irrelevant, nonrandom forces is raters' position in the organization relative to the target. For example, raters who have the same status as the targets they are rating have been shown to be more likely than higher-level raters to allow irrelevant information to influence their perceptions, which could lead to either a leniency or severity bias, depending on whether the irrelevant information is perceived positively or negatively by the rater. On the other hand, raters who provide assessments of higher-level targets (e.g., their supervisors) have been shown to be more likely to provide responses that are more lenient than their true perceptions (Borman, White, & Dorsey, 1995).

Further, raters from various levels of an organization might vary in the extent to which they are (a) exposed to the target in question, (b) attend to various characteristics of these targets, and/or (c)

assign different importance/meaning to the characteristics they attend to (LeBreton, Burgess, Kaiser, Atchley, & James, 2003). These different levels of exposure can influence the null distribution selection process in a variety of specific ways. For example, the mere exposure effect suggests that those who are exposed repeatedly to a stimulus will show more positive ratings than those who are exposed it less frequently (Bornstein, 1989; Zajonc, 1968), thereby suggesting that researchers should use a skewed distribution when calculating r_{wg} for those targets to which raters are frequently exposed (e.g., one's direct colleague), with the strength of the skew corresponding to their degree of familiarity. Again, this is not to say that all raters will rate those stimuli to which they are frequently exposed more favorably than novel ones but that in general, raters will tend to err on the side of leniency when rating a familiar target, so researchers should account for this tendency by selecting a skewed distribution that corresponds to the estimated strength of this bias. Additional who-based target-irrelevant, nonrandom forces can also be found in Table 1.

What? The question of *what* pertains to the nature of the targets being rated and how this nature may influence raters' deviations from their true perceptions. Perhaps the most important what-based consideration pertains to the emotional valence of stimuli, in that ratings of affective stimuli tend to be influenced by automatic and intuitive processes, whereas ratings of nonaffective stimuli (e.g., straightforward judgments of novel concepts) tend to be influenced by reflective and deliberative processes (Kahneman & Frederick, 2001). This distinction is relevant to the null distribution selection process because judgments made automatically are more likely to serve as proxy indicators of target-irrelevant characteristics, thereby skewing subsequent rater perceptions in a way that is concomitant with the strength and valence of evoked affect (Slovic, Finucane, Peters, & MacGregor, 2007). For example, raters who are asked to read a description of the financial fundamentals of a company whose logo elicits a positive emotional state may provide more favorable financial valuations than what a more rational assessment would yield, thereby making the rectangular null distribution a less theoretically viable option.

Although it has received surprisingly scant research attention, it might also be the case that raters are more likely to deviate from their actual perceptions in nonrandom ways when they assess people as compared to when they assess inanimate objects, ideas, and so on. Such a phenomenon makes theoretical sense based on the concepts of accountability and human targets' potential for eliciting affective reactions among raters. A potentially relevant example of this phenomenon comes from Harrison and McLaughlin (1993), who found that item context effects (i.e., systematic differences in responses to neutral items based on the content of the preceding items) were more likely among those items that referenced coworkers and supervisors and less likely among those items that assessed the work itself. That being said, this difference was not predicted a priori and has not been explored further to determine if it is evidence of a true effect or if it was spurious. Given the potential for fundamental differences between the psychology of rating people and the psychology of rating objects/ideas/concepts, however, such a line of future inquiry may prove to have direct implications for the null hypothesis selection process as well as the rating literature in general.

When? Although more esoteric than the other issues considered here, several effects that are relevant to the concept of time have been shown to influence human judgments in target-irrelevant, nonrandom ways. For example, research indicates that time pressure deteriorates cognitive control (Rothstein, 1986), encourages a focus on information with a negative valence, and increases anchoring effects (Edland & Svenson, 1993). Explanations of these phenomena revolve around the notion that time pressure creates a mindset wherein raters seek immediate closure, thereby putting them in a more prevention-focused mindset than they would experience given a more open-ended timeframe (Pennington & Roese, 2003). As such, raters who experience time pressure tend to seek information that will diminish the overall value of a target by fixating on its worst characteristics (Liberman,

Molden, Idson, & Higgins, 2001), thereby suggesting that a positively skewed null distribution should be used in the presence of time pressure, with the severity of this skew reflecting the amount of time pressure raters experience.

Another when-based consideration pertains to the order in which stimuli are presented. For example, the "focusing illusion" refers to the idea that raters tend to overemphasize the importance of specific factors when making an overall judgment if the specific factor is rated before making the overall judgment (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2006). In one published instance of this phenomenon, participants who rated their satisfaction with a specific aspect of their lives allowed this information to disproportionately affect judgments of their lives in general (Strack, Martin, & Schwarz, 1988). Specifically, the correlation between overall life satisfaction and raters' satisfaction with the number of dates they had in the previous month was not significantly different from zero when first asked about their satisfaction with their lives in general, then about their satisfaction with the number of dates they have recently had, but rose to .66 when the order was reversed. Thus, in those cases wherein r_{wg} is calculated on a general characteristic but specific judgments about a subordinate concept are collected first, observed variance is likely to be predictably skewed if raters' judgments of the specific issue trend in a particular direction (e.g., if employees are asked to rate their least preferred coworker before making judgments of their team as a whole).

Where? One issue to consider when addressing the question of *where* is whether rater responses were collected in an artificial or natural setting. This issue is potentially relevant to the null distribution selection process because raters have been shown to be more accurate in natural settings and less accurate in laboratory settings (Funder, 1987; Kruglanski, 1990; Swaan, 1984). It is important to point out, however, that "less accurate" is not synonymous with biased; it only becomes an issue to consider when selecting a null distribution when inaccurate judgments of a particular type/direction are systematically encouraged. For example, while in an aesthetically pleasing lab environment, individuals rated the personalities of targets who were not associated with the experiment more favorably than those who provided ratings of the same targets in a cluttered and dirty environment (Teven & Comadena, 1996), thereby suggesting a systematic (as opposed to random or simply inaccurate) effect.

Also, several culturally based effects are potentially relevant in various rating scenarios. For example, research suggests a modesty bias among raters of East Asian descent, such that when rating their own performance (and potentially other personal characteristics), raters from this region will, on average, rate themselves more harshly than their supervisors rate them (Farh, Dobbins, & Cheng, 1991). That being said, the universality of this effect has been questioned by some who suggest the presence of a general trend to overestimate one's own performance (Yu & Murphy, 1993), although this effect may be less pronounced among those of East Asian descent (Korsgaard, Meglino, & Lester, 2004). Further, some have argued that to the extent that this effect exists, it is best explained by differences in institutional collectivism, as opposed to differences in geography (Barron & Sackett, 2008), thereby suggesting this where-based consideration merely serves as a proxy for a deeper who-based consideration. Regardless of how one conceptualizes this effect though, a thorough application of the 5Ws and an H framework will permit researchers to identify it.

Why? Perhaps the most important question researchers should ask themselves when selecting a null distribution is "why are these ratings being collected?" (Jawahar & Williams, 1997; Murphy, Cleveland, Skattebo, & Kinney, 2004). The purpose for collecting ratings is particularly important in a typical rating context because humans tend to enter into information sharing settings with certain tacit assumptions about the types of responses that will be expected, based largely on the types of questions that are being asked (Schwarz, 1999). The overarching theme of these assumptions can

be summarized via the ''cooperativeness principle,'' which states that participants assume all parties involved will (a) make contributions that are relevant to the perceived theme/purpose of the discussion, (b) provide an appropriate level of detail in their responses, (c) communicate as clearly as possible, and (d) not conceal important circumstances regarding the intent of the discussion (Schwarz, 1999).

As such, an important first step when providing ratings is that participants attempt to discern the purpose of the discussion (i.e., determine why they are being asked the questions they are being asked) so that they can respond appropriately. This is not to say that respondents merely say what they think the researcher wants to hear, but instead suggests that the broader context surrounding why ratings are being provided will subtly shape the quantity, quality, and nature of the information provided (Bernardin & Orban, 1990). For example, participants who were asked to explain why a hypothetical murder occurred were more likely to focus on the murderer's personal characteristics when they were being questioned by ''The Institute for Personality Research,'' whereas they were more likely to focus on societal issues when being questioned by ''The Institute for Social Research'' (Norenzayan & Schwarz, 1999). Similarly, when raters know that their responses are likely to be used to make important organizational decisions (either for themselves or others), their subsequent judgments are likely to be more lenient than similar data collected explicitly for research purposes only (Cleveland, Murphy, & Williams, 1989; Murphy, Jako, & Anhalt, 1993).

Thus, potential why-based issues to consider include (but are not necessarily limited to) the extent to which raters can obtain or avoid important outcomes by providing judgments of one type or another, whether raters can reasonably be assumed to have any potential ulterior motives for providing ratings of one type or another, and whether there are any political considerations in the data collection setting that might influence raters' responses in a target-irrelevant, nonrandom way. Indeed, even notoriously error-prone self-assessments are substantially less self-serving (though not completely free of bias) when mundane information is collected for ostensibly unimportant purposes, but are predictably more biased when focused on important concepts collected for consequential reasons (Dunning, Heath, & Sulls, 2004). This line of research therefore suggests that r_{wg}-based agreement estimates based on perceptions collected for innocuous reasons should use a slightly skewed distribution, whereas those collected for consequential reasons should account for subsequent biases by using a moderately or heavily skewed distribution, depending on the importance of the consequences.

How? Perhaps the broadest issue to consider when selecting a null distribution is ''how are the ratings in question being collected?'' This issue has the potential to be particularly meaningful because it entails forming hypotheses about the ways in which specific methods that might ordinarily go unquestioned are likely to influence observed rating patterns in unintended ways. One of the most important how-based issues is whether data are collected anonymously or in an identifiable fashion. Not surprisingly, research suggests that socially desirable responses are more likely to occur when raters' specific responses can be paired with their identities, thereby suggesting that the most favorable portions of the response scale are likely to be over endorsed compared to identical judgments that are made anonymously (Joinson, 1999; London, Smither, & Adsit, 1997). Related to the issue of anonymity is whether or not the rater reasonably expects to have future interactions with the ratee. Although the extant literature does not specifically state that this consideration will affect rater response sets (e.g., by further increasing leniency), such a possibility represents not only the type of issue that researchers should consider (and discuss) when selecting a null distribution, but also an area of potentially fruitful future research.

Seemingly more innocuous characteristics of the data collection process such as response format and item quality have also been shown to systematically influence responses. For example, participants are more likely to self-report high success in life when using a –5 to +5 rating scale compared

to when they use a 0 to 10 scale, even when the wording of all questions and anchors is identical (Schwarz, Knauper, Hippler, Noelle-Neumann, & Clark, 1991). Consistent with the first point of the cooperativeness principle outlined previously, participants might incorrectly assume that even though the item in question only directly inquires about success, the presence of negatively scaled response options implies that the item's authors are also interested in failure. In this case then, participants make assumptions about the item's intent but in the process produce systematic deviations from their actual perceptions of the intended concept. Further, research also suggests that raters are likely to artificially gravitate toward the center of response scales when responding to ambiguous or poorly written items (Guilford, 1954; Guion, 1965; James et al., 1984), thereby suggesting that a triangular null distribution should be used when calculating r_{wg} using less than ideal scales (e.g., those that lack rigorous validity evidence).

Summary of the 5Ws and an H Framework. James et al. (1984) clearly explained that alternative null distributions should be used in the presence of response biases. For example, these authors stated that agreement estimates that are based on the rectangular null distribution will be inflated "if a common tendency exists among judges to select socially desirable response alternatives rather than response alternatives reflective of true judgments" (p. 86). Missing from the agreement literature, however, is a method of systematically identifying the substantive *causes* of these biases, thereby leaving authors to their own devices to identify relevant considerations. The 5Ws and an H framework is proposed here as an attempt to fill this void by providing a simple yet effective approach to identifying the target-irrelevant nonrandom forces that are most likely to be present under certain research conditions. Readers who are interested in exploring additional potential sources of bias and their likely effects on error distributions are referred to Table 1 and the literature cited therein.

Utilizing Resultant Evidence

Once the target-irrelevant, nonrandom forces that are most likely to be present in a given rating context have been identified using the 5Ws and an H framework, researchers should then estimate the most likely *net* effect of these influences. Here, four key questions must be addressed. First, what biases are most likely to result from the presence of each target-irrelevant, nonrandom force? Second, how strong is each bias likely to be? Third, how are these biases likely to combine to yield an ultimate effect (e.g., will they exacerbate each other or cancel each other out?)? Lastly, which type of null distribution best reflects the resulting effect? Once one or more null distributions have been identified as potential candidates, the corresponding amount error associated with this/these distribution(s) can be obtained in LeBreton and Senter (2008, pp. 832-833).

The previous content should not be interpreted to mean that researchers should rely exclusively on theoretical information about the likely shape of error. Indeed, where available, researchers should also draw from existing empirical evidence about the presence and strength of biases under similar circumstances. It is important to explicate here, however, that we do not believe that direct empirical evidence of bias in one's local context is a necessary precondition for selecting null distributions. In an ideal world, researchers would demonstrate the strength of the specific target-irrelevant, nonrandom forces that are operating, but such additional empirical efforts are (understandably) unlikely to be voluntarily enacted by researchers and unlikely to be insisted upon by journal editors.

Unfortunately, using the 5Ws and an H framework to identify target-irrelevant, nonrandom forces is unlikely to yield a single definite answer (i.e., one best error estimate). Instead, we argue that all available information should be used to identify the most optimistic and pessimistic null distributions, which readers, reviewers, and editors can critique for conceptual veracity. This perspective is consistent with that of LeBreton and Senter (2008) who advocated treating r_{wg} as a quasi–

confidence interval that is based on a thorough and rigorous assessment of the rating context. In those cases wherein only one potential bias is likely present, we recommend using its corresponding null as the most pessimistic option and the rectangular null as the most optimistic option; in those cases wherein no target-irrelevant, nonrandom forces are identified (or the biases that are present cancel each other out), we recommend using the rectangular null as the most optimistic option and a slightly skewed distribution as the most pessimistic option.

LeBreton et al. (2003, pp. 88-89) provide a particularly strong example of the type of evidence that should be used when selecting null distributions. Although these authors obviously did not use the term *target-irrelevant, nonrandom forces* and did not utilize the 5Ws and an H framework, they did dedicate several paragraphs to explaining the unique characteristics of their data collection context that were likely to systematically encourage one or more categories of alternative response options over others, discussed the likely resulting biases, and outlined the likely net results of these biases on the shape of rater error distributions. Specifically, these authors stated that because their data were collected for purposes of developmental feedback in a manner that made the rater unidentifiable by the target (i.e., the rater's manager), it was possible that raters would respond in an unbiased manner, but it was also possible that they would show a moderate leniency effect. As such, these authors ultimately calculated r_{wg} using rectangular, slightly skewed, and moderately skewed null distributions.

Study 1 Conclusion

We argue that those who calculate r_{wg} should use the 5Ws and an H framework to identify the target-irrelevant, nonrandom forces that are most likely to create biases in their specific research context, then use this information to create quasi–confidence intervals that represent r_{wg} under the most realistically optimistic and pessimistic considerations. To the extent that researchers put these recommendations into practice, the process of selecting null distributions when calculating r_{wg} should be better informed, more consistent, and more transparent, thereby leading to more valid decisions to aggregate. Although we suspect that those who calculate r_{wg} typically do not demonstrate this level of rigor when selecting a null distribution, it is important to examine contemporary practices to see what specific areas (if any) should be improved upon. Given that the tutorial provided here is the first source intended to provide practical guidance regarding how to go about selecting a null distribution, it would be unreasonable to hold the extant literature to a particularly high standard (e.g., by critiquing studies for not utilizing the 5Ws and an H framework). That being said, it is also likely that important insights can be gleaned from reviewing contemporary null distribution selection practices, which is the primary purpose of Study 2.

Study 2: A Review of Contemporary Null Distribution Selection Practices

Several researchers have suggested that instead of critically assessing the presence of potential response biases, researchers have tended to calculate r_{wg} using the rectangular null distribution in isolation (Brown & Hauenstein, 2005; LeBreton & Senter, 2008; Schmidt & DeShon, 2003). Given that using the rectangular null in the presence of response biases will overstate agreement estimates, any mismatch between associated assumptions and the reality of raters' deviations from their true perceptions has the potential to create what we call *meaningless agreement*. Meaningless agreement exists when obtained estimates meet a particular technical/mathematical standard (e.g., $r_{wg} \geq 70$), but they have been contaminated by one or more target-irrelevant nonrandom forces, which have influenced the observed variability in raters' responses (i.e., the numerator in Equation 1), but the

denominator assumes that randomness is the only force besides raters' actual perceptions affecting observed variance.

Given the critical importance of the assumptions underlying the use of the rectangular null distribution for both calculation and interpretation, a pervasive reliance on using it in isolation has the potential to represent an important "taken for granted assumption" (Hollenbeck, 2008, p. 19) because it highlights a conceptual disconnect between recommended best practices and standard operating procedures. Further, any such disconnect will lead to upwardly biased agreement estimates in those cases wherein target-irrelevant, nonrandom forces are present but unaccounted for in a specific research context. The following section outlines the methods we use to document contemporary null distribution selection practices, thereby empirically examining the extent to which researchers have or have not followed the recommendations of James et al. (1984), LeBreton and Senter (2008), and others who have supported moving beyond the assumptions of random error that underlie Finn's (1970) conceptualization of r_{wg}. Before describing our approach, however, it is first important to state that we expect that (a) null distribution choice will generally be underreported in the organizational science literature, (b) those studies wherein this decision is reported will heavily utilize the rectangular null, and (c) little relevant justification will be provided for this choice.

Method

Literature Search

In order to locate a sample of studies to help examine contemporary null distribution selection practices, we conducted a literature search using the Social Sciences Citation Index for all studies that cited any of the following seminal r_{wg} articles: James et al. (1984, 1993), LeBreton et al. (2005), or LeBreton and Senter (2008). We then narrowed our inclusion criteria further by focusing on those studies that were located in organizational science journals that routinely publish research in areas wherein r_{wg} is commonly used. Specifically, r_{wg} is often calculated in assessments of teams and leaders, which are especially common in journals such as *Journal of Small Group Research*, *Group and Organizational Management*, and *The Leadership Quarterly*. In addition, however, we also sampled from some of the field's more general journals in order to capture other less common uses of r_{wg}, such as assessing content validity in the instrument development process. Specifically, we also included (alphabetically): *Academy of Management Journal*, *Journal of Applied Psychology*, *Journal of Management*, *Journal of Organizational Behavior*, *Journal of Occupational and Organizational Psychology*, *Organizational Behavior and Human Decision Processes*, and *Personnel Psychology*.

As such, the present study's conclusions are relatively conservative for two reasons. First, they are based only on articles from sources wherein authors, reviewers, and editors are most likely to be familiar with the best practices surrounding the calculation of r_{wg}. Second, those studies that calculated r_{wg} but did not cite any seminal publications (and are therefore likely less familiar with contemporary recommendations) were not included here. These two factors combine to suggest that the present sample of studies represents an overly positive view of contemporary null distribution selection practices. Removing redundancies associated with empirical papers that cited multiple seminal sources yielded a total of 519 articles. One-quarter (130) of these studies were randomly selected for inclusion, a proportion that yields a hypothetical error rate of roughly $\pm 1\%$ (Bartlett, Kotrlik, & Higgins, 2001). Eighteen of these studies did not actually calculate r_{wg} and were therefore excluded from the data set, resulting in a total of 111 examined studies. Because r_{wg} is often used to assess agreement on multiple constructs within a single study, a total of 440 calculations were ultimately assessed.

Table 2. Descriptive Statistics for Those Target-Irrelevant, Nonrandom Forces That Comprise Each Calculation's Rectangular Probability Score.

Target-Irrelevant, Nonrandom Force	*M*	*SD*	% Present
Rater/ratee familiarity	.00	.00	.22
Organizational stature	.15	.35	14.7
Personal relationships	.00	.00	.22
Level of judgment	.03	.16	2.5
Training/motivation	.46	.50	46.3
Time pressure	.19	.39	18.9
Research setting (lab vs. field)	.09	.29	9.4
Purpose of assessment	.00	.00	.00
Anonymity	.01	.07	.45

Note: Target-irrelevant, nonrandom forces were coded such that scores of 0 suggest that the issue in question would likely not lead to a response bias, whereas scores of 1 suggest that the issue in question has the potential to lead to a response bias.

Coding Strategy

Studies were coded for two broad purposes: (a) description (i.e., recording each study's analytic/reporting practices) and (b) prescription (i.e., assessing the extent to which there were reasons to doubt the assumptions underlying the rectangular null distribution). Two separate coders assessed characteristics pertaining to each of these two considerations. In both cases, each coder was trained to use the codebook (described in the next two paragraphs), and an iterative process was used such that the trainer and the coder independently assessed a subset of studies and discussed subsequent discrepancies until a consensus was reached. After all discrepancies had been resolved, this process was repeated. The coder then finished the coding on her own, but made note of places of uncertainty, which were subsequently resolved through discussion with the trainer.

In terms of specific descriptive details, each study was coded for the journal in which it was published, its primary and secondary areas of research focus (e.g., leadership, teams, performance appraisal), the specific variable(s) for which r_{wg} was calculated, the specific r_{wg} index used, the number of items per scale, the number of scale response options, the reported r_{wg} value, and the ultimate aggregation decision. Subsequent analyses were conducted at the calculation level such that for those cases wherein r_{wg} was calculated on multiple variables, each use of r_{wg} was considered on its own (as opposed to, for example, taking an average of all reported r_{wg} scores).

In terms of specific prescriptive details, each study was also coded for the target-irrelevant, nonrandom forces starred in Table 1, which were used to create a "rectangular probability" score, wherein values closer to zero indicate that the rectangular null distribution was a potentially tenable option and values greater than zero indicated that the rectangular null distribution was less justifiable (see Table 2 for descriptive statistics associated with the components of rectangular probability score).[1] For example, a study that utilized a design wherein raters provided feedback about their supervisors in a non-anonymous fashion would receive a rectangular probability score of two, whereas a similar study wherein raters provided feedback about their coworkers in an anonymous fashion would receive a rectangular probability score of zero. The former value indicates the presence of two target-irrelevant nonrandom forces (i.e., a rater-ratee power distance, data provided non-anonymously), thereby suggesting that something other than the rectangular null distribution should be considered, whereas the latter suggests that the assumptions underlying the rectangular null distribution are potentially tenable. Unfortunately, however, it was not possible to code for all of the biases outlined in Table 1 due to reporting limitations.

Further, we also recorded which null distribution was used and the strength of the rationale provided for selecting this null. The strength of this rationale was assessed on a 4-point scale, wherein 0 = *no rationale was given* (i.e., the authors simply stated that they calculated r_{wg} without going into any additional detail), 1 = *weak rationale given* (i.e., the authors provided a superficial reference to previous theory and/or data), 2 = *moderate rationale given* (i.e., the authors discussed relevant theory and/or data in some depth—typically one or two sentences), 3 = *strong rationale given* (i.e., the authors discussed relevant theory and/or data in substantial detail—typically one or more brief paragraphs). Here, a concerted effort was made to account of the idea that rationales of various strengths could come in a variety of forms (e.g., theoretical justifications, assessments of past research), so strong versus weak rationales were distinguished based on the depth with which researchers made their arguments as well as the adequacy and meaningfulness thereof.

Because the only information needed to calculate r_{wg} is the observed variance and the expected variance, one can determine what r_{wg} *would* have been had the original researchers used a different null distribution. That is, one can substitute the variance associated with the null distribution that the original researchers used back into the r_{wg} equation in order to ascertain the observed variance on the variable(s) in question. This information can then be used in conjunction with the expected variance for various alternative nulls (based on those outlined in LeBreton & Senter, 2008) to calculate what r_{wg} would have been if one or more of these alternatives had been used. It is important to note here that the purpose of this review is not to criticize any single study, journal, or author but rather, to provide information about the corpus of studies that have calculated r_{wg}, the comprehensiveness of their null distribution selection process, and the practical effects (if any) of this choice on the ultimate decision to aggregate.

Results

Descriptive Analyses

Consistent with our first expectation, 75.9% of the calculations examined here did not report which null distribution was used. A conservative test of this prediction can be obtained by comparing this proportion to the value (i.e., 50%) that would be expected if reporting this decision occurred in a binary random manner. It is important to point out here that despite the fact that the nature of this significance test falls prey to the same underlying logic of utilizing a rectangular null distribution (i.e., assuming that a null pattern of results will occur in a purely random fashion), this assumption actually makes the present tests more conservative and those that use r_{wg} less conservative.[2] Results indicate support for our first prediction, in that the proportion of calculations wherein null distribution choice was not reported was significantly larger than that which would be expected due to chance factors alone, $\chi^2(1, N = 440) = 285.7$, $p < .001$.[3]

Consistent with our second expectation, of the 24.1% of r_{wg} values that were based on calculations wherein null distribution choice *was* reported (i.e., 106 calculations), 69.8% (i.e., 74 calculations) used a rectangular null. Again, a conservative test is provided by comparing the proportion of analyses wherein the rectangular null distribution was used to the proportion that would be expected if null distribution choice were a binary random decision (i.e., "rectangular" vs. "other"). Results indicate formal support for this expectation in that a significantly larger proportion of analyses used a rectangular null (i.e., 69.8%) compared to the proportion that would be expected based on chance factors alone (i.e., 16.67%), $\chi^2(1, N = 106) = 18.6$, $p < .001$.

Our third expectation was tested by examining the strength of the rationales used to justify researchers' choice of null distribution. Results indicate that the distribution of rationales for those analyses wherein null distribution was specifically reported was severely bimodal. Specifically, 53.8% of analyses were performed without reporting any rationale as to why the distribution in

question was used, 1.9% were based on a weak rationale, 3.8% were based on a moderate rationale, and 40.6% were based on a strong rationale. The proportion of analyses that were based on either no or a weak rationale was compared to the proportion that would be expected due to chance. Although this expectation was technically not supported, $\chi^2(1, N = 106) = .07, p = .79$, it is important to note that the strength of provided rationales was largely a function of which null distribution was selected. That is, those authors who chose a rectangular null distribution tended to provide less justification than those who used an alternative null distribution. Specifically, 19.2% of analyses that used a rectangular null provided a "moderate" or "strong" rationale, whereas 90.6% of analyses that used an alternative null provided this level of detail. Thus, it would appear that providing a strong justification for selecting one's null was not only the exception rather than the rule, but also an exception that was reserved primarily for those using something other than the rectangular null. That being said, the fact that more than two-fifths of estimates were based on a strong rationale is a finding that is both noteworthy and commendable.

Prescriptive Analyses

From a more prescriptive perspective, the present data are also able to help assess the proportion of analyses that utilized the rectangular null distribution in the presence of one or more target-irrelevant, nonrandom forces, thereby challenging the assumption that random error is the only factor influencing raters' deviations from their true perspectives. Although this issue is impossible to address with true certainty (i.e., not all forces that may influence the shape of null agreement can be known in advance nor would they be reported in every study), it is worth noting that more than two-thirds (i.e., 67.3%) of the 440 analyses assessed in this study contained at least one target-irrelevant, nonrandom force, thereby suggesting at least one reason why error might follow some nonrandom pattern (the mean rectangular probability score was equal to 1, with a standard deviation of .9). Furthermore, of those analyses wherein it was clear that the authors utilized the rectangular null distribution, the number of analyses that contained at least one target-irrelevant, nonrandom force approached three-quarters (i.e., 73.3%). It is important to note, however, that in all of these analyses, we gave the benefit of the doubt to the original authors' reporting practices (i.e., we assumed that when a given issue was not explicitly mentioned, the force in question was not present). Again, this assumption makes the present results a conservative test of our general research question.

The empirical data collected here were also able to help assess the proportion of analyses that would have come to a different aggregation decision if an alternative null distribution had been used. The analyses used to address this issue were conducted in two ways. First, initial conclusions were based on the assumption that the rectangular null was used in those analyses for which null distribution choice was not reported. In case this assumption is untenable, however, we also re-ran our analyses using *only* those calculations for which null distribution choice was actually reported. Under both of these approaches, however, analyses were only run on those calculations wherein evidence suggested that there was one or more reason to call into question the assumption of random error underlying the use of the rectangular null (i.e., a rectangular probability score greater than or equal to 1).

Results of the first analysis indicate that the proportion of r_{wg} values that were initially above the traditional .70 cutoff for aggregation but would have dropped below this threshold (thereby potentially reversing the author's original decision to aggregate) if an alternative null would have been used were: 33.8% (for a slightly skewed), 42.2% (triangular), 55.9% (moderately skewed), 51.5% (normal), and 80.4% (heavily skewed). As expected, results for the second analysis are similar in the sense that the proportion of r_{wg} values that would have dropped below .70 using only those analyses for which null distribution choice was reported was: 31.7% (slightly skewed), 39.0%

(triangular), 50.9% (moderately skewed), 56.0% (normal), and 83.0% (heavily skewed). It is important to note here, however, that our use of the traditional .70 cutoff does not constitute an endorsement of this standard as an absolute criterion for aggregation. Instead, it is used merely to demonstrate the general point that the common null distribution selection practice of relying almost exclusively on the rectangular null has the potential to affect one's decision to aggregate and therefore deleteriously impact the validity of resultant inferences.

Study 2 Discussion

The present review empirically supports the aforementioned possibility that the null distribution selection process represents an important "taken for granted assumption" (Hollenbeck, 2008, p. 19) in the calculation of r_{wg}. Specifically, at least three problematic trends exist in this literature; namely, researchers (a) too often fail to report which null distribution they utilized and (b) gravitate toward the rectangular null distribution without appropriately justifying this decision, even when (c) there may be reasons to believe its underlying assumptions are untenable. These practices are problematic because they (a) prevent readers from fully judging the construct validity of reported agreement estimates, (b) indicate that authors are choosing a simpler course of action instead of a more rigorous path that will help ensure valid agreement estimates, and (c) suggest that those who calculate r_{wg} do not have the tools and information necessary to decide which null distribution(s) to select (respectively). Further, additional analyses also suggest that more than 30% of decisions to aggregate may not have been empirically justified if one or more alternative null distributions had been selected. Thus, as opposed to simply representing less than ideal practices, we argue that the trends identified in Study 2 have the potential to inhibit our science by preventing readers and reviewers from adequately judging the validity of conclusions that are based on r_{wg}-based agreement estimates. The following section therefore summarizes how the content of the tutorial presented in Study 1 can be used in conjunction with the empirical findings from Study 2 to improve upon the present state of affairs.

General Discussion

Given the information outlined throughout this paper, we conclude that a lack of specific information, likely coupled with an incentive to utilize the rectangular null distribution, has led those who calculate r_{wg} to adopt less than ideal practices when selecting a null distribution. Specifically, researchers too often ignore this issue altogether, give it short shrift, or utilize the wrong type of information (e.g., focus on the shape of the focal construct's distribution as opposed to the shape of its associated error distribution). As such, one of the primary goals of the present treatment was to provide a tutorial that not only explicates the type of information that should be considered (i.e., target-irrelevant, nonrandom forces) when selecting null distributions but also presents a framework for doing so (i.e., the 5Ws and an H). As with all studies, however, the present treatment is not without limitations.

Limitations

The primary limitation of this study is that it is impossible to identify and explain all possible target-irrelevant, nonrandom forces that are likely to influence rater responses. This issue therefore adversely affected the tutorial portion of this study by curtailing available content. As such, the tutorial was instead designed to describe a *system* that enables researchers to assess a comprehensive list of broad categories of issues that are relevant to the null distribution selection process. This issue also limited the empirical portion of this study by making the rectangular probability

score (which we used to quantify the extent to which there were reasons to question the assumptions underlying the use of the rectangular null distribution) necessarily incomplete. Thus, rectangular probability scores were limited only to those target-irrelevant nonrandom forces that are indicated in Table 1 because published studies typically do not provide adequately rich information about the context in which data were collected to code for the other considerations listed therein.

Final Recommendations. Building upon the nature of the previous tutorial and subsequent empirical review, this section outlines several specific steps that researchers, readers, reviewers, and editors can take to ensure that the null distribution selection process is given its due diligence. First, it is critical that researchers report the null distribution that was used for every published r_{wg} calculation. Without this information, readers are unable to adequately judge the justifiability of a given decision to aggregate, thereby leaving open the question of a given phenomenon's ultimate viability as higher-order construct. Even if no empirical or theoretical rationale is provided for the null distribution that was used, it is a step in the right direction to at least specify one's decided upon course of action. As is evidenced by the empirical results of Study 2, however, such reporting practices are not yet commonplace.

Second, we recommend that researchers use the 5Ws and an H framework to report a summary of the target-irrelevant nonrandom forces that are likely present in their specific data collection context. In most cases, such a summary would only add a few sentences to the length of a manuscript. Although the particularly strong example outlined previously (i.e., LeBreton et al., 2003) utilized three brief paragraphs, we argue that using the system outlined here will permit more efficient explanations by providing researchers with a common framework, thereby ensuring that reviewers and editors are provided with adequate (and consistent) information to judge the validity of subsequent agreement estimates.

Consistent with James et al.'s (1984) original writing on the null distribution selection process, we agree that researchers should then "use this information to propose a small but inclusive *set* of null distributions that represent the major forms of anticipated response bias" (pp. 94-95), then calculate a range of r_{wg} values based on these alternative nulls. This recommendation is also consistent with that of LeBreton and Senter (2008) who likened this process to calculating "quasi–confidence intervals" (p. 837) that can be used to assess the justifiability of aggregation based on the range that emerges using multiple potentially viable nulls. In those cases wherein multiple theoretically justifiable null distributions exist, quasi–confidence intervals should be created based on the most optimistic and the most pessimistic options suggested by the 5Ws and an H framework; in those cases wherein no target-irrelevant, nonrandom forces are present (or the biases that are present cancel each other out), we recommend that authors use the rectangular null as their high-end estimate and a slightly skewed distribution as their low-end estimate.

Lastly, and consistent with others in this area (i.e., LeBreton & Senter, 2008, p. 830), we argue that the practice of *uncritically* using *any* null distribution in isolation should cease. Given that the findings of this study conservatively suggest that there is often at least one reason to believe that the rectangular null distribution may not be the only viable option, and that anywhere from one-third to more than three-quarters of aggregation decisions would have changed if an alternative null distribution would have been used, we see ample reason to utilize more caution when using r_{wg} to justify conclusions based on aggregated data. We acknowledge that the benefits of using the rectangular null (especially its ease of calculation and positive effects on agreement estimates) provide a strong incentive to utilize it in isolation, but it is ultimately incumbent on those who calculate r_{wg} to understand and critically vet the assumptions underlying its use because, in the words of the Canadian rock band *Rush*, "if you choose not to decide, you still have made a choice" (Peart, Lee, & Lifeson, 1980).

Declaration of Conflicting Interests

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

Funding

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

Notes

1. Although it would be ideal to code for every possible target-irrelevant, nonrandom force, two key issues prohibited this course of action. First, it is impossible to identify a priori every possible target-irrelevant, nonrandom force, so any list (including Table 1) will necessarily be incomplete. Second, standard methodological reporting practices limited our ability to code for all of the forces outlined in Table 1. As such, we elected to utilize those that were most likely to be regularly reported in published empirical studies.
2. Using what amounts of a rectangular distribution makes the present findings more conservative because this study compares the presence of the rectangular null to all possible alternatives simultaneously, which is a more stringent test than comparing the presence of the rectangular null to each potential alternative in a pair-wise fashion. In the case of r_{wg}, however, the rectangular null contains more variance than nearly all other alternatives, thereby leading to a spuriously inflated value in those cases wherein there are one or more reasons to question the validity of the assumptions underlying the use of the rectangular null distribution (James, Demaree, & Wolf, 1984).
3. Because data were often nested within study (i.e., many studies reported multiple r_{wg} values), some observations violated the assumption of independence. Thus, as opposed to using the standard (i.e., Pearson's) chi-square test, all analyses were conducted using McNemar's (1947) chi-square test, which accounts for violations of this assumption.

References

Barron, L. G., & Sackett, P. R. (2008). Asian variability in performance rating modesty and leniency bias. *Human Performance*, *21*, 277-290.

Bartlett, J. E., II, Kotrlik, J. W., & Higgins, C. C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, *19*, 43-50.

Bernardin, H. J., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology*, *85*, 232-236.

Bernardin, H. J., & Orban, J. A. (1990). Leniency effect as a function of rating format, purpose for appraisal, and rater individual differences. *Journal of Business and Psychology*, *5*, 197-211.

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology*, *80*, 168-177.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968-1987. *Psychological Bulletin*, *106*, 265-289.

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the r_{wg} indices. *Organizational Research Methods*, *8*, 165-184.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, *2*, 49-68.

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy- versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology*, *2*, 262-275.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, *74*, 130-135.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*, 69-106.

Edland, A., & Svenson, O. (1993). Judgment and decision making under time pressure: Studies and findings. In O. Svenson & J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 27-40). New York, NY: Plenum.

Farh, J., Dobbins, G. H., & Cheng, B. (1991). Cultural relativity in action: A comparison of self-ratings made by Chinese and U.S. workers. *Personnel Psychology*, *44*, 129-147.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, *30*, 71-76.

Forgas, J. P., & George, J. M. (2001). Affective influences on judgments and behavior in organizations: An information processing perspective. *Organizational Behavior and Human Decision Processes*, *86*, 3-34.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75-90.

Guilford, J. P. (1954). *Psychometric methods*. New York, NY: McGraw-Hill.

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

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. *Journal of Applied Psychology*, *78*, 129-140.

Hollenbeck, J. R. (2008). The role of editing in knowledge development: Consensus shifting and consensus creation. In Y. Baruch, A. M. Konrad, H. Aguinis, & W. H. Starbuck (Eds.), *Opening the black box of editorship* (pp. 16-26). New York, NY: Palgrave MacMillan.

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.

James, L. R., Demaree, R. G., & Wolf, G. (1993). r_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, *78*, 306-309.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, *50*, 905-925.

Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of Management Review*, *31*, 386-408.

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavioral Research Methods, Instruments and Computers*, *31*, 433-438.

Kahneman, D., & Frederick, S. (2001). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics of intuitive judgment: Extensions and applications*. New York, NY: Cambridge University Press.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., & Stone, A. A. (2006). Would you be happier if you were richer? *A focusing illusion.* Science, *312*, 1908-1910.

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of work environments. *Journal of Applied Psychology*, *86*, 3-16.

Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*, 211-236.

Korsgaard, M. A., Meglino, B. M., & Lester, S. W. (2004). The effect of other orientation on self-supervisor rating agreement. *Journal of Organizational Behavior*, *25*, 873-891.

Kruglanski, A. W. (1990). Motivations for judging and knowing: Implications for causal attribution. In E. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol. *2*, pp. 333-368). New York, NY: Guilford Press.

Kulik, C. T., & Perry, E. L. (1994). Heuristic processing in organizational judgments. In L. Heath, R. S. Tindale, J. Edwards, E. J. Posavac, F. B. Bryant, E. Henderson-King, Y. Suarez-Balcazar, & J. Meyers (Eds.), *Applications of heuristics and biases to social issues* (pp. 185-204). New York, NY: Plenum Press.

Kurman, J. (2001). Self-enhancement: Is it restricted to individualistic cultures? *Personality and Social Psychological Bulletin*, *27*, 1705-1716.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, *6*, 80-128.

LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding r_{wg}, r^*_{wg}, $r_{wg(j)}$, and $r^*_{wg(j)}$. *Organizational Research Methods*, *8*, 128-139.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815-852.

Liberman, N., Molden, D. C., Idson, L. C., & Higgins, E. T. (2001). Promotion and prevention focus on alternative hypotheses: Implications for attributional focus. *Journal of Personality and Social Psychology*, *80*, 5-18.

London, M., Smither, J., & Adsit, D. (1997). Accountability: The Achilles heel of multi-sources feedback. *Group and Organization Management*, *22*, 162-184.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30-46.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153-157.

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*, 121-140.

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology*, *89*, 158-164.

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, *78*, 218-225.

Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology*, *29*, 1011-1020.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Peart, N., Lee, G., & Lifeson, A. (1980). Freewill. On *Permanent Waves* [CD]. Morin Heights, Quebec: Le Studio.

Pennington, G. L., & Roese, N. J. (2003). Regulatory focus in temporal distance. *Journal of Experimental Social Psychology*, *39*, 563-576.

Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. *Organizational Behavior and Human Decision Processes*, *37*, 83-92.

Schmidt, A. M., & DeShon, R. P. (2003). *Problems in the use of r_{wg} for assessing interrater agreement*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

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.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*, 206-224.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93-105.

Schwarz, N., Knauper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570-582.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*. *177*, 1333-1352.

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, *18*, 429-422.

Swaan, W. B. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review*, *91*, 457-477.

Teven, J. J., & Comadena, M. E. (1996). The effects of office aesthetic quality on students' perceptions of teacher credibility and communicator style. *Communication Research Reports*, *13*, 101-108.

Yu, J., & Murphy, K. (1933). Modesty bias in self ratings of performance: A test of the cultural relativity hypothesis. *Personnel Psychology*, *46*, 357-363.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monographs*, *92*, 1-27.

Author Biographies

Rustin D. Meyer is an assistant professor of psychology at the Georgia Institute of Technology. He received his PhD in industrial-organizational psychology from Purdue University in 2009. His research interests focus broadly on the main and interactive effects of organizational context on work behaviors as well as humanitarian work psychology. His research has been published in outlets such as the *Journal of Management, Journal of Organizational Behavior, Journal of Business and Psychology,* and *Leadership Quarterly*.

Troy V. Mumford is an assistant professor in the Department of Management at Colorado State University. He received his PhD in organizational behavior and human resource management from the Krannert Graduate School of Management at Purdue University. He enjoys doing research exploring leadership at different organizational strata (e.g., supervisory, managerial, and executive), in teams (e.g., team roles, work synergies), and as drivers of organizational culture. His research has been published in such outlets as the *Journal of Applied Psychology, Personnel Psychology,* and *Leadership Quarterly*.

Carla J. Burrus is a graduate student in industrial-organizational psychology at the Georgia Institute of Technology. Her primary research interests include leadership, situational influences on behavior, and humanitarian work psychology.

Michael A. Campion is the Herman C. Krannert Professor of Management at Purdue University. He received his PhD in industrial and organizational psychology from North Carolina State University. His research interests include interviewing, selection, teams, work design, job analysis, testing, training, turnover, promotion, and motivation. He is the past president of the Society for Industrial and Organizational Psychology and past editor of *Personnel Psychology*.

Lawrence R. James is known for his work in psychological and organizational climate, statistics, and measurement of personality via conditional reasoning. He is the recent recipient of SIOP s Distinguished Scientific Contributions Award.