tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

Journal of Applied Psychology

0021-9010/16/$12.00 http://dx.doi.org/10.1037/apl0000099

Getting on the Same Page: The Effect of Normative Feedback Interventions on Structured Interview Ratings

AQ: au
AQ: 1

Christopher J. Hartwell
Utah State University

Michael A. Campion
Purdue University

This study explores normative feedback as a way to reduce rating errors and increase the reliability and validity of structured interview ratings. Based in control theory and social comparison theory, we propose a model of normative feedback interventions (NFIs) in the context of structured interviews and test our model using data from over 20,000 interviews conducted by more than 100 interviewers over a period of more than 4 years. Results indicate that lenient and severe interviewers reduced discrepancies between their ratings and the overall normative mean rating after receipt of normative feedback, though changes were greater for lenient interviewers. When various waves of feedback were presented in later NFIs, the combined normative mean rating over multiple time periods was more predictive of subsequent rating changes than the normative mean rating from the most recent time period. Mean within-interviewer rating variance, along with interrater agreement and interrater reliability, increased after the initial NFI, but results from later NFIs were more complex and revealed that feedback interventions may lose effectiveness over time. A second study using simulated data indicated that leniency and severity errors did not impact rating validity, but did affect which applicants were hired. We conclude that giving normative feedback to interviewers will aid in minimizing interviewer rating differences and enhance the reliability of structured interview ratings. We suggest that interviewer feedback might be considered as a potential new component of interview structure, though future research is needed before a definitive conclusion can be drawn.

Keywords: interviewer feedback, structured interview, control theory, normative feedback, interviewer differences

Regardless of industry, geography, or culture, the selection interview is perhaps the most widely utilized and preferred procedure for employee selection (Anderson & Witvliet, 2008; MacHatton, Van Dyke, & Steiner, 1997; Marcus, 2003; Moscoso & Salgado, 2004; Posthuma et al., 2014; Sanyal & Guvenli, 2005; Wilk & Cappelli, 2003). Much has been studied about interviews, including factors pertaining to the interview itself, applicant characteristics, and the decision-making processes (for recent reviews, see Posthuma, Morgeson, & Campion, 2002; Levashina, Hartwell, Morgeson, & Campion, 2014). Although many facets of the interview have received a great deal of research attention, the role of the individual interviewer seems to have been essentially forgotten on the research agenda (O'Brien & Rothstein, 2011; Pulakos, Schmitt, Whitney, & Smith, 1996; Schmitt, 1976).

There is a surprising dearth of research on the influence that feedback given to interviewers may have on their subsequent interview ratings. Feedback from mock interviews has been recommended as a component of interviewer training (Campion, Palmer, & Campion, 1997), but research has not given proper attention to feedback relating to actual interview ratings. This is surprising given that there has been so much research examining the role of feedback on job performance (e.g., Ilies & Judge, 2005; Kluger & DeNisi, 1996; Tolli & Schmidt, 2008), and that feedback serves as a central tenet of numerous motivational theories, including control theory (Campion & Lord, 1982; Carver & Scheier, 1981; Klein, 1989; Wiener, 1948). Thus, although performance feedback has been studied in various settings and situations, it has been largely ignored in the case of the interviewer.

The purpose of this research is to examine the effect of normative performance feedback given to interviewers regarding their actual structured interview ratings on their subsequent structured interview ratings. Using control theory and social comparison theory, we develop a regulatory model of normative feedback interventions (NFIs), which we test using multiple NFIs in an operational hiring setting. This research contributes to the literature on selection interviews by examining the effectiveness of normative feedback in combating interviewer leniency and severity and enhancing the psychometric properties of the structured interview. In addition to being the first of its kind, the longitudinal field-based nature of the study offers practical insight into how normative feedback might reduce individual differences in interviewer ratings, even after efforts have been made to reduce such differences by using other components of interview structure.

Individual Differences in Interview Ratings

Prior research has shown that there are individual differences in how interviewers rate applicants (e.g., Dougherty, Ebert, & Call-

Christopher J. Hartwell, Jon M. Huntsman School of Business, Utah State University; Michael A. Campion, Krannert School of Management, Purdue University.

Correspondence concerning this article should be addressed to Christopher J. Hartwell, Jon M. Huntsman School of Business, Utah State University, Logan, UT 84322–3555. E-mail: chris.hartwell@usu.edu

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

endar, 1986; Dreher, Ash, & Hancock, 1988; Huffcutt & Woehr, 1999; Mayfield, 1964; Mayfield, Brown, & Hamstra, 1980; Mullins, 1982; Van Iddekinge, Sager, Burnfield, & Heffner, 2006). These differences may arise as a result of errors that impact individual interviewers' ratings, including similarity bias (Graves & Karren, 1996; Harris, 1989; Zedeck, Tziner, & Middlestadt, 1983), confirmatory bias (Harris, 1989), contrast effects (Arvey & Campion, 1982; Schmitt, 1976; Zedeck et al., 1983), halo effects (Heneman, Schwab, Huett, & Ford, 1975; Zedeck et al., 1983), primacy and recency biases (Arvey & Campion, 1982), first impression bias (Arvey & Campion, 1982), and distributional errors (Pulakos et al., 1996). These differences pose challenges when assessing the validity of employment interviews because they introduce systematic error into the rating process and reduce reliability. Thus, researchers have contended that examining interview validity without considering individual differences may underestimate the validity of the interview (Arvey & Campion, 1982; Melchers, Lienhardt, von Aarburg, & Kleinmann, 2011; Van Iddekinge et al., 2006).

Most prior research on rating differences has focused on determining whether such differences exist and understanding their causes. Only minor attention has been paid to how they might be minimized. Two of the most widely recommended approaches to minimizing rating differences are structuring interviews and training interviewers (Arvey & Campion, 1982; Campion et al., 1997; Graves & Karren, 1996; Huffcutt & Woehr, 1999; Melchers et al., 2011; Pulakos et al., 1996), though interviewer training is sometimes considered part of interview structure (e.g., Campion et al., 1997). These methods have been shown to be effective in increasing the psychometric properties of the interview (e.g., Huffcutt & Woehr, 1999; Pulakos et al., 1996), but some individual rating differences remain (Huffcutt & Arthur, 1994; O'Brien & Rothstein, 2011; Schmidt & Zimmerman, 2004; Van Iddekinge et al., 2006). This leaves open the possibility that further reductions in individual rating differences might be made through additional mechanisms, such as providing performance feedback regarding actual interview ratings.

Distributional Errors

While we recognize that a variety of the previously described errors may exist in an interviewing context, this study focuses on interviewer distributional errors—specifically leniency and severity. These errors have received little attention in the interviewing literature, but have been found to be present in interviewing contexts (Heneman, 1975)—even when the interviews are structured (O'Brien & Rothstein, 2011). Distributional errors occur when a rater favors one section of the rating scale instead of utilizing a wide range of ratings. In an employee selection process, these rating errors can impact reliability and validity, particularly when different interviewers are involved in the hiring process. Job applicants interviewed by a lenient interviewer have an advantage over others, while applicants interviewed by a severe interviewer are disadvantaged (O'Brien & Rothstein, 2011). In addition, there is less differentiation between candidates if an interviewer relies on certain parts of the rating scale over others. It should be noted that we are referring to leniency *error*, in which ratings are inflated higher than what the behavior warrants. This is distinguishable from a leniency *effect*, which may occur when ratings tend to be higher than the midpoint of the rating scale, but may still validly represent behavior. For example, average interview scores may be higher than the rating scale midpoint when prior selection processes (such as résumé screening or hiring assessments) validly result in a high quality applicant pool.

AQ: 2

Although little research has examined leniency and severity in an interviewing context, prior research on performance appraisal may be instructive. In that context, leniency occurs when a supervisor tends to give inflated ratings, and it is often operationalized as a supervisor that gives above-average ratings relative to others (Murphy & Cleveland, 1995). Because of the tie between appraisals and salary decisions, as well as a desire for a positive continued relationship with subordinates and coworkers, leniency is a common issue (Bretz, Milkovich, & Read, 1992; Jawahar & Williams, 1997). It has been extensively studied in the context of supervisor ratings (e.g., Bol, 2011; Kane, Bernardin, Villanova, & Peyrefitte, 1995), self-ratings (e.g., Farh & Dobbins, 1989; Yu & Murphy, 1993), and 360-degree feedback ratings (e.g., Furnham & Stringfield, 1998; Ng, Koh, Ang, Kennedy, & Chan, 2011). In contrast, leniency is only briefly cited in the interviewing literature as a potential issue (Dreher et al., 1988; McIntyre, 1990; Zedeck et al., 1983), though it was found to exist in a study where judges rated mock videotaped interviews (Heneman et al., 1975), and another recent study found that it persisted even in a structured interview setting (O'Brien & Rothstein, 2011).

On the opposite end of the rating spectrum is severity, which occurs when raters tend to give lower ratings. Like leniency, severity can be manifest and/or operationalized by examining mean rater differences (Pulakos et al., 1996), where those who give below average ratings relative to others would be considered severe raters. We could find no empirical research specifically examining severity in the management literature, perhaps because it rarely occurs in performance appraisals, the context where distributional errors have been most often examined. However, there are theoretical arguments for why severity would exist in selection interviews. First, the risk of hiring a bad employee outweighs the risk of not hiring a good employee (Jagacinski, 1991, 1995), which could lead to conservative (severe) ratings of applicants (Motowidlo, 1986). In addition, other psychological research has pointed to an asymmetry effect between positive and negative information in which negative information is weighted more heavily (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). Early research demonstrated that this effect is present in simulated interview contexts, specifically finding that unfavorable information influences interviewer ratings more strongly than favorable information (Bolster & Springbett, 1961; Hollmann, 1972). Severity may also be more likely to occur in interviews, as compared with performance appraisals, because candidates are usually not given specific feedback on their interview ratings and because interviewers do not have an ongoing interpersonal relationship with applicants that would make them reluctant to assign low ratings.

A Regulatory Model of Interview NFIs

Normative feedback—also referred to as comparative feedback (e.g., Schmiege, Klein, & Bryan, 2010)—is information regarding individual past behavior that includes similar information about

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

referent others, allowing comparative inferences to be made. This normative information can be at the individual and/or group level.

The effect of normative feedback on individual intentions and motivation has been studied extensively. For example, studies have shown that normative feedback affects intentions and behaviors concerning individual safety (Klein, 1997), personal health (Klein, 1997; Schmiege, Klein, & Bryan, 2010), alcohol consumption (Neighbors, Lewis, Bergstrom, & Larimer, 2006), and recycling behaviors (Schultz, 1998). Within organizational contexts, normative feedback has been found to lessen absenteeism (Gaudine & Saks, 2001) and increase both task learning (Lewthwaite & Wulf, 2010) and self-perceptions of task ability (Klein, 1997). In addition, managers' turnover intentions and career satisfaction were both found to be influenced by comparisons of their individual accomplishments with those of others (Eddleston, 2009), and normative feedback from subordinates was found to have a more significant effect on managers' reactions than nonnormative feedback (Atwater & Brett, 2006). Perhaps most relevant to the current study, one unpublished dissertation (Davidson, 2003) examining managers' ratings in a performance appraisal context found that both lenient and severe raters who were given normative feedback significantly changed ratings in the direction of the overall group mean. However, NFIs have not always proven to be effective. In a review of the literature, Kluger and DeNisi (1996) found that feedback reduced future performance in a third of the cases, and the same authors suggested that normative feedback may lead to performance decline (DeNisi & Kluger, 2000). Moreover, NFIs typically show desired results only when perceived by the recipient as useful (Gaudine & Saks, 2001; Klein, 1997).

F1 In Figure 1, we draw upon control theory (Campion & Lord, 1982; Carver & Scheier, 1981) and social comparison theory (Festinger, 1954; Wood, 1989) to propose a model of how normative feedback can be useful in regulating interview ratings. Control theory posits that feedback regarding one's own behavior acts as a sensor signal, which can be compared with some sort of standard or referent. This is referred to in control theory frameworks as the comparator mechanism. When the sensor and the referent signals are compared and a discrepancy is found between the two, the individual is motivated to reduce that discrepancy by either changing individual behavior or changing the referent.

Integrating social comparison theory with control theory aids in understanding how normative feedback provides the recipient with both sensor and referent signals. In the current context, for example, the interviewers are each provided with normative feedback that includes his or her own mean interview ratings (the sensor signal in our model), as well as information regarding other individual interviewers' mean ratings and overall mean ratings across all interviewers (the referent signals in our model). Social comparison theory explains that individuals compare their own abilities to those of others, particularly when a meaningful true score is unknown (Festinger, 1954), being motivated through a desire for self-evaluation, self-improvement, and/or self-enhancement (Wood, 1989). Of these motivations, behavioral change based on job-related normative feedback is driven by self-improvement, in which the individual seeks to improve his or her abilities and performance. Self-improvement has been typically viewed in contexts where higher scores are always better scores, but in our model and context, this self-improvement motivation is focused on providing accurate ratings. Thus, self-improvement is not realized by achieving higher scores, but by reducing the discrepancy between the sensor and referent signals (Carver & Scheier, 1981).

Social comparison is included in our model (see Figure 1) as the specific type of comparator mechanism enacted as the interviewer compares his or her mean ratings with that of other interviewers and the overall referent group mean. Motivated by self-improvement to reduce discrepancies present in the social comparison, the individual decides on appropriate behavioral change. This behavioral change includes ratings changes that effect the individual's mean ratings and variance in interview ratings. True to the control theory framework (Campion & Lord, 1982; Carver & Scheier, 1981), there is a feedback loop present in our model. However, while the individual's behavioral change affects the individual's sensor signal in the subsequent NFI, the behavioral changes across interviewers also affect the referent signals in the subsequent NFI. Our model further proposes that behavioral changes across interviewers lead to increased agreement and reli-

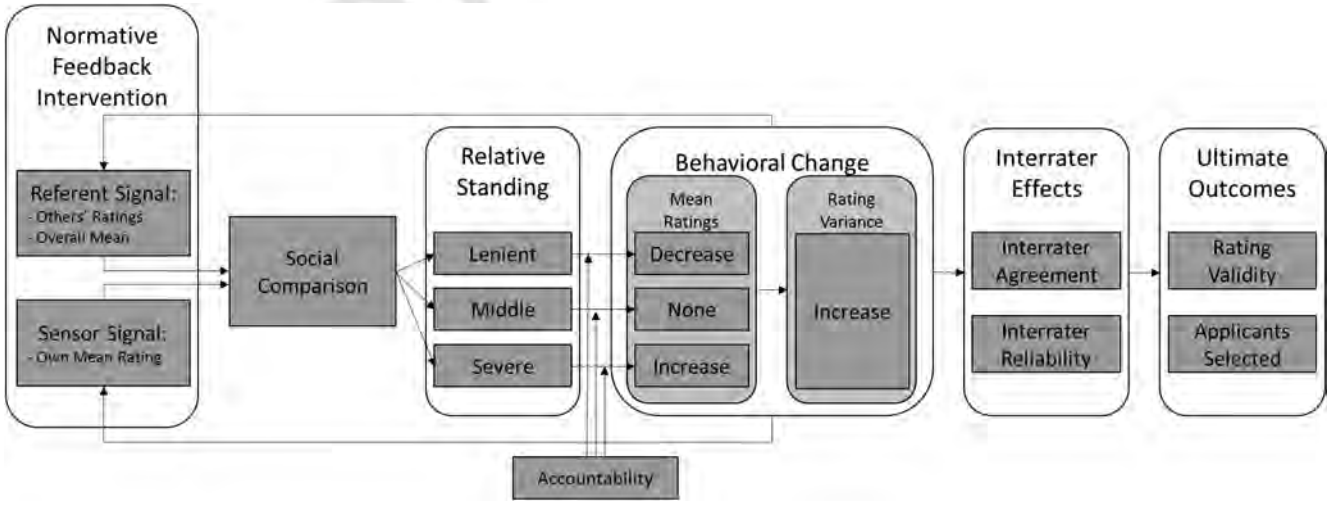

Figure 1. A regulatory model of normative feedback interventions and interviewer ratings.

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

ability between interviewers, and ultimately impacts interview validity and hiring decisions. Accountability, a factor in our model that moderates the behavioral change in response to social comparison, will be discussed in more detail later.

A number of the studies have utilized social comparison theory to explain how normative feedback may operate (e.g., Klein, 1997; Schmiege et al., 2010), such as in the context of peer review ratings (Mumford, 1983). The feedback provides a common frame of reference for an evaluation of personal competence based on a comparison with others, and encourages behavioral changes when there is a discrepancy between the individual and the group norms (Johnson, Turban, Pieper, & Ng, 1996). By integrating social comparison theory and control theory, our model sheds light on how the NFI provides the information on both the self and referent others (i.e., providing the sensor and the referent signals) that allows for social comparison (i.e., the comparator mechanism), and how a motivation for self-improvement drives self-regulation to improve rating accuracy.

Effects of Normative Feedback on Interviewer Mean Ratings

Our model proposes that normative feedback provides salient cues that trigger social comparison and leads to behavioral change for interviewers with comparatively high or low mean interview ratings. In the absence of true scores to compare with interview ratings, the normative feedback provides useful information that allows a comparison between interviewers, and between an individual interviewer and the overall mean across interviewers (acting as an aggregate true score). This aggregate true score provides a specific and meaningful referent signal that allows each interviewer to more accurately assess his or her output (average interview rating). Without this comparative information as a referent signal, individual average interview ratings would have little meaning. Given this comparative referent, however, interviewers feel a cognitive dissonance driven by a motivation for self-improvement when discrepancies are present (Bandura & Cervone, 1983; Festinger, 1957).

We predict that the overall mean interview rating across interviewers will act as an aggregate true score to which interviewers will compare their individual mean interview rating, and that the motivation for self-improvement will cause the specific behavior that results from the normative feedback to differ depending on the individual's relative standing. Specifically, those interviewers whose mean interview ratings are comparatively high (demonstrating relative leniency) or low (demonstrating relative severity) will be compelled to regulate their subsequent ratings to be more in line with the overall mean interview rating.

> *Hypothesis 1:* Normative feedback will affect interviewers' subsequent mean ratings, such that (a) lenient raters' mean ratings will decrease, (b) severe raters' mean ratings will increase, and (c) middle raters' mean ratings will not significantly change.

Normative feedback provides a salient cue for interviewers, but our model posits that an additional factor—that we label *accountability*—moderates the behavioral changes that occur as a result of social comparison, such that behavioral change is more likely when interviewers perceive that they are accountable for their ratings. Our concept of accountability refers to the perception of the individual regarding the importance of task performance in the overall context, as evidenced by contextual cues (e.g., incentives, consequences, centrality to the job, etc.). Prior control theory frameworks have hinted at the importance of accountability, but it has been directly integrated. For example, Campion and Lord (1982) mention that an individual's interpretation of the work environment impacts individual motivation, while Carver and Scheier (1985) explain that the degree of incentive associated with a task will impact individual efforts. However, by integrating social comparison theory with control theory, the importance of accountability is manifest more clearly. Social comparison theory (Festinger, 1954) explicitly acknowledges that the more important or relevant an ability is perceived by the individual, the more pressure the individual will feel to reduce the discrepancies concerning that ability. Thus, high accountability regarding the performance of a specific job task (e.g., interviewing and rating job applicants) will lead to high pressure to conform to performance expectations regarding the task.

It is important to note that the NFIs in our context were given for informational use; there were no direct consequences (bonuses, discipline, etc.) tied to them in our study. Thus, accountability was likely not aroused to the extent that it could be if such consequences were attached. However, interviewers were aware that their managers viewed their NFI data. In addition, given that decisions made during the interview process have long-term effects on the organization, and that such decisions are a central part of an interviewer's purpose and job identity, it is safe to assume that all interviewers feel a certain amount of accountability for their ratings (Wood & Bandura, 1989), and that the NFI makes that accountability more salient. We hypothesize, however, that while receipt of normative feedback should increase the perceived accountability of all raters, some will likely feel more accountability than others.

There are two major reasons that lenient interviewers (relative to other interviewers) will likely have higher accountability perceptions. First, given the high costs associated with hiring a low-performing employee, doing so is typically viewed as a more serious error than mistakenly rejecting a good candidate (Jagacinski, 1991, 1995; Motowidlo, 1986), and thus, accountability perceptions may be stronger for lenient raters than severe raters. Specifically, lenient raters may feel more accountability because leniency increases the chances of hiring applicants without the proper skills and abilities to perform the job well (Rowe, 1984), which would reflect poorly on the interviewer's ability and decrease overall organizational effectiveness. Conversely, a severe rater that rejects qualified candidates may feel less accountability, because it is typically not possible to know for sure how a rejected candidate would have performed.

Second, in a vast majority of structured interview field studies (including ours), there is a leniency effect in which the mean interview rating across interviewers is above the midpoint of the rating scale. We were able to locate 17 field studies from the past 20 years in which both rating scale ranges and overall mean interview ratings were reported. In 15 of those studies, the absolute mean interview rating was significantly higher than the midpoint of the scale (13 were significant at $p < .0001$).[1] Given this, most Fn1

[1] Full results and references available from the first author.

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

interviewers could be considered lenient in an absolute sense, such that their individual mean interview rating falls above the midpoint of the scale. While this leniency effect may be completely valid, interviewers will feel some accountability for being lenient in an absolute sense inasmuch as the midpoint of the scale is used as a comparative reference point. This would compound the perceived accountability for lenient raters (who are lenient in both a relative and absolute sense), while diminishing the perceived accountability for relatively severe raters (who are severe in a relative sense, but still lenient in an absolute sense). Even middle raters would feel some accountability for being lenient in an absolute sense. However, because ratings are based on anchored rating scales (discussed in more detail later) and because previous selection procedures typically weed out many unqualified candidates (resulting in fewer low interview ratings), the perceived accountability relative to the scale midpoint is likely to be much smaller than that relative to other interviewers.

Given these accountability effects, we expect that the change in subsequent interview ratings after the NFI will be most pronounced for lenient raters, though changes in severe interviewers' ratings will be greater than those raters that fall in the middle.[2] Fn2

Hypothesis 2: The magnitude of change in interview ratings after receipt of normative feedback will be (a) greater for lenient raters than for both severe raters and middle raters, and (b) greater for severe raters than for middle raters.

Effects of Normative Feedback on Interviewer Rating Variance

As noted, interviewers who exhibit leniency and severity tend to rely on, or disproportionally utilize, certain parts of the rating scale, while using other parts less frequently. Consistent with Kane's (1994) definition of nonvolitional systematic rating error, we view these tendencies as an unconscious flaw or deficiency in the process of observing, processing, and evaluating information. The NFI makes these unconscious tendencies more salient to the interviewer, and we propose that the effect of such feedback will do more than simply shift ratings toward the mean (creating a central tendency effect). In contrast, when distributional errors are reduced, we expect an increased range of ratings because raters will more often assign ratings from these underutilized parts of the rating scale while still using their previously favored parts of the rating scale, thus increasing the distribution of ratings. In other words, when a rater is made aware of his or her unconscious rating tendency, the rater is likely to recognize that tendency when making future ratings and make changes accordingly. When a rater's leniency is reduced, the rater will continue to use the upper part of the scale (though with more discretion), but will increase differentiation by recognizing their rating tendency and utilizing the lower ratings more often. Similarly, reducing a rater's severity will increase the use of higher ratings, while the rater continues to use lower ratings as well. Even interviewers in the middle, who do not significantly change their mean ratings, will likely see that others give higher and lower average ratings, feeling enabled or even compelled to broaden the range of future ratings. Thus, by minimizing individual rating differences though normative feedback, we expect all interviewers to exhibit a larger range in their interview ratings.

Hypothesis 3: Within-interviewer variance in ratings will increase after receipt of normative feedback.

Effects of Normative Feedback on Interrater Agreement and Reliability

As proposed in our model in Figure 1, normative feedback also may affect both the interrater agreement and reliability of interviewer ratings. Prior research on both interviewer training and performance appraisal has shown that frame-of-reference (FOR) training that utilizes feedback increases rating accuracy and interrater reliability (Bernardin & Buckley, 1981; Melchers et al., 2011). We expect that the NFI will highlight discrepancies between an individual's mean ratings and that of the referent group. Similar to FOR training, NFIs provide information that allows the establishment of a better frame of reference regarding the accuracy of ratings because each interviewer can frame his or her mean rating within the context of the ratings of other interviewers. The behavioral changes that result from the NFI will better align mean ratings and rating distributions across interviewers, increasing both the interrater agreement and interrater reliability (Weekley & Gier, 1989). In our context, interrater agreement is a test of whether the absolute values of ratings are comparable across interviewers, while interrater reliability is a test of whether the rank ordering of applicants is relatively consistent across interviewers (see LeBreton & Senter, 2008). We expect interviewers' mean ratings to more closely converge after feedback; thus, we anticipate increased interrater agreement. In addition, increased rating variance (as predicted above) and increased attention to others' interview ratings will create more covariation between interviewers, suggesting higher interrater reliability as interviewers utilize the rating scales more uniformly.

Hypothesis 4a: Interrater agreement will increase after receipt of normative feedback.

Hypothesis 4b: Interrater reliability will increase after receipt of normative feedback.

Relative Effects of Multiple Waves of Normative Feedback

Consistent with the discrepancy-reducing feedback loop of control theory (Campion & Lord, 1982), each subsequent NFI gives important information on the individual's progress toward reducing the discrepancy and is viewed by the individual as most relevant for making further behavioral changes. It can be expected that changes in interview ratings would be most closely tied with the most recent NFI. However, when multiple pieces of information are included in the NFI, the question arises as to the specific information used to make comparisons.

In our study, subsequent NFIs included not only mean ratings from the most recent time period, but also overall mean ratings combined over multiple years. Either of these two pieces of information could act as a cue for social comparison, but there is a

[2] As one anonymous reviewer pointed out, this hypothesis may not hold true in contexts where it is so difficult to find candidates that false negatives might be considered more detrimental. This is likely a boundary condition of our model.

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

question as to which would be more salient to the interviewers. The mean rating from the most recent time period may drive behavioral change because this feedback corresponds to the most recent behavior. The combined mean rating over multiple years may instead be most salient if it is viewed as a more reliable indicator of behavior over a longer period of time.

One of the conceptual underpinnings of control theory is that, over time, people organize their environment and feedback into patterns and schemas, to interpret future environmental stimulation (Carver & Scheier, 1982). While recent feedback is important, it is likely added to a composite of information to activate the comparator mechanism (Taylor, Fisher, & Ilgen, 1984). For example, research has found that repeated negative feedback significantly increased the likelihood of subsequent lowering of self-set goals (Campion & Lord, 1982). In another study, repeated negative feedback similarly increased the expectation of future failure (Zikmund-Fisher, 2004). This suggests that individuals utilize multiple past waves of feedback in activating the comparator mechanism and making behavioral changes. In our study, the combined mean ratings in subsequent NFIs represent a summary of all previous feedback, and social comparison is theorized to be driven by this overall representation of past performance. Therefore, we predict that this information will be more correlated with future behavioral changes than will the most recent feedback.

Hypothesis 5: Normative feedback based on the combined mean ratings over multiple time periods will be more strongly correlated with changes in future ratings than will feedback based on the most recent mean ratings.

Effects of Normative Feedback Over Multiple Time Periods

The novelty of the NFI the first time it is implemented will likely direct behavior toward improving task performance when a discrepancy exists. However, repeated negative feedback (where a discrepancy continues to persist) could turn attention away from the task and toward personal attributions of the self (Kluger & DeNisi, 1996). The regulatory effect of NFIs in our model would suggest that even positive feedback that shows reduced discrepancies would lead to smaller subsequent reactions to feedback. Therefore, we hypothesize that the incremental positive effects of the feedback—including changes in mean interview ratings, variation, agreement, and reliability—will decrease over time.

Hypothesis 6: The magnitude of changes in (a) mean interview ratings, (b) variance, (c) interrater agreement, and (d) interrater reliability will decrease over multiple NFIs.

Effects of Normative Feedback on Interview Validity and Applicant Selection

Because reliability creates an upper bound for validity (Conway, Jako, & Goodman, 1995), we expect higher interrater reliability (as expected in Hypothesis 4b) to lead to higher validity of the structured interview in predicting job performance, all else being equal. In other words, if the constructs being measured are held constant, the increased accuracy of the selection method will lead to higher rating validity. This is similar in principle to the increase in validity found by Bass and Avolio (1989) when partialing leniency scores out of correlations between transformational leadership styles and leadership effectiveness or satisfaction criteria. In addition, as discussed earlier, the selection of applicants is affected by distributional errors, such that those being rated by lenient interviewers generally receive higher ratings and are ultimately more likely to be selected than those being rated by severe interviewers. Thus, we expect that the decisions regarding which applicants are selected will be impacted by leniency and severity errors.

Hypothesis 7a: The validity of interview ratings in predicting job performance will increase when distributional errors are reduced.

Hypothesis 7b: Which applicants are selected will be significantly affected by leniency and severity errors.

Unfortunately, we were unable to obtain information from the organization regarding which applicants were ultimately hired, nor could we obtain subsequent job performance data, making Hypothesis 7a untestable using our data. Even in a case where such data were available, making inferences about changes in applicants selected (Hypothesis 7b) would not be possible, because applicants would not be the same at different time periods. Given these constraints, we developed a simulated data set (Study 2) to test these hypotheses statistically. To augment our simulation, we also reanalyze the data of two published data sets (Morgeson, Reider, & Campion, 2005; Van Iddekinge et al., 2006) in which interviewers conducted a high volume of interviews. By standardizing the ratings at the interviewer level, we expected that the validity of interview ratings would increase (in line with Hypothesis 7a). Thus, Study 2 allows a test of the final relationship in the regulatory model outlined in Figure 1: whether reducing interviewer-level error increases the validity of interview ratings and affects the ranking and ultimate selection of applicants.

Study 1: Method

Study Context

Data for this study were gathered as part of a large organization's operational interview process over a period exceeding four years, in which experienced interviewers rated applicants applying for professional positions in multiple career paths (e.g., management, economic analysis, and public relations). This setting is ideal for conducting our research for a variety of reasons. First, social comparison theory makes the assertion that the "importance of an ability . . . will increase the pressure toward reducing discrepancies concerning [it]" (Festinger, 1954, p. 130). The subjects of the current study are full-time interviewers, and providing accurate interview ratings is the main purpose of their jobs. Second, conducting the study in an operational selection setting with hiring consequences gives high fidelity and generalizability to the study. Finally, our interest is in the benefits of feedback above and beyond other known forms of interview structure. These interviews were structured following the 15 components of interview structure outlined by Campion et al. (1997). Specifically, (a) a job analysis acted as the basis for the interview, (b) the same questions were asked of all candidates, (c) very limited prompts and follow-up questions were allowed, (d) validated question types

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

were used, (e) a large number of questions were asked, (f) ancillary information about the candidate was limited, (g) no questions from the candidate were allowed, (h) each candidate response was rated individually, (i) multiple types of rating anchors were used (examples, descriptions, adjectives, etc.), (j) detailed notes were taken by interviewers, (k) two interviewers interviewed each candidate, (l) interviewers each conducted a high volume of interviews, (m) there was no comparison of candidates between interviews, (n) interviewers were highly trained, and (o) final interview ratings were determined as a unit-weighted average of the three interview components (discussed later in more detail).

A couple of the structural components deserve specific attention to understand how they are likely affected by the NFI. First, anchored rating scales, which provided behavioral examples, key words, and descriptions tied to points on the rating scale, were utilized to maximize standardization in interviewers' numerical ratings. Through 6 hr of annual interviewer training focused on rating videotaped candidates, interviewers were explicitly trained to focus on and reliably utilize the scale anchors when making their ratings. Attention to the rating anchors was the most central part of the interview training. The anchors on the 7-point scale corresponded to low (1–2), medium (3–5), and high (6–7) responses. The NFIs worked in conjunction with (not in place of) these rating anchors. In essence, interviewers were trained to use the anchors as common frames of reference (Bernardin & Buckley, 1981; Melchers et al., 2011), but there was discretion when assigning ratings to responses that fell in between anchors, or when applicant answers did not perfectly correspond with the anchors. Thus, the NFIs gave insight to interviewers, allowing them to better understand their overall tendencies to rate leniently or severely in these ambiguous cases. This awareness, in turn, should lead to more thoughtful consideration when assigning ratings, while still relying on the anchors as the major driver of the ratings.

The second component that warrants attention is the interviewer training. Interviewers in our study were late-career job incumbents and recent retirees, generally with 20–30 years of work experience in the organization. Together, they received 12 hr of mandatory training every September devoted specifically to interviewing. This training included 3 hr of lecture and discussion regarding the basics of personnel selection science and structured interviews (what they are, how they are developed, why they are important, etc.), along with 3 hr focusing on each of the three interview question types (situational, past-behavioral, and experience-based). The 3 hr for each question type entailed 1 hr of lecture and 2 hr of practice using anchored rating scales to rate videotaped interviews. The NFI was not emphasized in the training, though interviewers were told during training that they would receive annual normative feedback and that they should pay attention to it. As with the anchored ratings scales, the NFIs were designed to complement annual training, not replace it. The NFI provides practical feedback on actual interview ratings to utilize in conjunction with the training.

Normative Feedback Interventions

A feedback report was emailed by managers to each interviewer as a Microsoft Excel spreadsheet on an annual basis, with three reports (NFI1, NFI2, and NFI3) given over the course of this study (timeline described below). For each feedback report, each interviewer was assigned a random ID code, and this code was known only to the interviewer (the interviewers did not know each other's IDs) and management. This is an important factor in this study, as some theoretical research suggests that the use of normative scores may not be an effective means of social comparison when individuals can choose to compare themselves to others based on more salient factors, such as similarity, liking, and so forth (Wood, 1996). In this study, the normative feedback scores were the only information available, as all other factors were controlled by using confidential ID codes. Therefore, the interviewers were only able to compare themselves anonymously to the other interviewers. While we acknowledge that the possibility of interviewers comparing ratings still existed, the effort required to do so was substantially increased with the use of the confidential IDs. Instead of simply scanning all of the names on the NFI report to find all desired comparisons, interviewers would have had to mutually divulge information and compare ratings individually.

The feedback report given to each individual interviewer consisted of the average interviewer rating for each interviewer and the mean overall interview rating across all interviewers. Thus, interviewers were able to find their ID code and view their average interview rating for the prior time period, and compare that rating to that of their peers and the overall average across interviewers. In NFI2 and NFI3, in addition to information from the most recent rating period, similar information was included for all available prior time periods, along with overall mean ratings that combined ratings from all prior time periods. A sample NFI (representing NFI3) is provided in the Appendix.

Sample

There were 118 interviewers (54% men, 72% White) who completed over 20,000 interviews during a 4.5 year period. Of these, 62 provided ratings during the time period before the initial feedback report (Time 0), 45 also provided ratings in the time period before the second feedback report (Time 1), 36 provided ratings for three consecutive time periods (Time 0, Time 1, and Time 2), and 24 provided ratings over all time periods (Time 0–Time 3).

Measures

Time. Time 0 includes 18 months of interview ratings before NFI1, Time 1 includes interview ratings in the 12 months between NFI1 and NFI2, Time 2 includes interview ratings in the 12 months between NFI2 and NFI3, and Time 3 includes 10 months of interview ratings after NFI3. It should be noted that these time periods are not the same as those contained in the NFI reports (see Figure 2). There was a 5 month lag between when data were gathered for the NFIs and when they were distributed. This was partly before the time it took to receive the necessary raw data, conduct analyses, and distribute the feedback reports; part was because of the time period between assessment cycles (when very few interviews were conducted). The effect of this lag on our results is likely negligible, because the time periods discussed here (Time 0–Time 3) coincide with the point at which each NFI was distributed to the participants. In addition, the interviewers were given the NFI report just before the next assessment cycle, so they could utilize the feedback when making their subsequent ratings.

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

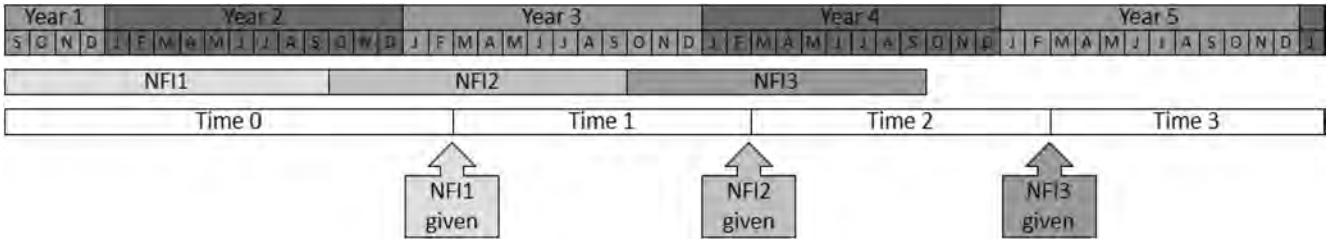

Figure 2. The timeline of the current study, highlighting the differences between the NFI periods (NFI1–NFI3) and the time periods in which we measure behavioral change (Time 0–Time 3).

Interview ratings. The structured interviews included three components: situational (nine items), past-behavioral (seven items), and experience-based (three items). Situational items asked applicants how they would respond to a hypothetical situation (Latham, Saari, Pursell, & Campion, 1980), past-behavioral items asked for specific examples of past behavior in a given type of situation (Janz, 1982), and experience-based items were more biographical in nature—such as asking applicants about prior work experience and education. Each of these components tapped different job-related competencies. Calculating component scores was accomplished by taking the mean of all component items. The *SD*s of the component scores were similar, so the overall interview rating was calculated by taking the mean of the situational, past-behavioral, and experience-based component scores, to achieve approximately equal weighting of the components.[3] Fn3

Analyses

Neither interviewer race nor gender were significantly related to interview ratings or NFI ratings at any time period, nor were there significant differences in race, gender, or average interview ratings between interviewers with complete data at all four time periods and those with missing data. Therefore, we do not consider these demographic variables further in our analyses. Power analyses (Cohen, 1988) indicated sufficient power (60–70% or more) to detect medium and large effects for most of our hypotheses. We have taken into consideration the instances in which low power may have impacted our results (specifically Hypothesis 2 and Hypothesis 6a and 6b); more details are found in the results for those specific hypotheses.

For a majority of our analyses, we used random coefficient modeling (RCM), following the steps outlined in Bliese and Ployhart (2002) and utilizing the Multilevel package of the R statistical program (Bliese, 2013).[4] Fn4 In the context of our study, the first step of this process (Step 1) was to calculate the intraclass correlation coefficient (ICC1) for the criterion variable (e.g., mean interview ratings) to ensure that significant between-subjects variability existed. Step 2 involved estimating the relationship between time and the criterion variable (i.e., does the overall mean of the criterion change over time?). In analyses that included more than two time periods, this step required testing both linear and nonlinear relationships (i.e., is the rate of change over time constant or does it vary?). Step 3 required analysis of whether significant mean differences (e.g., between-rater differences in initial mean ratings) and differences in change over time (e.g., between-rater changes in mean ratings after the NFI) were found. Step 4 (conducted only for analyses that extend beyond two time periods) examined the error structure of the criterion variable (autocorrelation and heteroskedasticity) to arrive at the final Level 1 model. Throughout these steps, log likelihood ratios were calculated to determine the best-fitting model.

Once the final Level 1 model was obtained, Step 5 added a Level 2 variable (NFI ratings) to examine between-subjects differences in intercepts. Finally, Step 6 tested for cross-level interactions, using the Level 2 predictor variable to examine between-subjects differences in slope over time (i.e., is the change over time in the criterion related to the Level 2 predictor?). This step provided a final Level 2 model and allowed us to test many of our hypotheses.

Study 1: Results

Descriptive statistics and intercorrelations of all interviewer-level variables are found in Table 1. T1

Time 0 Versus Time 1 (Hypotheses 1–4)

The first round of analyses examined the difference in interviewers' average ratings between Time 0 and Time 1 as a result of NFI1. Forty-five interviewers completed interview ratings for both time periods and received feedback during NFI1. These interviewers completed an average of 73 interviews during Time 0 (M = 72.96, SD = 53.43) and 110 interviews during Time 1 (M = 110.20, SD = 83.82), with a total of 8,242 interviews during these two time periods.

Hypothesis 1. Our first hypothesis predicted that we would see (a) a decrease in average interview ratings for the lenient raters, (b) an increase in ratings for severe raters, and (c) no significant change in ratings for the average raters as a result of the NFI. The ICC(1) calculated in Step 1 of the RCM analysis (.76) showed a high level of between-subjects compared to within-subject variability in mean interview ratings. Results of Step 2 indicated that the overall average interview rating did not change significantly between Time 0 and Time 1 ($t(45) = -1.14$), providing initial support for Hypothesis 1c. The model in Step 3, which allowed for variability in slopes, fit the data better than the model in step two

[3] We also conducted analyses using the interview components, and the results were generally similar to those reported for the overall interview ratings. Results of these analyses are available from the first author.

[4] Completing the analyses in SPSS using the steps outlined in Peugh and Enders (2005) revealed virtually identical results.

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

T2

T3

F3

Table 1
Descriptive Statistics and Intercorrelations of Variables at the Interviewer-Level of Analysis

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
Mean rating															
1. Time 0	66	5.34	.18	—											
2. Time 1	67	5.34	.12	.72* (45)	—										
3. Time 2	73	5.26	.13	.57* (37)	.71* (54)	—									
4. Time 3	59	5.24	.13	−.15 (26)	.15 (28)	.50* (46)	—								
Mean variance															
5. Time 0	66	.19	.09	−.42* (66)	−.14 (46)	.09 (37)	−.38 (26)	—							
6. Time 1	67	.22	.07	−.08 (46)	−.21 (67)	.09 (54)	−.04 (28)	.55* (46)	—						
7. Time 2	73	.20	.07	−.24 (37)	−.32* (54)	−.26* (73)	.24 (46)	.06 (37)	.22 (54)	—					
8. Time 3	59	.16	.07	.48* (26)	.26 (28)	.10 (46)	−.18 (59)	.37 (26)	.19 (28)	.21 (46)	—				
NFI feedback															
9. NFI1	62	5.33	.22	.94* (62)	.70* (45)	.68* (36)	−.06 (25)	−.32* (62)	−.08 (45)	−.15 (36)	.48* (25)	—			
10. NFI2: Recent	63	5.30	.15	.54* (41)	.79* (63)	.65* (52)	.04 (27)	−.07 (42)	−.28* (63)	−.16 (52)	.21 (27)	.45* (41)	—		
11. NFI2: Combined[a]	41	5.25	.16	.87* (41)	.87* (41)	.71* (32)	−.32 (22)	−.10 (41)	−.25 (41)	−.31 (32)	.45* (22)	.83* (41)	.76* (41)	—	
12. NFI3: Recent	57	5.21	.15	.38* (33)	.47* (38)	.82* (57)	.65* (46)	.07 (33)	.04 (38)	.03 (57)	.06 (46)	.49* (32)	.49* (36)	.46* (28)	—
13. NFI3: Combined[a]	40	5.22	.14	.79* (33)	.87* (38)	.74* (40)	.02 (30)	.10 (33)	−.12 (38)	−.16 (40)	.38* (30)	.84* (32)	.85* (36)	.97* (40)	.59* (40)

Note. Parentheses refer to the *n* for the specific bivariate relationship. NFI = normative feedback intervention.
[a] Combined ratings include only interviewers with ratings at both the most recent time period and at least one prior time period.
* $p < .05$.

($\chi^2_{diff}(2) = 11.62$, $p < .01$), indicating that changes in mean ratings from Time 0 to Time 1 varied across individuals. This model was retained as our final Level 1 model (see Table 2).

The model in Step 5 showed that centered mean interview ratings at NFI1 were related to initial mean interview ratings at Time 0, $t(60) = 20.54$, $p < .001$. Because ratings used in NFI1 are a large part of the ratings that comprised Time 0 (see Figure 2), this was to be expected. The final Level 2 model created in Step 6 (see Table 2) indicated a significant interaction between time and the mean ratings at NFI1, $t(43) = -3.94$, $p < .001$, indicating that ratings at NFI1 affected how interviewers' ratings changed from Time 0 to Time 1. We plotted this interaction by using the final Level 2 model parameters and the *SD* of mean ratings at NFI1 (.22) to predict Time 1 interview ratings for middle (mean), lenient (1 *SD*), and severe (−1 *SD*) raters (Bliese & Ployhart, 2002; Cohen, Cohen, West, & Aiken, 2003). As Figure 3 shows, lenient raters reduced their ratings by .08 from Time 0 to Time 1 (5.52 to 5.44), while severe raters increased their ratings by .03 over the same time period (5.20 to 5.23). Middle ratings decreased by .02 (5.36 to 5.34). Table 3 illustrates that using ratings on NFI1 in our final Level 2 model explains 19% of the variance in individual rating changes from Time 0 to Time 1. These findings provide strong initial support for Hypothesis 1a, though the results are not as strong for Hypotheses 1b.

To determine whether rating changes were more than simple regression to the mean, we mean-centered the average interview ratings at Time 0 and used ±1 *SD* (.18) to represent our lenient and severe raters at Time 0. Similar to Smither et al. (1995), we used the correlation between interview ratings at Time 0 and Time 1 ($r = .72$) as a conservative estimate of reliability in a bivariate regression equation (i.e., Time1 = −.02 + [.72][Time0]), where −.02 is the change in overall mean ratings between Time 0 and Time 1 (see the linear trend in Table 2).

For severe raters, the equation was as follows: $\text{Time1}_{SEV} = -.02 + (.72)(-.18)$. The resulting Time1_{SEV} value (−.15) was then added to the estimated mean rating at Time 0 (5.36; see the Level 2 intercept in Table 2) to give us 5.21, the expected mean rating for the severe raters at Time 1 because of regression. The actual rating increase of severe ratings was from 5.20 to 5.23, three times what would be expected solely by regression to the mean (5.20 to 5.21), indicating that the increase could not be explained by regression to the mean. A similar equation for lenient ratings ($\text{Time1}_{LEN} = -.02 + [.72][.18]$) resulted in an expected decrease in the average lenient interview rating of .05 from 5.52 to 5.47. The observed decrease of .08 to 5.44 is 60% more than would be expected from regression effects. Therefore, accounting for regression to the mean could not explain the change in either lenient or severe ratings from Time 0 to Time 1, further supporting Hypothesis 1.

Hypothesis 2. Our second hypothesis predicted the magnitude of change as a result of the NFI would be (a) greater for lenient raters when compared with both severe and middle raters, and (b) greater for severe raters than middle raters. As there exists no test of significance to compare slopes computed at different points along a continuum (Cohen et al., 2003), we took an alternative approach to test this hypothesis. We split the 62 interviewers with feedback at NFI1 into three relatively equal groups. The lenient interviewers ($n = 21$) were operationalized as those with average interview scores in the top third and severe interviewers ($n = 21$) were those with average interview scores in the bottom third. The

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

Table 2
Random Coefficient Models Predicting Mean Interview Rating at Time 1

Model and parameter	Estimate	*SE*	*df*	95% CI (lower)	95% CI (upper)	*t*
Final Level 1 model (variable slopes)						
Intercept	5.344	.020	86	5.305	5.383	263.55***
Linear trend of time	−.011	.016	45	−.042	.020	−.69
Final Level 2 model (using mean ratings at NFI1 to predict rating changes)						
Intercept	5.357	.008	60	5.327	5.359	667.65***
Linear trend of time	−.022	.014	43	−.043	.011	−1.558
Mean rating at NFI1	.730	.035	60	.661	.799	20.87***
Linear Trend × Mean rating at NFI1	−.267	.068	43	−.402	−.132	−3.94***

Note. CI = Confidence interval; NFI = normative feedback intervention.
*** $p < .001$.

middle third ($n = 20$) were included as a group that was neither lenient nor severe. Of these 62 total interviewers, 45 had data available during both time periods (Time 0 and Time 1), including lenient ($n = 18$), middle ($n = 15$), and severe ($n = 12$) raters. Because of the reduction in power associated with making this necessary split, we interpret significance in findings of $p < .10$ for this hypothesis.

For the raters with data at both time periods, we first calculated the raw value of the change between Time 0 and Time 1 interview ratings as a new variable. While change scores have been criticized in some circumstances (Edwards, 1994; Edwards & Van Harrison, 1993), the use of these scores is appropriate and necessary in within-subjects research where there is an expected Participant × Treatment interaction (Hogan, Barrett, & Hogan, 2007; McFarland & Ryan, 2000), which we predicted and subsequently found in Hypothesis 1. After computing these change scores, we calculated the mean and *SD* of this new variable for each group. Finally, we used the absolute value of each group's mean change, along with the *SD* and *n* of each group, in a series of independent sample *t* tests.

The results indicated that the magnitude of change in lenient raters was significantly larger than the change in middle raters, $t(31) = 2.73$, $p < .05$ and marginally greater than the change in severe raters, $t(28) = 1.73$, $p < .10$, providing support for Hypothesis 2a. The magnitude of change between severe and middle raters was not significant, $t(25) = .69$, which does not support Hypothesis 2b.

Hypothesis 3. The next hypothesis predicted that within-interviewer variance would significantly increase as a result of the NFI. We conducted RCM analyses, using mean variance as the criterion variable, and the results are found in Table 4. Steps 1 and 2 indicated between-interviewer differences in interviewers' rating variance, and that the overall average rating variance significantly increased by .029 from Time 0 and Time 1, $t(45) = 3.04$, $p < .01$, supporting Hypothesis 3. Modeling variability in slopes in Step 3 did not fit the data better ($\chi^2_{diff}(2) = .02$), indicating that interview-

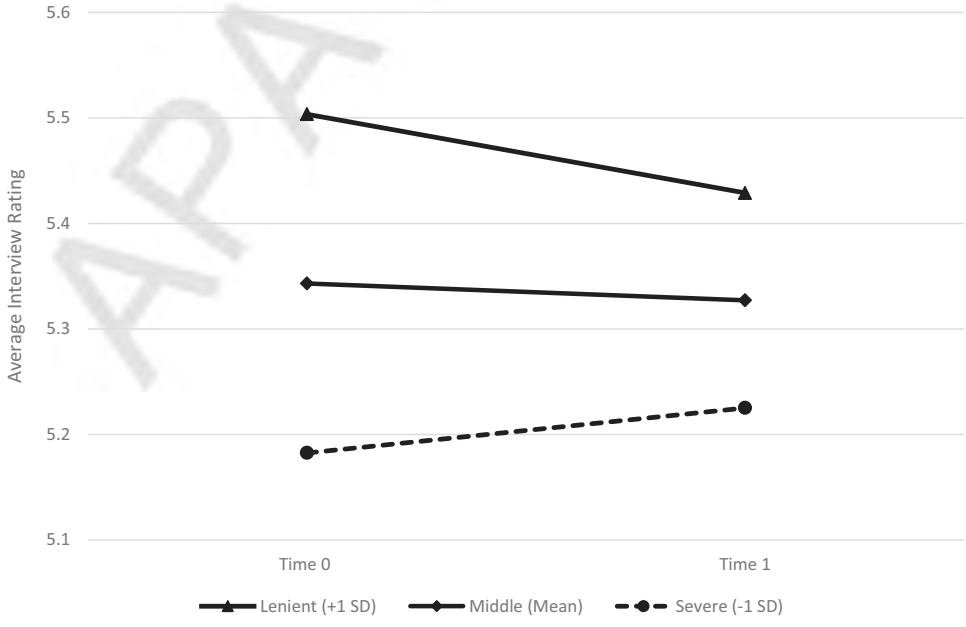

Figure 3. Time × Feedback interaction on interview ratings (over two time periods).

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

Table 3
Proportion of Variance in Interview Ratings Explained by Models (Time 0–Time 1)

Model	Observed variance	PVE (pseudo R^2)
Level 1		.03
Unconditional on time	.0060	
Conditional on time	.0058	
Level 2 slope differences		.19
Unconditional on mean rating at NFI1	.0094	
Conditional on mean rating at NFI1	.0076	

Note. PVE = proportion of variance explained; NFI = normative feedback intervention.

ers tended to increase rating variance from Time 0 to Time 1, regardless of individual mean rating at Time 0. Results using NFI1 ratings to predict changes in variance in our Level 2 analysis were insignificant, confirming that the change in variance was similar across interviewers.

Hypothesis 4. We expected interrater agreement and interrater reliability between interviewers to increase. It should be noted that the same pair of interviewers did not rate a high volume of applicants together. Rather, interviewer pairings were rotated according to interviewer schedules and other job duties. Therefore, it was not possible to use multilevel analyses to test this hypothesis. We instead used the interview as the unit of analysis, and calculated the absolute difference between the ratings of the two interviewers that interviewed each applicant. We then compared the mean absolute differences between raters at each time period. For interviews in which both interviewers had data at Time 0 and Time 1 (n = 1380 and 2061, respectively), the absolute difference between raters reduced from Time 0 (M = .22, SD = .17) to Time 1 (M = .19, SD = .15). In essence, the discrepancy between interviewers (already considerably low) was reduced by an additional 14% after receipt of feedback. Results of a t test indicated that this increased agreement was significant: $t(3439) = 4.26$, $p < .01$, supporting Hypothesis 4a.

To test Hypothesis 4b (that interrater reliability would increase after the NFI), we calculated the correlation between the two interviewers who interviewed each applicant. This analysis showed that, as predicted, the correlation between interviewers increased from Time 0 (r_{xy} = .85) to Time 1 (r_{xy} = .89). Thus, the proportion of variance explained (PVE) increased from .72 to .79. Fisher's r to z' transformation and a test of the difference between these correlations yielded significant results ($z = -4.77$, $p < .01$), supporting Hypothesis 4b.

Subsequent Time Periods (Hypothesis 5)

The first four hypotheses examined changes in interviewer behavior after the initial NFI; Hypothesis 5 looked at the changes resulting from subsequent NFIs. To do this, we conducted multiple RCM analyses following the steps outlined previously. We examined the change between Time 1 and Time 2 using NFI2, and the change between Time 2 and Time 3 was examined using NFI3. Results examining the change between Time 1 and Time 2 are found in the top half of Table 5. The Level 1 model showed an overall reduction in mean interview ratings, $t(33) = -3.94$, $p < .001$. When using the final Level 2 model to examine whether the change in interview ratings differed across interviewers, results using the most recent mean rating in NFI2 were not significant ($t(30) = -1.62$), but results using the combined rating over all available time periods were marginally significant, $t(30) = -1.83$, $p < .10$. This provides some initial support for Hypothesis 5, though the difference between the two t values is small and the PVE of the combined ratings (.06) is only slightly higher than that of the most recent ratings (.05; see Table 6).

RCM analyses examining the change between Time 2 and Time 3 (see the bottom half of Table 5) indicated that the combined mean interview rating over all available time periods interacted with time to significantly predict changes in subsequent mean interview ratings, $t(44) = -2.94$, $p < .01$, while the interaction between time and the most recent mean interview rating did not ($t(44) = -0.99$). As Table 6 indicates, the combined mean rating explained 11% of the variance in individual mean ratings from Time 2 to Time 3, while the most recent mean ratings did not explain any variance. These results provide further support for Hypothesis 5 by showing that the combined mean rating over all available time periods exhibits a stronger relationship with subsequent rating changes over time, as compared with the most recent mean rating.

All Time Periods (Hypothesis 6)

Hypothesis 6 predicted that the magnitude of the changes in (a) interview ratings, (b) variance, (c) interrater agreement, and (d)

Table 4
Random Coefficient Models Predicting Mean Interviewer Variance at Time 1

Model and parameter	Estimate	*SE*	*df*	95% CI (lower)	85% CI (upper)	*t*
Final Level 1 model (fixed slopes)						
Intercept	.190	.010	86	.170	.210	19.61***
Linear trend of time	.029	.009	45	.011	.047	3.04**
Final Level 2 model (using mean ratings at NFI1 to predict variance changes)						
Intercept	.185	.009	60	.167	.203	20.71***
Linear trend of time	.025	.010	43	.005	.045	2.53**
Mean rating at NFI1	−.102	.039	60	−.178	−.026	−2.61*
Linear Trend × Mean rating at NFI1	−.024	.048	43	−.118	.070	−.50

Note. CI = Confidence interval; NFI = normative feedback intervention.
* $p < .05$. ** $p < .01$. *** $p < .001$.

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

Table 5
Random Coefficient Models Predicting Mean Interview Rating (Subsequent Time Periods)

Model and parameter	Estimate	*SE*	*df*	95% CI (lower)	95% CI (upper)	*t*
Time 1 to Time 2 (NFI2)						
Final Level 1 model (fixed slopes)						
Intercept	5.410	.029	46	5.353	5.467	185.21***
Linear trend of time	−.066	.017	33	−.099	−.033	−3.94***
Final Level 2 model (using mean recent ratings at NFI2 to predict rating changes)						
Intercept	5.378	.023	39	5.333	5.423	238.83***
Linear trend of time	−.055	.016	30	−.086	−.024	−3.34**
Mean recent rating at NFI2	.701	.139	39	.429	.973	5.05***
Linear Trend × Mean recent rating	−.169	.104	30	−.373	.035	−1.62
Final Level 2 model (using mean combined ratings at NFI2 to predict rating changes)[b]						
Intercept	5.361	.023	45	5.316	5.406	236.82***
Linear trend of time	−.049	.017	32	−.082	−.016	−2.92**
Mean combined rating at NFI2	.844	.134	45	.581	1.107	6.29***
Linear Trend × Mean combined rating	−.183	.100	32	−.381	.015	−1.83†
Time 2 to Time 3 (NFI3)						
Final Level 1 model (fixed slopes)						
Intercept	5.212	.067	39	5.081	5.343	77.83***
Linear trend of time	−.024	.027	29	−.077	.029	.88
Final Level 2 model (using mean recent ratings at NFI3 to predict rating changes)						
Intercept	5.185	.067	38	5.054	5.316	76.93***
Linear trend of time	.028	.029	28	−.029	.085	.98
Mean recent rating at NFI3	1.214	.570	38	.085	2.343	2.13*
Linear Trend × Mean recent rating	−.243	.246	28	−.725	.239	−.99
Final Level 2 model (using mean combined ratings at NFI3 to predict rating changes)						
Intercept	5.209	.054	38	5.103	5.315	95.58***
Linear trend of time	.026	.024	28	−.021	.073	1.06
Mean combined rating at NFI3	1.884	.454	38	.994	2.774	4.15***
Linear Trend × Mean combined rating	−.611	.208	28	−1.019	−.203	−2.94**

Note. CI = confidence interval; NFI = normative feedback intervention.
† $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

interrater reliability would decrease over subsequent NFIs. We examine each part of this hypothesis separately.

Interview ratings. To test Hypothesis 6a, we conducted RCM analyses using data over all four time periods (Time 0–Time 3), looking for a nonlinear relationship between interviewer mean ratings and time. Steps 1 and 2 indicated differences in mean interview ratings across interviewers, and a significant linear relationship in which the overall mean interview rating declined by an average of .03 each time period, $t(147) = -4.71$, $p < .001$. This is possibly an effect of the perceived accountability of all raters as absolutely lenient raters (relative to the midpoint of the scale). A significant nonlinear relationship was not found, $t(146) = 1.11$, but we retained the quadratic trend to test our hypothesis by examining the effect of our Level 2 variable. The model in Step 3, which allowed for variability in slopes, fit the data better than the model in step two ($\chi^2_{diff}(2) = 28.02$, $p < .001$). The models in Step 4 that allowed for autocorrelation and heteroskadicity did not fit the data better than the model in Step 3 ($\chi^2_{diff}(2) = .03$ and $\chi^2_{diff}(2) = 2.54$, respectively). Therefore, we retained the quadratic model in Step 3 as our final Level 1 model, shown in Table 7. T7

We retained the mean rating at NFI1 as our Level 2 predictor variable through all time periods, as information from NFI1 is the only piece of feedback that was available to the interviewers at all time periods. Level 2 results indicated a significant interaction between the linear effect of time and centered mean ratings at NFI1, $t(102) = -5.01$, $p < .001$, but the interaction between the quadratic trend of time and ratings at NFI1 was not significant, $t(102) = 0.12$. These results (shown in Table 7) do not support Hypothesis 6a, though it is worth noting the significant linear interaction effect, which shows that feedback at NFI1 may still be related to overall interview changes after three time periods and two subsequent NFIs (as highlighted in Figure 4). While this effect was not predicted, it does match prior research regarding the primacy effect, in which initial information has a stronger and longer-lasting effect on the individual than subsequent information (Bond, Carlson, Meloy, Russo, & Tanner, 2007; Cable & Gilovich, 1998). However, these results should be interpreted cautiously because of the small number of interviewers with data at all time periods and high intercorrelations between NFI1 mean ratings and combined mean ratings in NFI2 and NFI3 ($r = .83$ and $r = .84$, respectively; see Table 1). F4

Variance. We expected a nonlinear relationship between time and within-individual mean rating variance, such that the magnitude of the increase in variance diminishes over time. Level 1 RCM analysis of the change in overall average mean variance over the four time periods showed an insignificant linear relationship ($t(105) = -1.03$). However, adding a quadratic term for our time variable in Step 2 was significant, $t(104) = -3.14$, $p < .01$, indicating the presence of a nonlinear relationship between time and mean variance. However, this relationship does not match our prediction in Hypothesis 6b. Interview ratings initially become more variable (see Hypothesis 3), but this was a fleeting effect. Variation of interviewers' ratings had actually declined by the fourth time period (after NFI 3). This result may have occurred

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

Table 6
Percentage of Variance in Interview Ratings Explained by Models (Subsequent Time Periods)

Model	Observed variance	PVE (pseudo R^2)
Time 1 to Time 2 (NFI2)		
Level 1		.27
Unconditional on time	.0060	
Conditional on time	.0044	
Level 2 slope differences[a]		.05
Unconditional on mean rating at NFI2	.0040	
Conditional on mean rating at NFI2	.0038	
Level 2 slope differences[b]		.06
Unconditional on mean rating at NFI2	.0050	
Conditional on mean rating at NFI2	.0047	
Time 2 to Time 3 (NFI3)		
Level 1		.00[c]
Unconditional on time	.0089	
Conditional on time	.0091	
Level 2 slope differences[a]		.00[c]
Unconditional on mean rating at NFI3	.0118	
Conditional on mean rating at NFI3	.0119	
Level 2 slope differences[b]		.11
Unconditional on mean rating at NFI3	.0097	
Conditional on mean rating at NFI3	.0086	

Note. PVE = proportion of variance explained; NFI = normative feedback interventions.
[a] Mean ratings from the most recent time period; [b] combined mean ratings over all available time periods; and [c] negative values were reset to .00 as suggested by Snijders and Bosker (1994) and done in similar past research (Thoresen, Bradley, Bliese, & Thoresen, 2004).

because variance was not part of the NFI feedback and was not something consciously managed by interviewers, or it could be an artifact of the reduced number of interviewers with data at all time periods. If these results represent a true reduction in the differentiation between candidates by individual interviewers, this could negatively affect the validity of interview ratings.

Interrater agreement. Using interviewers with data at all time periods and similar analyses as Hypothesis 4a, we sought to determine how interrater agreement changed over the course of the study. Mean absolute differences of interviewers rating the same candidate decreased over each time period, from Time 0 (.209) to Time 1 (.174) to Time 2 (.161) to Time 3 (.160). The continual decrease in rating differences signifies an increase in interrater agreement over each time period. However, the changes get smaller at each time period and only the change from Time 0 to Time 1 is significant: $t(928) = 3.14$, $p < .01$. These results show a pattern of results that support Hypothesis 6c.

Interrater reliability. We conducted analyses similar to Hypothesis 4b (using interviewers with data at all time periods) to determine how interrater reliability changed through multiple time periods and NFIs. Results indicate that reliability increased significantly from Time 0 to Time 1 (from .81 to .88; $z = -3.64$, $p < .01$) and from Time 1 to Time 2 (from .88 to .92; $z = -3.41$, $p < .01$), but then decreased from Time 2 to Time 3 (from .92 to .87). This decrease, while similar in magnitude to the prior increases was only marginally significant ($z = 1.92$, $p < .10$) because of the smaller relative n in Time 3. Therefore, we conclude that Hypothesis 6d is partially supported.

Study 2: Method

Study 2 allowed for a test of the final relationships in the regulatory model outlined in Figure 1 by examining the effect of interviewer-level error (or lack thereof) on rating validity and applicants selected. The focus of this study is a data simulation, in which interviewer-level error is added to simulated interview ratings. As a follow-up, interview ratings from two published studies in which interviewers rated a high volume of applicants are standardized to examine the effect of a statistical reduction of interview-level error on the validity of interview ratings.

First, to examine the effects of leniency and severity on interview validity and applicant selection, we simulated a data set with properties similar to Study 1 data, and added a simulated job performance variable to examine Hypotheses 7a and 7b. Specifically, our simulated data set consisted of 100 interviewers and a total of 9,576 simulated interview ratings.

At the interviewer level, the mean interview ratings (I), mean rating *SD* (σ_I), and number of interview ratings (N) per interviewer were randomly generated using the SIMPLAN command in SPSS, with distributions modeled after the distributions of our Study 1 data at Time 1 (when our theory and data suggest that NFI1 reduced distributional errors). A randomly generated job perfor-

Table 7
Random Coefficient Model Predicting Mean Interviewer Rating Over All Time Periods

Model and parameter	Estimate	*SE*	*df*	95% CI (lower)	95% CI (upper)	*t*
Final Level 1 model (variable slopes)						
Intercept	5.287	.012	146	5.263	5.311	428.15***
Linear trend of time	−.524	.138	146	−.794	−.254	−3.80***
Quadratic trend of time	.097	.087	146	−.074	.268	1.11
Final Level 2 model (using mean ratings at NFI1 to predict rating changes over all time periods)						
Intercept	5.320	.008	102	5.304	5.336	635.22***
Linear trend of time	−.322	.113	102	−.543	−.101	−2.86**
Quadratic trend of time	.107	.083	102	−.056	.270	1.30
Mean rating at NFI1	.483	.040	60	.405	.561	12.00***
Linear Trend × Mean rating at NFI1	−2.856	.571	102	−3.975	−1.737	−5.01***
Quadratic Trend × Mean rating at NFI1	.050	.438	102	−.808	.908	.12

Note. CI = Confidence interval; FNI = normative feedback intervention.
** $p < .01$. *** $p < .001$.

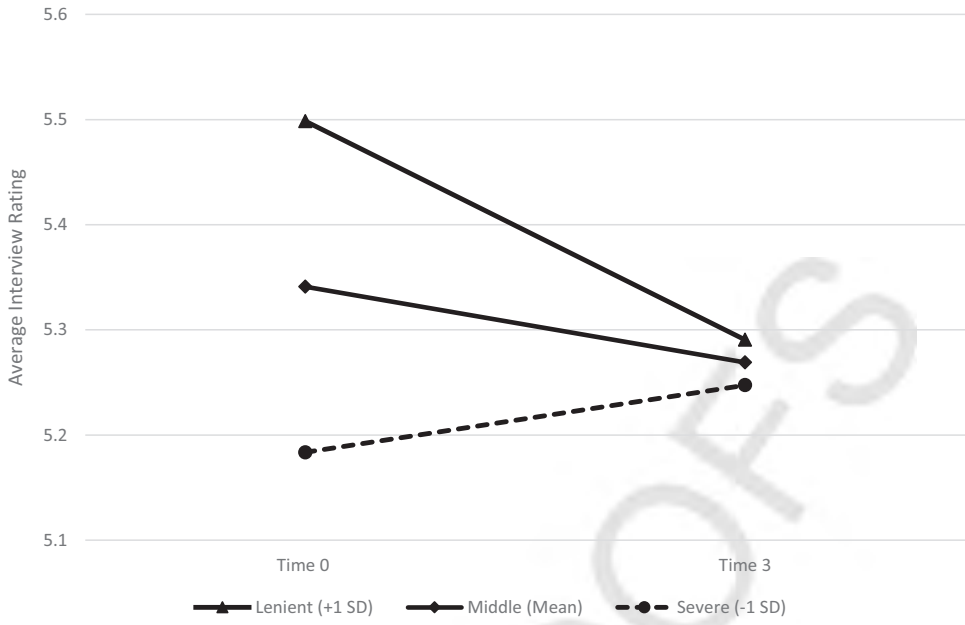

Figure 4. Linear Time × Feedback interaction on interview ratings (over all time periods).

mance variable was created to have a correlation with interview ratings corresponding to the meta-analytic validity coefficient of the structured interview ($\rho = .44$) found in McDaniel, Whetzel, Schmidt, and Maurer (1994). This was accomplished by first standardizing the mean interview rating variable (I) into a *z* score (Z_I) and creating a normally distributed variable (using the NORMINV command in Excel with a mean of 0 and *SD* of 1) to represent standardized mean job performance (Z_J). Following this, we calculated a raw performance variable (J) that correlated with I by using the following equation: $J = \rho Z_I + \sqrt{1-\rho^2} Z_J$. The resulting scatterplot (see Figure 5) shows a distribution of relationships similar to what would be expected from a field study. Finally, we randomly generated a normally distributed validity variable (V) across interviewers using the NORMINV command in Excel (using the mean and *SD* [$\sigma_\rho = .27$] from McDaniel et al., 1994).

To populate the data at the interview level, we randomly generated two standardized variables (Z_1 and Z_2) for each of the 100 interviewers (using the same procedure as Z_J above), with the number of rows for each interviewer being equal to the *N* generated in the interviewer-level data. The raw individual interview ratings (X) were calculated separately for each interviewer as follows: $X = I + \sigma_I Z_1$. The individual job performance ratings (Y) were calculated using the following equation: $Y = J + \sigma_J(Z_1 V + Z_2\sqrt{(1-V)^2})$, with σ_J being a constant (.83) taken from the *SD* of performance ratings in Van Iddekinge et al. (2006). The resulting correlation between interview ratings and job performance at the individual level was $r = .41$.

To represent the leniency and severity errors at the interviewer level, we randomly assigned each interviewer a constant error term with a normal distribution, mean, and *SD* that matched the change in ratings from Time 0 to Time 1 in Study 1 ($M = .02$, $SD = .06$). This mean is .02 (rather than −.02; see Table 2) because we are adding error into our simulated data. One way to look at it is that we are using this information to recreate a simulated Time 0, given our simulated Time 1 data. Each interviewer's constant error term was then added to all of the interview ratings for that interviewer to represent interview ratings that included interviewer-level leniency and severity biases. Because it is probable that these biases are not completely constant in interviewers, we also tried giving each interviewer a normal distribution ($SD = .03$) around their constant error term, and results of the analyses were nearly identical. For parsimony sake, we report the results with only the constant error term.

Study 2: Results

Hypothesis 7a predicted the correlation between interview ratings and job performance would increase as distributional errors are reduced and interrater effects (agreement and reliability) are increased. To test this hypothesis, we compared the correlation between interview ratings and job performance ratings both before and after the distributional error was introduced, with the expectation that the correlation would be higher before adding the error to the interview ratings. Results indicated that the validity remained the same ($r = .41$) in the two conditions, providing no support for Hypothesis 7a.

In essence, the simulation in this study is akin to adding interviewer-level error to otherwise standardized interview ratings, expecting (but not seeing) a decrease in validity. The reverse of what we attempted would be to standardize interview ratings at the interviewer level, theoretically reducing interview-level bias and increasing validity. Thus, in addition to our simulated results, we examined the effect of interviewer-level standardization by reporting the results of one published research study and reanalyzing the data from two existing studies in which interviewers each rated enough applicants to infer that mean rating differences could be symptomatic of interviewer effects, rather than true variation in

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

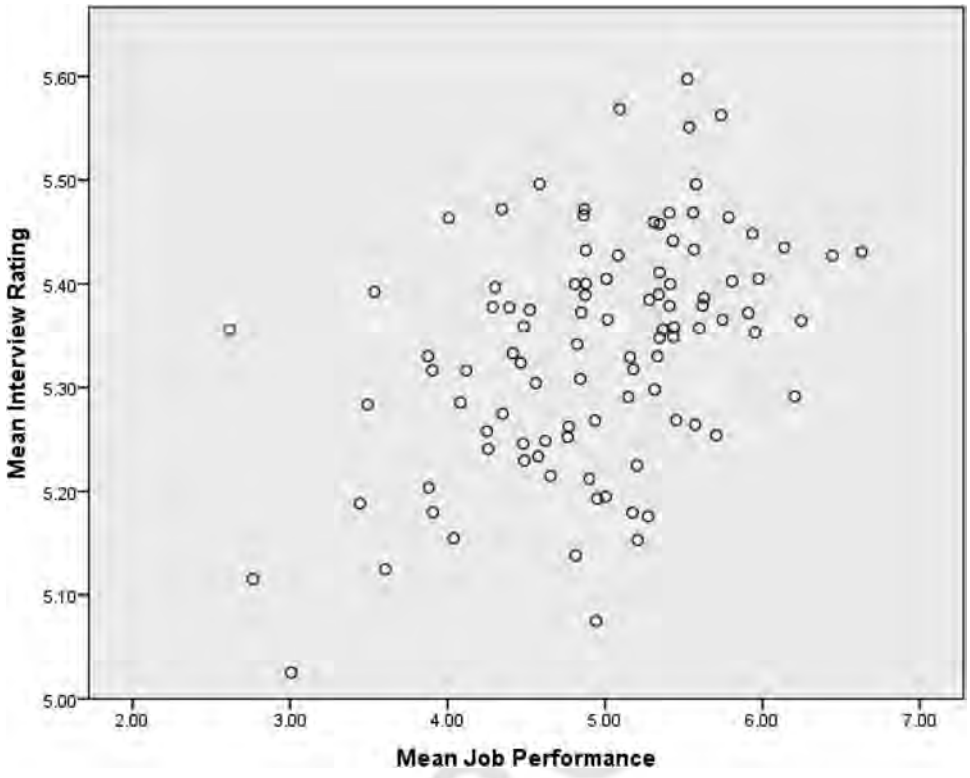

Figure 5. Scatterplot of simulated mean interview rating and job performance variables at the interviewer level ($n = 100$).

applicants. This was done to determine whether statistical standardization increased interview rating validity. First, similar to our simulated data, Pulakos et al. (1996), using 62 interviewers who rated between 11 and 48 applicants ($M = 24.9$), reported that using raw interview ratings or interview ratings standardized at the interviewer level resulted in a virtually identical correlation with supervisor-rated job performance. Second, we reanalyzed data in two existing published studies (Morgeson et al., 2005; Van Iddekinge et al., 2006). For Morgeson et al. (2005), we limited our analysis to the 20 interviewers who conducted at least five interviews ($M = 9.8$). Standardizing structured interview ratings at the interviewer level in a manner similar to Pulakos et al. (1996) led to no change in correlation between ratings and performance for these data. For Van Iddekinge at al. (2006), 64 interviewers each conducted between 14 and 43 interviews ($M = 29.4$). Interviewer-level standardization netted a nonsignificant increase of .02.[5] These results follow the same trend as the Study 2 simulation and do not support Hypothesis 7a.

Fn5

It should be noted that, in all of these cases, the interviews were very highly structured, including high structure on at least 10 of Campion et al.'s (1997) 15 components of interview structure, such as all using anchored rating scales. Less structured or unstructured interviews would likely have more rater error and stronger validity effects. For example, early research on rating scales in both selection and performance appraisal contexts showed that using only graphic rating scales resulted in about three times the amount of distributional error across raters as using anchored ratings scales (Burnaska & Hollmann, 1974; Vance, Kuhnert, & Farr, 1978). Increasing the distributional error by the same magnitude in our simulation study significantly reduced validity to $r = .39$ ($z = 1.65$, $p < .05$), highlighting that these effects would likely be greater in a less structured or unstructured interview context.

The final Hypothesis (7b) suggested that there would be a notable difference in which applicants are selected, dependent upon the presence of distributional errors. We set four different selection cutoffs to test this hypothesis: the top 5% ($n = 479$), 10% ($n = 958$), 25% ($n = 2394$), and 50% ($n = 4788$) of applicants. Nine percent of applicants selected differ when selecting either the top 5 or 10% ($ns = 41$ and 82, respectively), 6% differ ($n = 132$) when selecting the top 25%, and 4% differ ($n = 183$) when selecting the top 50%. All of these proportional differences in applicants selected are significantly different from zero ($p < .05$) and the percentages would likely be relevant in a true selection context. These results support Hypothesis 7b, and it is worth noting that the proportion of applicants affected by rater error tends to increase as the selection rate decreases. Thus, the more selective the hiring process, the greater the impact that rater error may have on selection.

Discussion

We developed a regulatory model of normative feedback interventions in the context of interview ratings, which we tested in a large-scale longitudinal study and a follow-up study using simu-

[5] We thank Chad Van Iddekinge for reanalyzing these data.

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

lated data. In the first study, using field data from actual structured interviews, our first set of hypotheses examined change over a 2-year period. Results indicated that lenient and severe interviewers made adjustments to their subsequent ratings that were greater than what would be expected from regression to the mean, and that the reductions in lenient ratings were significantly greater than the increase of severe raters. According to our model, this most likely occurred because the relationships between social comparison and behavioral change is moderated by accountability pressures, such that the behavioral change is stronger for lenient raters. A second outcome of the NFI was that within-interviewer variance in ratings increased across interviewers, indicating more differentiation between applicants. This finding shows that the response to normative feedback was more than a simple shift in interviewers' ratings toward the overall mean (that could increase the threat of central tendency error), but that feedback also broadened the interviewers' range of ratings, regardless of their relative standing in the NFI. Third, interviewer ratings also showed greater convergence through increased interrater agreement and interrater reliability after receipt of the NFI. These positive outcomes of the NFI result from closer alignment in interviewers' mean ratings and rating distributions. Greater agreement and reliability enhance the psychometric properties of the interview and extend the upper boundary of validity, thus opening up the possibility that interview ratings could be more valid. Overall, these results support the notion that interviewers' individual rating differences can be reduced through receipt of normative feedback, and that these minimized differences have practical benefits for at least some of the psychometric properties of the interview.

Our second set of hypotheses examined how multiple NFIs with different pieces of information affect subsequent structured interview ratings. Our analyses indicated that, over time, combined mean ratings over multiple time periods gained relative importance over the mean ratings of the most recent time period in predicting subsequent behavioral changes. Finally, we hypothesized that the magnitude of change in mean interview ratings, variance, interrater agreement, and interrater reliability of the interview ratings would decrease over subsequent NFIs. A consistent pattern of decreased magnitude was found only for interrater agreement, and partial support was found for interrater reliability. A nonhypothesized linear relationship between changes in interview ratings and the relative standing on the first wave of feedback (NFI1) revealed a potential primacy effect for the first receipt of feedback. However, this conclusion is speculative and more research is needed on the topic.

Given our findings in Study 1 and our desire to understand whether the changes in interviewer ratings, rating variance, agreement, and reliability practically impact the hiring process, we simulated a data set in Study 2 with similar properties to our Study 1 data, and introduced a simulated job performance variable that correlated with our interview rating variable. The findings revealed that adding a constant error to the simulated interview ratings did not reduce the validity of the interview rating (though adding error that matched previous research on less structured ratings did), but it did change a meaningful percentage of the candidates hired depending on the selection ratio. In addition to the simulated data, reanalyzing two previously published data sets by standardizing ratings at the interviewer level did not significantly increase interview rating validity.

In retrospect, our lack of significant criterion-related validity results is similar to what has been found in past interviewing research focusing on standardization of ratings. Early research on reducing interviewers' halo error showed that training was effective in reducing halo, but did not impact the validity of interview ratings (Borman, 1975). Pulakos et al. (1996) similarly found that standardizing interviewer ratings did not affect interview validity. Our findings regarding validity and hiring effects mirrored prior selection research examining the impact of correcting for response distortion (i.e., faking, socially desirable responding) in personality assessments (Christiansen, Goffin, Johnston, & Rothstein, 1994; Rosse, Stecher, Miller, & Levin, 1998). For example, Christiansen et al. (1994) found that correcting for response distortion did not impact the criterion-related validity of the personality test, but it did impact a similar proportion of hiring decisions as our simulation. Rosse et al. (1998) also found that corrections for response distortion on personality tests had a large impact on potential hiring decisions, particularly in highly selective contexts. In a Monte Carlo study similar to our Study 2 simulation (though larger in scope), Paunonen and LeBel (2012) also found very little reduction in criterion-related validity coefficients when adding response distortion to personality assessment scores. This discrepancy between validity and hiring effects may occur because correlation coefficients are typically very robust and not sensitive to small changes in rank ordering, particularly when validity is of moderate-to-low magnitude (Conger & Jackson, 1972; Drasgow & Kang, 1984; Rosse et al., 1998). However, even without reduced validity coefficients, individual rating differences may provide less valid information at the individual level (Ben-Porath & Waller, 1992; Paunonen & LeBel, 2012). Overall, our results suggest that reducing distributional interviewer errors may have a practical impact on which individual applicants are selected, even if overall criterion-related validity remains relatively unchanged.

Theoretical and Practical Implications

This study extended control theory (Campion & Lord, 1982; Carver & Scheier, 1981) and social comparison theory (Festinger, 1954; Wood, 1989) by formulating and testing a regulatory model of normative feedback interventions in a structured interview context that incorporates both theories. This extends control theory by showing (a) how a control theory framework, by utilizing normative information, can operate in the absence of an absolute standard or referent from the environment; (b) how normative feedback can provide both the sensor and the referent signals that are necessary for control theory to operate in such a context; and (c) that the motivation for self-improvement is central to the behavioral change. Social comparison theory is extended because control theory helps to explain how social comparisons operate in a context where self-improvement is not realized by achieving higher scores, but by reducing discrepancies, and by similarly showing that comparisons over longer periods of time may be given more weight (e.g., the comparison over all time periods was more predictive of change than the comparison of the most recent time period). Both theories are extended through the addition of accountability as a moderator in our model that helps explain differences in behavioral change as a response to feedback.

Integrating and extending control and social comparison theories can help us understand behavioral change in response to

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

normative feedback in a selection context, matching similar findings both inside and outside of the management literature (e.g., Gaudine & Saks, 2001; Klein, 1997; Neighbors et al., 2006; Schmiege et al., 2010). Thus, another major implication of this research is that control theory and social comparison theory can be usefully applied together to explain the impact of normative feedback on distributional errors in interviewer ratings. Finally, given that normative feedback reduced interviewer rating differences in a manner consistent with other interview structuring methods (e.g., Campion et al., 1997; Levashina et al., 2014), this study extends interviewing theory by suggesting that providing the interviewer with performance feedback on actual interview ratings should be considered as a potential new component of interview structure. Future research is needed, however, to better understand normative feedback (especially over time), and to examine other types of interviewer feedback, before a firm conclusion can be drawn.

Our research also has practical implications. First, results showed that normative feedback can reduce interviewer leniency and severity, even in a structured interview context using trained and experienced interviewers. Interviewer feedback could be provided by organizations in an effort to minimize interviewer individual rating differences and make interview ratings more reliable. However, the value of repeated normative feedback is unclear and practical recommendations cannot be made until future research is conducted. Second, our finding that combined mean ratings predict subsequent rating changes better than mean ratings from the most recent time period suggests that organizations should consider giving both types of normative feedback to employees. Third, though no effect was found on interview rating validity, our simulated data illustrated that interviewers' behavioral changes as a response to normative feedback are likely to impact which applicants are hired. However, as discussed in the next section, more research using nonsimulated data is needed regarding the practical effects of normative feedback on interview validity. Finally, given that our results indicated that interviewers providing lenient ratings consistently had larger rating changes than other interviewers, normative feedback may be particularly effective in combating leniency in other rating contexts where it is a common problem, such as performance appraisals.

Study Limitations

Our study is built on many strengths that add to the discussion on structured interviews, such as the applied setting and examination of multiple NFIs. However, we also recognize the limitations of our study. One limitation is that the relatively small effect sizes in our study are likely to be underestimates of the effects that could be realized in a typical structured interviewing context because of a couple of reasons. First, our context was one of comparatively high interview structure—our interviews were based on all 15 components of structure (Campion et al., 1997), while the average structured interview includes only six components (Levashina et al., 2014). Second, interviewers in our study were required to have a consensus discussion on any rating disparity of two points or greater on the seven-point scale (though agreement on ratings was not required). This consensus discussion could have provided feedback to interviewers regarding their relative leniency or severity that is not captured in our data (the ratings recorded in our data would have occurred after any consensus discussion). Therefore, a similar study in a less structured, or even unstructured, interview context and/or a context void of a consensus discussion would likely produce even larger effects. A third limitation is that the number of interviewers in the study was somewhat small, particularly for analyses extended over all time periods. While we believe the results are reliable, having been based on ratings from several thousands of interviews, a larger number of interviewers would have increased our ability to detect significant changes.

There were also limitations to the study design. While using real interviewers conducting actual high-stakes interviews as part of an organization's selection process gives the study strong fidelity, we were constrained by the organization as to what information we were able to collect. Specifically, we were unable to obtain data regarding which applicants were ultimately hired or subsequent job performance information. Therefore, we were unable to examine changes in the criterion-related validity of interviewer ratings (one of the outcomes in our model). Our simulated second study allowed us to make some validity judgments, but research using nonsimulated data to directly measure the effects of NFIs on interview rating validity is needed.

Another limitation to the longitudinal design is historical threat—outside events occurring during the course of the study that could confound the results (Shadish, Cook, & Campbell, 2002). For example, a widespread economic recession took place during the course of this study. This may have contributed to a surge in applicants that would explain the increase in interviews conducted during Time 1, which could have negatively affected the quality of applicants, contributing to the overall decrease in interview ratings over time. However, this is unlikely, as there are several previous hurdles in the hiring process before the interview that would eliminate low-quality candidates (e.g., an aptitude test and evaluation of education or work experience). Furthermore, the hiring needs of the organization during this recession resulted in the selection ratio staying roughly the same throughout the study. A second historical threat to consider is the interviewer training discussed earlier. This full-day training occurred on an annual basis in the organization even before the current study, and does not likely constitute a historical threat. In addition, the NFI took place in March and the training took place 6 months later (in September). Post hoc analyses of our Study 1 data indicated that a vast majority of the rating changes occurred in the first 6 months after feedback was given each year. This strengthens the case for the annual NFI, not the annual training, being the driver behind those changes.

Next, the feedback given to interviewers was informational in nature, and was not tied to any job-related criteria. The NFI was not packaged in a way that highlighted important information for the interviewers' attention. Rather, while interviewers were informed during annual training that they should pay attention to the NFI, it was left to the interviewer to determine the most important takeaways from the feedback. In addition, there were no tangible incentives or punishments (e.g., promotion, raises, or disciplinary action) that resulted from the NFIs. This lack of directed attention and job-related consequences likely reduced the accountability that interviewers felt in regards to the NFIs. According to our model, this lack of accountability would have reduced the likelihood of behavioral change in response to social comparison, particularly as the novelty of the feedback wore off in subsequent NFIs. According to control theory, discrepancies can be reduced through be-

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

havioral adjustment and/or through goal adjustment (Campion & Lord, 1982). The process we have tested in our study is a process of behavioral adjustment (changes in interview ratings). However, goal adjustment can also reduce the discrepancies felt by interviewers if the interviewer's goal is adjusted away from trying to have an average rating similar to others. This could simply be the result of a change in perspective over time (e.g., "I believe my ratings are accurate, regardless of how they compare to others"). Additionally, however, given that the specific regulatory process we describe is only one subsystem in a larger system of multiple goals, expectations, and behaviors that constitute the overall job, interviewers' attention in our study was likely turned from our feedback—that could have been viewed as less useful and/or relevant (Gaudine & Saks, 2001; Klein, 1997)—to regulate behavior toward other goals that are more closely tied to work-related consequences (Taylor et al., 1984; Vancouver, 2005). In other words, the specific goal of reducing interview rating discrepancies likely lost priority over time as novelty wore off and other goals more explicitly tied to job-related outcomes became more salient.

Finally, one statistical limitation was the use of mean ratings and mean variance as outcomes of distributional errors. It should be noted that these distributional indices are statistical artifacts that may be the result of a variety of processes, not just distributional error. Future experimental research should attempt to operationalize and/or measure distributional errors in multiple ways, to determine the best indicators of such errors for future field research.

Future Research Directions

We believe the literature can be served well by a variety of potential future research directions. First of all, we previously mentioned our inability to collect data that would have allowed us to test changes in the criterion-related validity of interview ratings. Studying this outcome is essential to fully understanding the impact of normative feedback, and this should be a priority for future research. Second, our NFIs focused on differences in mean ratings. Examining other forms of normative feedback information (such as rating variance) and different modes of delivery (such as in a training context) would be useful.

Third, the patterns that resulted from multiple NFIs in our study did not fully support our predictions. This lack of support over multiple time periods is likely explained by environmental factors influencing the accountability construct in our model, in that a lack of job-related outcomes tied to the NFIs may have reduced accountability over time. Future research should attempt to replicate and further explain what happens over multiple feedback periods in different scenarios and/or contexts. For example, an unexpected linear relationship between ratings at NFI1 and subsequent ratings over all time periods revealed a potential primacy effect that should be examined more directly.

Fourth, our feedback interventions were spaced a year apart. Feedback research has, with differing results, examined whether the frequency of feedback affects resulting behaviors and attitudes (Chhokar & Wallin, 1984; Cook, 1968; Lurie & Swaminathan, 2009). Cohen et al. (2003) explain that, in general, the longer the time period, the weaker the expected influence of the independent variable(s) on behavioral change. Bandura's (1991) social–cognitive theory also suggests that behavioral change is more readily achieved when the feedback and/or consequences are temporally proximal. As discussed previously, a majority of the rating changes occurred in the first 6 months after each NFI. Therefore, it would be worth examining whether increasing the frequency of normative feedback would have a greater effect on interview behavior.

A fifth area for future research is an examination of how individual differences may impact interviewers' reactions to feedback. Although we did not find any significant race or gender differences in our data, other individual attributes may impact interviewers' reactions. For example, less-experienced interviewers may have stronger behavioral reactions to normative feedback because they do not have the training and experience that likely bring more confidence to experienced interviewers' ratings. Unfortunately, the restriction of range in our sample (all highly experienced interviewers) precluded a test of this hypothesis. Personality variables may also play a pivotal role in how interviewers respond to feedback. Interviewers who are more agreeable, conscientious, and open to feedback may be more likely to make changes as a result of normative feedback.

Future research might extend control and social comparison theories by moving beyond leniency and severity to explain how central tendency could be reduced by normative feedback. For example, giving normative feedback on rating variance could draw interviewers' attention and encourage more differentiation between applicants. Feedback on rating patterns might be useful in reducing other interviewer rating errors, such as primacy and recency biases if early and/or late ratings in a given time period (such as when a large volume of interviews are conducted in a day, week, or longer assessment cycle) show a tendency to be higher than other ratings. In addition, while we operationalized leniency by examining mean rater differences, other operationalizations of leniency (e.g., comparing ratings to actual applicant behaviors) could provide valuable information in contexts where a high volume of comparative interviewers is not available, or where raters may all tend toward leniency errors (e.g., performance appraisals).

The likelihood that the manner in which feedback is presented, along with giving different types of feedback, is another avenue for future research. In our context, the NFI was anonymous and delivered to individual interviewers via email. Changing the NFI in various ways (removing the anonymity, presenting it to the interviewers as a group, etc.) would impact the felt accountability that acts as a moderator of behavioral change resulting from social comparison. In addition, presenting and/or highlighting different pieces of information (such as the difference between individual and overall mean ratings, or—as discussed in the previous paragraph—within-interviewer variance and/or rating patterns) would impact the perceived discrepancy between the inputs for social comparison (the referent and sensor signals). Research is warranted regarding how these changes in the normative feedback information and context affect subsequent behavioral change.

One final suggestion for future research is to examine how normative feedback affects unstructured interviews. While there is overwhelming evidence that structured interviews provide more valid and reliable results than unstructured interviews (summarized in Campion et al., 1997), we also know that unstructured interviews are frequently used in practice. We believe that a similar study in the context of unstructured interviews could be

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

practically beneficial and would likely show that normative feedback has much larger effects on subsequent interviewer ratings.

References

Anderson, N., & Witvliet, C. (2008). Fairness reactions to personnel selection methods: An international comparison between the Netherlands, the United States, France, Spain, Portugal, and Singapore. *International Journal of Selection and Assessment, 16,* 1–13. http://dx.doi.org/10.1111/j.1468-2389.2008.00404.x

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35,* 281–322. http://dx.doi.org/10.1111/j.1744-6570.1982.tb02197.x

Atwater, L., & Brett, J. (2006). Feedback format: Does it influence manager's reaction to feedback? *Journal of Occupational and Organizational Psychology, 79,* 517–532. http://dx.doi.org/10.1348/096317905X58656

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50,* 248–287. http://dx.doi.org/10.1016/0749-5978(91)90022-L

Bandura, A., & Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology, 45,* 1017–1028. http://dx.doi.org/10.1037/0022-3514.45.5.1017

Bass, B. M., & Avolio, B. J. (1989). Potential biases in leadership measures: How prototypes, leniency, and general satisfaction relate to ratings and rankings of transformational and transactional leadership constructs. *Educational and Psychological Measurement, 49,* 509–527. http://dx.doi.org/10.1177/001316448904900302

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323–370. http://dx.doi.org/10.1037/1089-2680.5.4.323

Ben-Porath, Y. S., & Waller, N. G. (1992). "Normal" personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment, 4,* 14–19. http://dx.doi.org/10.1037/1040-3590.4.1.14

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205–212.

Bliese, P. D. (2013). *Multilevel: Multilevel functions* (R package, version 2.5). Retrieved from http://CRAN.R-project.org/package=multilevel

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods, 5,* 362–387. http://dx.doi.org/10.1177/109442802237116

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *Accounting Review, 86,* 1549–1575. http://dx.doi.org/10.2308/accr-10099

Bolster, B. I., & Springbett, B. M. (1961). The reactions of interviewers to favorable and unfavorable information. *Journal of Applied Psychology, 45,* 97–103. http://dx.doi.org/10.1037/h0048316

Bond, S. D., Carlson, K. A., Meloy, M. G., Russo, J. E., & Tanner, R. J. (2007). Information distortion in the evaluation of a single option. *Organizational Behavior and Human Decision Processes, 102,* 240–254. http://dx.doi.org/10.1016/j.obhdp.2006.04.009

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60,* 556–560. http://dx.doi.org/10.1037/0021-9010.60.5.556

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management, 18,* 321–352. http://dx.doi.org/10.1177/014920639201800206

Burnaska, R. F., & Hollmann, T. D. (1974). An empirical comparison of the relative effects of rater response biases on three rating scale formats. *Journal of Applied Psychology, 59,* 307–312. http://dx.doi.org/10.1037/h0036536

Cable, D. M., & Gilovich, T. (1998). Looked over or overlooked? Prescreening decisions and postinterview evaluations. *Journal of Applied Psychology, 83,* 501–508. http://dx.doi.org/10.1037/0021-9010.83.3.501

Campion, M. A., & Lord, R. G. (1982). A control systems conceptualization of the goal-setting and changing process. *Organizational Behavior and Human Performance, 30,* 265–287. http://dx.doi.org/10.1016/0030-5073(82)90221-5

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50,* 655–702. http://dx.doi.org/10.1111/j.1744-6570.1997.tb00709.x

Carver, C. S., & Scheier, M. F. (1981). *Attention and self-regulation: A control-theory approach to human behavior*. New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4612-5887-2

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin, 92,* 111–135. http://dx.doi.org/10.1037/0033-2909.92.1.111

Carver, C. S., & Scheier, M. F. (1985). A control-systems approach to the self-regulation or action. In J. Kuhl & J. Beckman (Eds.), *Action control: From cognition to behavior* (pp. 237–265). Berlin: Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-69746-3_11

Chhokar, J. S., & Wallin, J. A. (1984). A field study of the effect of feedback frequency on performance. *Journal of Applied Psychology, 69,* 524–530. http://dx.doi.org/10.1037/0021-9010.69.3.524

Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47,* 847–860. http://dx.doi.org/10.1111/j.1744-6570.1994.tb01581.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Routledge.

Conger, A. J., & Jackson, D. N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement, 32,* 579–599. http://dx.doi.org/10.1177/001316447203200303

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579. http://dx.doi.org/10.1037/0021-9010.80.5.565

Cook, D. M. (1968). The impact on managers of frequency of feedback. *Academy of Management Journal, 11,* 263–277. http://dx.doi.org/10.2307/254752

Davidson, R. N. (2003). *Providing comparative feedback to raters as a training intervention to reduce strictness and leniency bias in performance appraisals*. (Unpublished doctoral dissertation). Alliant International University, San Diego, CA.

DeNisi, A. S., & Kluger, A. N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? *The Academy of Management Executive, 14,* 129–139.

Dougherty, T. W., Ebert, R. J., & Callendar, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71,* 9–15. http://dx.doi.org/10.1037/0021-9010.71.1.9

Drasgow, F., & Kang, T. (1984). Statistical power of differential validity and differential prediction analyses for detecting measurement nonequivalence. *Journal of Applied Psychology, 69,* 498–508. http://dx.doi.org/10.1037/0021-9010.69.3.498

Dreher, G. F., Ash, R. A., & Hancock, P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. *Personnel Psychology, 41,* 315–325. http://dx.doi.org/10.1111/j.1744-6570.1988.tb02387.x

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

Eddleston, K. A. (2009). The effects of social comparisons on managerial career satisfaction and turnover intentions. *Career Development International, 14,* 87–110. http://dx.doi.org/10.1108/13620430910933592

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes, 58,* 51–100. http://dx.doi.org/10.1006/obhd.1994.1029

Edwards, J. R., & Van Harrison, R. (1993). Job demands and worker health: Three-dimensional reexamination of the relationship between person-environment fit and strain. *Journal of Applied Psychology, 78,* 628–648. http://dx.doi.org/10.1037/0021-9010.78.4.628

Farh, J., & Dobbins, G. H. (1989). Effects of self-esteem on leniency bias in self-reports of performance: A structural equation model analysis. *Personnel Psychology, 42,* 835–850. http://dx.doi.org/10.1111/j.1744-6570.1989.tb00677.x

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140. http://dx.doi.org/10.1177/001872675400700202

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

Furnham, A., & Stringfield, P. (1998). Congruence in job-performance ratings: A study of 360° feedback examining self, managers, peers, and consultant ratings. *Human Relations, 51,* 517–530. http://dx.doi.org/10.1177/001872679805100404

Gaudine, A. P., & Saks, A. M. (2001). Effects of an absenteeism feedback intervention on employee absence behavior. *Journal of Organizational Behavior, 22,* 15–29. http://dx.doi.org/10.1002/job.73

Graves, L. M., & Karren, R. J. (1996). The employee selection interview: A fresh look at an old problem. *Human Resource Management, 35,* 163–180. http://dx.doi.org/10.1002/(SICI)1099-050X(199622)35:2<163::AID-HRM2>3.0.CO;2-W

Harris, M. M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personnel Psychology, 42,* 691–726. http://dx.doi.org/10.1111/j.1744-6570.1989.tb00673.x

Heneman, H. G., III, Schwab, D. P., Huett, D. L., & Ford, J. J. (1975). Interviewer validity as a function of interview structure, biographical data, and interviewee order. *Journal of Applied Psychology, 60,* 748–753. http://dx.doi.org/10.1037/0021-9010.60.6.748

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92,* 1270–1285. http://dx.doi.org/10.1037/0021-9010.92.5.1270

Hollmann, T. D. (1972). Employment interviewers' errors in processing positive and negative information. *Journal of Applied Psychology, 56,* 130–134. http://dx.doi.org/10.1037/h0032661

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79,* 184–190. http://dx.doi.org/10.1037/0021-9010.79.2.184

Huffcutt, A. I., & Woehr, D. J. (1999). Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior, 20,* 549–560. http://dx.doi.org/10.1002/(SICI)1099-1379(199907)20:4<549::AID-JOB921>3.0.CO;2-Q

Ilies, R., & Judge, T. A. (2005). Goal regulation across time: The effects of feedback and affect. *Journal of Applied Psychology, 90,* 453–467. http://dx.doi.org/10.1037/0021-9010.90.3.453

Jagacinski, C. (1991). Personnel decision making: The impact of missing information. *Journal of Applied Psychology, 76,* 19–30. http://dx.doi.org/10.1037/0021-9010.76.1.19

Jagacinski, C. (1995). Distinguishing adding and averaging models in a personnel selection task. *Organizational Behavior and Human Decision Processes, 61,* 1–15. http://dx.doi.org/10.1006/obhd.1995.1001

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67,* 577–580. http://dx.doi.org/10.1037/0021-9010.67.5.577

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50,* 905–925. http://dx.doi.org/10.1111/j.1744-6570.1997.tb01487.x

Johnson, D. S., Turban, D. B., Pieper, K. F., & Ng, Y. M. (1996). Exploring the role of normative- and performance-based feedback in motivational processes. *Journal of Applied Social Psychology, 26,* 973–992. http://dx.doi.org/10.1111/j.1559-1816.1996.tb01120.x

Kane, J. S. (1994). A model of volitional rating behavior. *Human Resource Management Review, 4,* 283–310. http://dx.doi.org/10.1016/1053-4822(94)90016-7

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal, 38,* 1036–1051. http://dx.doi.org/10.2307/256619

Klein, H. J. (1989). An integrated control theory model of work motivation. *The Academy of Management Review, 14,* 150–172.

Klein, W. M. (1997). Objective standards are not enough: Affective, self-evaluative, and behavioral responses to social comparison information. *Journal of Personality and Social Psychology, 72,* 763–774. http://dx.doi.org/10.1037/0022-3514.72.4.763

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284. http://dx.doi.org/10.1037/0033-2909.119.2.254

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65,* 422–427. http://dx.doi.org/10.1037/0021-9010.65.4.422

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852. http://dx.doi.org/10.1177/1094428106296642

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the recent literature. *Personnel Psychology, 67,* 241–293. http://dx.doi.org/10.1111/peps.12052

Lewthwaite, R., & Wulf, G. (2010). Social-comparative feedback affects motor skill learning. *Quarterly Journal of Experimental Psychology, 63,* 738–749. http://dx.doi.org/10.1080/17470210903111839

Lurie, N. H., & Swaminathan, J. M. (2009). Is timely information always better? The effect of feedback frequency on decision making. *Organizational Behavior and Human Decision Processes, 108,* 315–329. http://dx.doi.org/10.1016/j.obhdp.2008.05.005

MacHatton, M. T., Van Dyke, T., & Steiner, R. (1997). Selection and retention of managers in the US restaurant sector. *International Journal of Contemporary Hospitality Management, 9,* 155–160. http://dx.doi.org/10.1108/09596119710185837

Marcus, B. (2003). Attitudes towards personnel selection methods: A partial replication and extension in a German sample. *Applied Psychology, 52,* 515–532. http://dx.doi.org/10.1111/1464-0597.00149

Mayfield, E. C. (1964). The selection interview—A re-evaluation of published research. *Personnel Psychology, 17,* 239–260. http://dx.doi.org/10.1111/j.1744-6570.1964.tb00065.x

Mayfield, E. C., Brown, S. H., & Hamstra, B. W. (1980). Selection interviewing in the life insurance industry: An update of research and practice. *Personnel Psychology, 33,* 725–739. http://dx.doi.org/10.1111/j.1744-6570.1980.tb02365.x

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616. http://dx.doi.org/10.1037/0021-9010.79.4.599

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85,* 812–821. http://dx.doi.org/10.1037/0021-9010.85.5.812

McIntyre, R. M. (1990). Spurious estimation of validity coefficients in composite samples: Some methodological considerations. *Journal of*

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

Applied Psychology, 75, 91–94. http://dx.doi.org/10.1037/0021-9010.75.1.91

Melchers, K. G., Lienhardt, N., von Aarburg, M. V., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology, 64,* 53–87. http://dx.doi.org/10.1111/j.1744-6570.2010.01202.x

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology, 58,* 583–611. http://dx.doi.org/10.1111/j.1744-6570.2005.655.x

Moscoso, S., & Salgado, J. F. (2004). Fairness reactions to personnel selection techniques in Spain and Portugal. *International Journal of Selection and Assessment, 12,* 187–196. http://dx.doi.org/10.1111/j.0965-075X.2004.00273.x

Motowidlo, S. J. (1986). Information processing in personnel decisions. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 4, pp. 1–44). Greenwich, CT: JAI Press.

Mullins, T. W. (1982). Interviewer decisions as a function of applicant race, applicant quality and interviewer prejudice. *Personnel Psychology, 35,* 163–174. http://dx.doi.org/10.1111/j.1744-6570.1982.tb02192.x

Mumford, M. D. (1983). Social comparison theory and the evaluation of peer evaluations: A review and some applied implications. *Personnel Psychology, 36,* 867–881. http://dx.doi.org/10.1111/j.1744-6570.1983.tb00516.x

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational and goal-oriented perspectives*. Newbury Park, CA: Sage.

Neighbors, C., Lewis, M. A., Bergstrom, R. L., & Larimer, M. E. (2006). Being controlled by normative influences: Self-determination as a moderator of a normative feedback alcohol intervention. *Health Psychology, 25,* 571–579. http://dx.doi.org/10.1037/0278-6133.25.5.571

Ng, K. Y., Koh, C., Ang, S., Kennedy, J. C., & Chan, K. Y. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology, 96,* 1033–1044. http://dx.doi.org/10.1037/a0023368

O'Brien, J. O., & Rothstein, M. G. (2011). Leniency: Hidden threat to large-scale, interview-based selection systems. *Military Psychology, 23,* 601–615.

Paunonen, S. V., & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology, 103,* 158–175. http://dx.doi.org/10.1037/a0028165

Peugh, J. L., & Enders, C. K. (2005). Using the SPSS mixed procedure to fit cross-sectional and longitudinal multilevel models. *Educational and Psychological Measurement, 65,* 717–741. http://dx.doi.org/10.1177/0013164405278558

Posthuma, R. A., Levashina, J., Lievens, F., Schollaert, E., Tsai, W., Wagstaff, M. F., & Campion, M. A. (2014). Comparing employment interviews in Latin America with other countries. *Journal of Business Research, 67,* 943–951. http://dx.doi.org/10.1016/j.jbusres.2013.07.014

Posthuma, R. A., Morgeson, F. P., & Campion, M. A. (2002). Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Personnel Psychology, 55,* 1–81. http://dx.doi.org/10.1111/j.1744-6570.2002.tb00103.x

Pulakos, E. D., Schmitt, N., Whitney, D., & Smith, M. (1996). Individual differences in interviewer ratings: The impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Personnel Psychology, 49,* 85–102. http://dx.doi.org/10.1111/j.1744-6570.1996.tb01792.x

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83,* 634–644. http://dx.doi.org/10.1037/0021-9010.83.4.634

Rowe, P. M. (1984). Decision processes in personnel selection. *Canadian Journal of Behavioural Science, 16,* 326–337. http://dx.doi.org/10.1037/h0080865

Sanyal, R., & Guvenli, T. (2005). Personnel selection practices in a comparative setting: Evidence from Israel, Slovenia, and the USA. *Journal of East-West Business, 10,* 5–27. http://dx.doi.org/10.1300/J097v10n04_02

Schmidt, F. L., & Zimmerman, R. D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89,* 553–561. http://dx.doi.org/10.1037/0021-9010.89.3.553

Schmiege, S. J., Klein, W. M. P., & Bryan, A. D. (2010). The effect of peer comparison information in the context of expert recommendations on risk perceptions and subsequent behavior. *European Journal of Social Psychology, 40,* 746–759.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29,* 79–101. http://dx.doi.org/10.1111/j.1744-6570.1976.tb00404.x

Schultz, P. W. (1999). Changing behavior with normative feedback interventions: A field experiment on curbside recycling. *Basic and Applied Social Psychology, 21,* 25–36. http://dx.doi.org/10.1207/s15324834basp2101_3

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Smither, J. W., London, M., Vasilopoulos, N. L., Reilly, R. R., Millsap, R. E., & Salvemini, N. (1995). An examination of the effects of an upward feedback program over time. *Personnel Psychology, 48,* 1–34. http://dx.doi.org/10.1111/j.1744-6570.1995.tb01744.x

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research, 22,* 342–363. http://dx.doi.org/10.1177/0049124194022003004

Taylor, M. S., Fisher, C. D., & Ilgen, D. R. (1984). Individuals' reactions to performance feedback in organizations: A control theory perspective. In K. R. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 2, pp. 81–124). Greenwich, CT: JAI Press.

Thoresen, C. J., Bradley, J. C., Bliese, P. D., & Thoresen, J. D. (2004). The big five personality traits and individual job performance growth trajectories in maintenance and transitional job stages. *Journal of Applied Psychology, 89,* 835–853. http://dx.doi.org/10.1037/0021-9010.89.5.835

Tolli, A. P., & Schmidt, A. M. (2008). The role of feedback, casual attributions, and self-efficacy in goal revision. *Journal of Applied Psychology, 93,* 692–701. http://dx.doi.org/10.1037/0021-9010.93.3.692

Vance, R. J., Kuhnert, K. W., & Farr, J. L. (1978). Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. *Organizational Behavior and Human Performance, 22,* 279–294. http://dx.doi.org/10.1016/0030-5073(78)90017-X

Vancouver, J. B. (2005). The depth of history and explanation as benefit and bane for psychological control theories. *Journal of Applied Psychology, 90,* 38–52. http://dx.doi.org/10.1037/0021-9010.90.1.38

Van Iddekinge, C. H., Sager, C. E., Burnfield, J. L., & Heffner, T. S. (2006). The variability of criterion-related validity estimates among interviewers and interview panels. *International Journal of Selection and Assessment, 14,* 193–205. http://dx.doi.org/10.1111/j.1468-2389.2006.00352.x

Weekley, J. A., & Gier, J. A. (1989). Ceilings in the reliability and validity of performance ratings: The case of expert raters. *Academy of Management Journal, 32,* 213–222. http://dx.doi.org/10.2307/256428

Wiener, N. (1948). *Cybernetics: Control and communication in the animal and the machine*. Cambridge, MA: MIT Press.

tapraid5/z2j-aplpsy/z2j-aplpsy/z2j00116/z2j3030d16z | xppws | S=1 | 1/23/16 | 10:08 | Art: 2014-0208

Wilk, S. L., & Cappelli, P. (2003). Understanding the determinants of employer use of selection methods. *Personnel Psychology, 56*, 103–124. http://dx.doi.org/10.1111/j.1744-6570.2003.tb00145.x

Wood, J. V. (1989). Theory and research concerning social comparisons of personal attributes. *Psychological Bulletin, 106*, 231–248. http://dx.doi.org/10.1037/0033-2909.106.2.231

Wood, J. V. (1996). What is social comparison and how should we study it? *Personality and Social Psychology Bulletin, 22*, 520–537. http://dx.doi.org/10.1177/0146167296225009

Wood, R., & Bandura, A. (1989). Social cognitive theory of organizational management. *The Academy of Management Review, 14*, 361–384.

Yu, J., & Murphy, K. R. (1993). Modesty bias in self-ratings of performance: A test of the cultural relativity hypothesis. *Personnel Psychology, 46*, 357–363.

Zedeck, S., Tziner, A., & Middlestadt, S. E. (1983). Interviewer validity and reliability: An individual analysis approach. *Personnel Psychology, 36*, 355–370. http://dx.doi.org/10.1111/j.1744-6570.1983.tb01443.x

Zikmund-Fisher, B. J. (2004). De-escalation after repeated negative feedback: Emergent expectations of failure. *Journal of Behavioral Decision Making, 17*, 365–379. http://dx.doi.org/10.1002/bdm.478

Appendix

Sample Normative Feedback Report

This form is designed to give feedback to interviewers on their average interview ratings. To maintain anonymity, each interviewer is identified only by identification code.

The following table shows the average rating by interviewer, broken down by period. There are also summary data across periods. For each period, two values are presented. "N" refers to the number of interviews conducted, and "Mean" refers to the average overall rating across those interviews.

The spreadsheet can be easily sorted in different ways. To see how your ratings compare to other interviewers, select the column representing the information of interest and then select the sort icon under the data tab.

	NFI1 period		NFI2 period		NFI3 period		All periods	
ID code	*N*	Mean	*N*	Mean	*N*	Mean	*N*	Mean
01	93	5.67	93	5.64	85	5.53	271	5.61
02			140	5.90	89	5.68	229	5.81
03	14	5.47					14	5.47
04	58	5.42					58	5.42
05			135	5.31	64	5.37	199	5.33
06					59	5.30	59	5.30
07	27	5.22	43	5.14	59	5.25	129	5.21
08	172	5.42	102	5.42	119	5.39	393	5.41
09	101	5.31	57	5.19			158	5.27
10			162	5.03			162	5.03
11	185	5.20	130	5.13	33	5.37	348	5.19
12	48	5.53	29	5.65	37	5.57	114	5.57
13	222	5.43					222	5.43
14	27	5.39					27	5.39
15	50	5.66	27	5.64	78	5.71	155	5.68
16	33	5.67	208	5.73	149	5.67	390	5.70
17			123	4.98	68	5.12	191	5.03
18	40	5.22	37	5.21	77	5.50	154	5.36
19	111	5.55					111	5.55
20					246	5.39	246	5.39
21			141	5.28	49	5.23	190	5.27
22	148	5.37	58	5.34	27	5.33	233	5.36
23			133	5.43	113	5.69	246	5.55
24	213	5.73					213	5.73
25			139	5.11	44	5.24	183	5.15
Overall	1,542	5.45	1,757	5.36	1396	5.43	4695	5.41

Received March 25, 2014
Revision received December 21, 2015
Accepted January 13, 2016 ■