



The International Journal of Human Resource Management

ISSN: 0958-5192 (Print) 1466-4399 (Online) Journal homepage: <https://www.tandfonline.com/loi/rjih20>

The performance of performance appraisal systems: understanding the linkage between appraisal structure and appraisal discrimination complaints

Ellen V. Rubin & Amani Edwards

To cite this article: Ellen V. Rubin & Amani Edwards (2020) The performance of performance appraisal systems: understanding the linkage between appraisal structure and appraisal discrimination complaints, The International Journal of Human Resource Management, 31:15, 1938-1957, DOI: [10.1080/09585192.2018.1424015](https://doi.org/10.1080/09585192.2018.1424015)

To link to this article: <https://doi.org/10.1080/09585192.2018.1424015>



Published online: 11 Jan 2018.



Submit your article to this journal [↗](#)



Article views: 7023



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 9 View citing articles [↗](#)



The performance of performance appraisal systems: understanding the linkage between appraisal structure and appraisal discrimination complaints

Ellen V. Rubin and Amani Edwards

Department of Public Administration and Policy, State University of New York, University at Albany, Albany, NY, USA

ABSTRACT

Employee perceptions of the fairness of performance evaluations are critical to the success of any appraisal system. Research on performance management includes extensive studies on how bias emerges in the appraisal process. Despite this, there is no empirical evidence linking formal discrimination complaint filings – a key measure of bias – and performance appraisals. To close this gap, we conduct an empirical analysis using information on appraisal systems and discrimination complaints from the US federal government. Our findings suggest that agencies with better-designed appraisal systems will experience higher appraisal-related discrimination complaints, contrary to expectations. In particular, an expansion of training and voice opportunities are associated with an increase in appraisal discrimination complaints, while increased differentiation between performers is not consistently associated with complaints. For managers struggling with performance management, it may be better to spend time on improving performance-focused interpersonal communication rather than redesigning appraisal systems.

KEYWORDS

Performance appraisal; discrimination; forced distribution rating systems

Employee performance appraisal systems serve as a key tool for holding individual employees accountable in the public, private, and nonprofit sectors. Appraisals often serve multiple purposes: to provide documentation and justification for future personnel actions, to motivate employees, to identify areas where professional development is needed, and to communicate organizational priorities, to name just a few (Daley, 2004; Levy & Williams, 2004; Murphy & Cleveland, 1995). These multiple purposes often clash, leading to interpersonal conflict.

While the private sector is more actively questioning the utility of employee performance appraisal systems, most government personnel systems require some

sort of annual performance rating. Civil service reforms undertaken by the US states over the last two decades are largely driven by concerns that government needs stronger tools for holding employees accountable (Battaglio & Condrey, 2009; Kellough & Selden, 2003). Many of these reforms focus on making it easier to fire public employees and strengthening the consequences for strong or weak performance. There has also been increasing attention on performance appraisal in the US federal government. Over the last fifteen years, the U.S. Office of Personnel Management (OPM) developed standards for assessing the strength of performance appraisal systems (2006). More recently, OPM announced the Performance Management Plus initiative to focus more attention on the relationship between managers and employees, emphasizing on-going feedback and coaching (US Office of Personnel Management 2016).

Scholars do not fully agree on how to measure the effectiveness of performance appraisal systems. Potential options include the accuracy of ratings and a host of behavioral, attitudinal, and cognitive reactions to ratings, including fairness perceptions (Levy & Williams, 2004; Murphy & Cleveland, 1995). Among the potential measures available, discrimination complaints alleging biased treatment in the performance appraisal process can serve as one of many behavioral signals of fairness perceptions. Discrimination complaints in the appraisal process can suggest any myriad of problems, including inaccurate or biased ratings, inconsistent application of performance standards, and/or interpersonal conflict between the manager and the employee. Based on this, the research presented in this article asks the following question: what is the relationship between performance appraisal structure and the rate at which employees claim discrimination in the appraisal process?

To understand the link between appraisal structure and appraisal discrimination complaints, this analysis combines two sets of data from US federal agencies. Information on agency appraisal structure comes from scores assigned by OPM to agency appraisal systems that were graded under the Performance Assessment and Accountability Tool (PAAT) framework. The PAAT grades are combined with data on agency discrimination complaints, which are reported publically on an annual basis under the requirements of the Notification and Federal Employee Antidiscrimination and Retaliation Act of 2002 (No FEAR Act).

The article begins with a review of the key themes in the research linking performance appraisal and bias. We next explore the literature on specific aspects of performance appraisal system design that might be linked to perceptions of bias. Following the presentation of hypotheses, our data are described in more detail. The study will then turn to our empirical analysis. We conclude by reflecting on the implications of our findings for current research, and outline steps for future studies.

Literature review

A claim of discrimination in the appraisal process is a purposive behavior resulting from a negative perception of the appraisal, and signals an employee believes her appraisal was negatively impacted because of her gender, race, age, disability status, etc. According to the psychology literature, this perception of harm might arise from a rating viewed as inaccurate and/or biased. Both may lead an employee to claim discrimination, and scholarship provides detailed studies on the different manifestations of each.

The perceived accuracy of an appraisal system is based on the evaluation process used by managers to recall and document performance information (Landy, Barnes, & Murphy, 1978; Landy, Barnes-Farrell, & Cleveland, 1980). This process includes the supervisor's 'frequency of evaluation, identification of goals, and the knowledge of their employees' performance and job duties' (Landy et al., 1980, p. 355). To improve accuracy in documenting performance scholars have identified multiple strategies managers can use, including writing notes about performance on a regular basis and storing them in an organized way so they can easily be found for the writing of the appraisal (Feldman, 1981).

Frequent evaluations and interactions with employees will also ensure that both the supervisor and employee are aware of each other's goals, which can affect the appraisal score as well (Wang, Wong, & Kwong, 2010).

A second key reason an employee may file a discrimination claim relating to her performance appraisal is because she believes the rating is inaccurate due to biased treatment by the supervisor. There is a long history of research on the ways in which rater bias distorts ratings based on the employee's race/ethnicity, gender, age, and other factors (Cox & Nkomo, 1986; Heilman & Eagly, 2008; Leslie, King, Bradley, & Hebl, 2008). The stereotype-fit model suggests bias enters into the appraisal rating, or other personnel decisions, when raters attribute stereotypes to an individual ratee (Dipboye, 1985). For example, if a rater attributes an executive to being a white male, then she will most likely promote a white male to that position, even though an ethnic minority is more qualified. Raters evaluate ratees to the extent that their perceptions of the ratee fit the job stereotype, resulting in discrimination (Dipboye, 1985). Landy (2008) moves Dipboye's (1985) stereotype-fit model further and notes many of these stereotypes are negative, particularly when applied to women and minorities. Furthermore, the subjective nature of many personnel decisions, including appraisals, enables bias to emerge. Additionally, organizations oftentimes lack safeguards that prevent biased thinking (Landy, 2008).

There are a number of things organizations can do to design their appraisal systems to increase fairness perceptions by reducing opportunities for inaccuracy or bias. These are discussed in detail below, followed by hypotheses for this study. For example, setting clear expectations for performance allows employees to know what behaviors are most valued and how they should work to achieve

them. Involving employees throughout the appraisal process can serve to increase the level of trust that employees have in the final ratings. Training managers and employees in the requirements of the appraisal process, and educating both on the norms and expectations underlying the system, can help to reduce conflict and ensure standards are applied consistently. Finally, recognizing differences in levels of performance among employees acknowledges the success of strong performers. However, when taken to the extreme, differentiations between performers may reflect bias rather than actual differences in contributions.

Setting clear expectations

Clear performance standards are linked to greater perceptions of appraisal system fairness (Dickinson, 1993; Erdogan, Kraimer, & Liden, 2001; Folger, Konovsky, & Cropanzano, 1992). Clarity is enhanced when employees can see how their work contributes to the accomplishment of organizational goals. Repeated studies in motivation theory consistently show that clear goals exert greater motivational effects than vague 'do your best' goals (Latham, 2012). Setting clear goals in the public sector can be particularly challenging, but also particularly important, because goal ambiguity has been tied to decreased organizational productivity (Chun & Rainey, 2005; Pandey & Wright, 2006; Tubre & Collins, 2000).

One way agencies can set clearer performance goals is by creating a 'line of sight,' explaining to employees how their performance is related to program performance. Similar to setting clear goals, alignment of employee effort to organizational performance helps employees to prioritize work and seek out relevant knowledge needed to achieve the goals (Hitt, Bierman, Shimizu, & Kochhar, 2001; Locke & Latham, 2002). Goal alignment occurs when the objectives and purposes of individuals are aligned with the objectives and purposes of the organization (Ayers, 2013; Rubin, 2015). As alignment between employee appraisal and organizational goals increases, employees indicate higher satisfaction with appraisal systems (Decramer, Smolders, & van der Straeten, 2013). This leads to the following hypothesis:

H1: As the clarity of performance expectations increases, the rate of appraisal discrimination complaints will decrease.

Employee participation

Employee participation in appraisal has long been acknowledged as important: 'Employee participation in appraisal is one of the most widely researched factors ... Research suggests that allowing employees to participate in the appraisal process is associated with positive employee reactions to the appraisal system' (Cawley, Keeping, & Levy, 1998, p. 616). Participation is useful because it gives employees a better understanding of what aspects of performance will be rated,

the standards expected, and increases buy-in to both the content and the rating (Silverman & Wexley, 1984).

Importantly, participation can occur at multiple points in the appraisal process. Depending on the study, participation has been operationalized as participation in system design, goal setting, self-assessment and/or providing examples of performance to the rater, participation in feedback conversations, or exerting influence over the final rating (Cawley et al., 1998; Greenberg, 1986; Roberts, 2003; Roberts & Reed, 1996). Managers who involve employees in setting goals and expectations find that employees understand expectations better, are more confident that they can achieve those expectations, and perform at a higher level (Roberts & Reed, 1996). When considered jointly, employee participation in setting goals, discussing actual performance, and presenting work examples to managers to inform appraisal ratings are all related to procedural justice perceptions of the appraisal process (Greenberg, 1986; Roberts & Reed, 1996). Combined, the research suggests the following hypothesis:

H2: As opportunities for participation in the appraisal process increase, the rate of appraisal discrimination complaints will decrease.

Training

Training on performance appraisal systems should be one piece of developing managerial competencies. Appraisal training provided to raters has received more attention than appraisal training provided to employees. In particular, appraisal training provided to raters generally focuses on improving rating accuracy. A recent meta-analysis finds that rater training improves appraisal accuracy more so than changes to the appraisal instrument, and the most frequently studied rater training technique is frame-of-reference training (Roch, Woehr, Mishra, & Kieszczynska, 2012). Frame-of-reference training uses examples, practice, and feedback to help raters develop shared understandings of performance standards (Arvey & Murphy, 1998; Roch et al., 2012). The training ensures managers will rate similar levels of performance consistently across individual ratees, which has been shown to be a key component of appraisal procedural justice perceptions (Greenberg, 1986). In addition to improving accuracy, training managers in performance appraisal increases satisfaction with the appraisal system, resolution of workplace conflict, and decreases rating distortions (Taylor, Tracy, Renard, Harrison, & Carroll, 1995).

The few studies that examine appraisal training provided to employees describe training specifically as contributing to perceptions of the fairness of the appraisal. These studies argue that training employees helps them to understand how the system works, understand performance standards, and understand what performance information may be relevant to provide to raters to improve accuracy of the rating (Folger et al., 1992). Knowledge of performance appraisal criteria does increase fairness perceptions of the appraisal, including overall appraisal system

fairness, satisfaction with the appraisal system, and satisfaction with the rating (Erdogan et al., 2001; Taylor et al., 1995). Combined, the research on appraisal training for managers and employees suggests:

H3: As training on the performance appraisal systems made available to employees and managers increases, the rate of appraisal discrimination complaints will decrease.

Differentiation between performers

If an organization intends to use its appraisal systems to inform administrative decisions such as pay raises, promotions, or discipline, it is important that the assigned ratings allow managers to differentiate between levels of performance. Unfortunately, decades of empirical evidence indicate that when ratings are used to inform administrative decisions with significant positive or negative consequences, raters tend to inflate ratings, or issue inaccurate ratings more generally (Jawahar & Williams, 1997; Murphy & Cleveland, 1995). To address this, organizations may choose to implement appraisal system policies to encourage differentiation in both the ratings assigned and the consequences of the ratings. Such policies are grounded in expectancy and equity theories. Equity theory suggests individuals will view appraisal decisions as more fair when they receive more positive consequences for high performance (Blume, Baldwin, & Rubin, 2009; Latham, 2012). Similarly, expectancy theory suggests that individuals compare their own performance to their peers. When an individual perceives she is performing at a higher level, but sees her lower-performing peer receive a similar rating and/or consequence, she will view the appraisal system as less fair (Blume et al., 2009; Latham, 2012).

Organizations may care about the distribution of their performance ratings because they may perceive the ratings do not accurately reflect performance. To address this, organizations may implement a relative rating system (as opposed to an absolute rating system), and/or a forced distribution rating system (FDRS). A relative rating system asks managers to rate employees against each other, often placing them into ordered lists, while an absolute rating systems evaluates employee performance against pre-determined elements and standards (Roch, Sternburgh, & Caputo, 2007). Similarly, FDRS requires raters to evaluate employees by sorting them into predetermined performance categories (Dominick, 2009). Scholars are currently debating the positive and negative impact of these rating approaches (as summarized in Dominick, 2009; Moon, Scullen, & Latham, 2016). In the short term, forced rankings may improve employee motivation and task performance, and may attract high performing applicants to join the organization because they see that high performance is recognized (Blume et al., 2009). Conversely, relative rankings under an FDRS scheme may result in lower employee procedural justice perceptions, as compared to absolute rating against pre-defined standards (Roch et al., 2007). Other criticisms suggest raters view forced rankings

as more difficult to use and less fair (Schleicher, Bull, & Green, 2008), and result in ratings that are less accurate (Park, 2014).

In addition to pushing managers to change the distribution of ratings, agencies may choose to link ratings to positive or negative management decisions. To recognize strong performers, organizations typically consider pay-for-performance schemes. Unfortunately pay-for-performance schemes in government rarely deliver the expected increase in motivation and performance. In the US federal government pay-for-performance did not increase public manager motivation and did not improve performance in multiple studies (Perry, Engbers, & Jun, 2009). More recent efforts in US state governments have led to perceptions of favoritism and failed promises (Battaglio & Condrey, 2009; Kellough & Nigro, 2002). Beyond the public sector experience, the use of ratings for administrative decisions may result in biased ratings against a particular group. One study found ratings assigned to women were biased only when those ratings were used to inform merit pay decisions (Dobbins, Cardy, & Truxillo, 1988). Since differentiation between performers may lead to favoritism and/or discrimination, we propose the following hypothesis:

H4: When differentiation between levels of performance increases, the rate of appraisal discrimination complaints will increase.

Data and methods

To analyze the relationship between performance appraisal structure and discrimination complaints resulting from the appraisal process, two agency-level data sets are merged. For the independent variables, the data on federal agency performance appraisal systems was provided by OPM to one of the authors in 2011.¹ For the dependent variable, information on agency discrimination complaints are calculated from agency No FEAR Act reports. Under this law, agencies are required to publically report on an annual basis the number of discrimination complaints they received in the fiscal year, including the basis for the discrimination (was the unfair treatment due to the complainant's gender, race, age, etc.) and the manner in which the individual believes they were harmed (was the individual denied a promotion, were they inappropriately disciplined, was their appraisal unfair, etc.).² Agencies report aggregate complaint rates. We do not know which individual employees filed which complaints, and we cannot determine what portion of appraisal-related discrimination complaints allege bias in the form of race-, gender-, age-based, etc. discrimination. As a result, the unit of analysis is the agency.

Independent variables

OPM evaluates agency performance appraisal systems using ten criteria identified in the PAAT. These criteria include: alignment, results focus, credible measures, distinctions in levels of performance, consequences, employee involvement,

feedback and dialogue, and training. OPM includes two additional criteria related to its responsibilities to audit agency personnel management systems, including organizational assessment and guidance, and oversight and accountability. Based on OPM's definition and measurement of each criterion, the variables used in this analysis were selected based on their fit with the theory presented above.

Each criterion includes multiple questions/components, and OPM assigns specific points to each component of each criterion. Combined, OPM assigns ten points to each of the ten criteria, such that an agency with an outstanding performance appraisal system could earn a high score of 100 points. OPM designates as system that scores 80 points or higher, out of a possible 100 points, as effective.

To document its efforts on each of the ten criteria, agencies answer multiple questions and provide extensive documentation to OPM. OPM assigns scores for each criteria based on their assessment of the evidence provided by the agencies against the detailed criteria. The data-set includes the points assigned to each component of each criterion. For example, the criterion of alignment includes four components, each scored separately, with the maximum potential point total equaling ten points. Additionally, multiple OPM raters evaluated agency PAAT submissions to derive the final score for the assessment of the system. The data-set does not include any notations on inter-rater reliability.

As noted in Rubin (2015), the PAAT data have many limitations. First, use of the PAAT evaluation system is entirely voluntary and is initiated by the agency itself. Many federal agency appraisal systems have not been evaluated via the PAAT and are thus not included in the data-set, while others have been evaluated under the PAAT framework multiple times. For the analysis conducted here, we use the most recent PAAT score for those agencies that were evaluated multiple times. However, given the age of the data it is possible that the scores used in this analysis do not represent the current systems in use. Importantly, the voluntary nature of the PAAT assessment means that we cannot assume that participation in the PAAT review is a completely random process. Second the data-set only includes the scores assigned by OPM for the individual criterion; the data-set does not include the extensive documentation agencies provided to OPM.

Independent variables were derived from the PAAT criteria based on their fit with theory, and informed by confirmatory factor analysis. The variable *clarity* was formed by calculating the sum of the total scores from the PAAT criteria on (a) alignment, (b) results-focus, and (c) credible measurement. Alignment considers the degree to which employee performance plans create a line of sight between individual goals and organizational goals, while the criterion results-focus considers whether employees are actually held accountable for their contributions to organizational goal achievements (OPM, 2006). The criterion on credible measurement considers the degree to which the appraisal program ensures that performance measures for employees are specific, measurable, and objective. We combined these into one variable, called *clarity*, because employees in agencies that

are successful in all three areas should have a better understanding of expectations for their individual performance.

The empirical models operationalize agency efforts to differentiate between levels of performance with multiple measures. As a global measure of differentiation, the variable *distinctions* calculates the sum from the OPM criteria (a) distinctions among levels of performance and (b) consequences based on performance. Agencies are asked to report the distribution of performance ratings for the most recent appraisal period, with more points given to agencies with a broader distribution. To document the consequences for performance appraisal ratings, agencies report the use of both positive and negative consequences resulting from ratings, including cash awards and time off, or the placement of employees into performance improvement periods. For a more fine-grained analysis, we also consider three different forms of performance distinctions separately. Specifically, in an alternative model we drop the global *distinctions* variable and instead include three variables: (a) *ratings* which uses the OPM score assigned for the distribution of performance ratings, with a broader distribution receiving higher points; (b) *rewards* which uses the OPM score assessing the degree to which agencies provide a variety of positive awards for high performance; and (c) *consequences for poor performers*, which includes the OPM PAAT score for the formal consequences imposed for poor performance, with more negative consequences earning an agency more points. Interestingly, in the overall PAAT scoring scheme, agencies earn more points for providing positive recognition than they do for imposing negative consequences.

The variables *voice* and *training* were derived using one criterion from the PAAT for each variable. *Voice* is calculated as the sum of the points assigned by OPM for the criterion employee involvement, which considers the degree to which employees participated in the design of the appraisal program overall, and the level of employee involvement in setting their individual performance standards (OPM, 2006). The variable *training* reflects the sum assigned by OPM for its criterion training. This criteria asks agencies to report whether the appraisal system requires both employees and managers be trained on the appraisal process, and whether they actually received training in the two years prior to PAAT completion.

Confirmatory factor analysis with varimax rotation assessed the combination of PAAT criteria for the variables *clarity* and *distinctions* and supported the creation of the two variables. The two-variable structure explains 78% and 31% of the variance respectively. Additionally, the two variable structure obtained an acceptable fit ($\chi^2(10) = 247.78$ $p < .01$). The variable *clarity* produced a Cronbach's alpha scale reliability coefficient of .92. The variable *consequences* produced a Cronbach's alpha scale reliability coefficient of .73. Both variables have a high degree of consistency within the model. We initially considered combining the OPM criteria of (a) feedback and dialogue and (b) employee involvement into a composite voice variable, but the factor structure did not support such a combination. As a result, *voice* is assessed using the single OPM criterion employee

involvement because OPM's definition of this criterion is more consistent with the theory discussed previously.

Dependent variables

For the dependent variables, information on agency discrimination complaints is calculated from agency No FEAR Act reports. Since 2003, agencies are required to report annually on the bases and issues associated with discrimination complaints. These reports include information on the number of discrimination complaints alleging biased treatment in the performance appraisal process, in addition to the total complaints filed during the year, and whether employees claimed they were discriminated against based on their gender, race, etc. (for more information see Rubin & Kellough, 2012).

The dependent variable, developed from the No FEAR Act data, is a ratio calculated by dividing the number of appraisal discrimination complaints by the number of employees in that agency. The number of employees was gathered from OPM Fedscope data, using the September employee counts. We use a ratio instead of the number of appraisal discrimination complaints because 50 complaints in a small agency like the National Science Foundation means something very different than in a large agency like the Veterans Health Administration. With both the complaint and appraisal system data coming from agency reports, our level of analysis is the federal agency; there are 79 agencies in the sample.

In addition to the ratio of complaints to employees we develop two alternatives dependent variables for assessing the impact of appraisal system design on discrimination complaints related to performance appraisals. The ratio is recoded as a dummy variable with an agency receiving a score of 0 if it experienced no discrimination complaints tied to appraisal, or scored as 1 if it received any appraisal-related discrimination complaints. A third dependent variable is developed where an agency receives a score of zero if it experienced no complaints; a score of 1 if its complaint rate was greater than zero but equal to or less than the median complaint rate; or a score of 2 if it experienced a complaint rate greater than the median. Based on these independent variables, we analyze our hypothesized relationships using (1) Ordinary Least Squares (OLS) regression using the ratio dependent variable, (2) a logit model with the dummy dependent variable, and (3) a multinomial logit model with the three-category dependent variable. The logit and multinomial logit modes are interpreted using predicted probabilities.

It is critical to describe how we manage time in our analysis. Agencies completed the PAAT in different years and at different intervals. As noted above, this analysis only includes the most recent PAAT scores. In the models, PAAT scores, developed in year t , were linked with discrimination complaints in year $t + 1$. As a result the dates for the dependent variables differ across agencies, but follow a similar structure. For example, the 2008 PAAT score from the Federal Highway Administration is paired with discrimination data from 2009; the 2009 Patent and

Trademark Office PAAT data are paired with discrimination data from 2010, etc. As a result, this allows us to avoid problems associated with same source bias and interpret any relationships as: a structural component of the appraisal system in year x causes an increase/decrease in discrimination complaints in the next year.

Analysis

The analysis begins with a review of the descriptive statistics for our key variables (see Table 1). Agencies included in the sample average approximately 15,000 employees. As expected, there is a very small portion of federal employees filing discrimination complaints tied to the appraisal process. Twenty-one agencies reported no discrimination complaints tied to appraisals. This is not surprising; this simply means that employees filing complaints indicated the unfair treatment happened in other parts of the personnel process. The descriptive statistics in Table 1, specifically the standard errors and the differences between the mean and median value of the dependent variable, indicates this variable is highly skewed, with a limited number of extremely high values. We chose to not drop these observations from the sample so that we would not be ‘selecting on the dependent variable’ and thus biasing our results. Instead, to manage this skew in the dependent variable, we use robust standard errors in all models. In the OLS model we transform all variables by calculating the natural log.

The total PAAT score, assigned by OPM, could have a maximum potential value of 100 points, where OPM defines a successful appraisal system as one that achieves a score of 80 points or higher. In this sample, the average PAAT score is 64 points. Less than half of the sample (33 agencies) earn a PAAT score of 80 points or higher. Within the PAAT scores, we see interesting variations between the different aspects of appraisal system design. On average, agency systems are strongest on providing voice opportunities, i.e. involving employees throughout the appraisal process. Not surprisingly, agencies struggle most in making distinctions between employees. Agency efforts to make distinctions earn an average of 12 out of 20 points.

The bivariate analyses suggest minimal support for the hypotheses (see Table 2). The difference of means test indicates there is no difference in complaint rates

Table 1. Descriptive statistics.

	Mean	Std. Dev.	Median	Min.	Max.	<i>N</i>
Appraisal complaints/empl.	.0008	.0009	.0005	0	.0053	79
Clarity	17.620	9.110	19	0	30	79
Voice	7.614	2.431	8	0	10	79
Training	6.237	2.910	7	1	10	79
Distinctions	11.536	3.752	12	0	17	79
Ratings	5.592	2.273	6	0	8	79
Awards	3.163	1.044	3	0	4	79
Conseq. for poor perf.	1.740	.649	2	0	2	79
Total OPM PAAT score	63.883	22.108	71.4	12	93	79
Agency employees	15,051	33,568	3,106	46	251,392	79

Table 2. Correlations.

	Appr compl/ empl	Clarity	Distinctions	Voice	Training
Appr compl/empl	1				
Clarity	.045	1			
Distinctions	.019	.492*	1		
Voice	.151	.633*	.381*	1	
Training	.243*	.668*	.295*	.680*	1

Notes: All correlations calculated with variables transformed using the natural log.

* $p < .05$.

between those agencies rated as below OPM appraisal standards, and those agencies meeting or exceeding OPM's target of 80 points on the PAAT ($t = -.834$). The correlations also suggest weaknesses in the hypotheses. Specifically, the dependent variable appraisal complaints per employees exhibits a significant and positive correlation only with the training score. This is in the opposite direction hypothesized, but suggests that when employees understand the appraisal system they are better able to judge when they have been treated unfairly.

The results of the OLS models are very modest; they explore only a small amount of variation (see Table 3). We empirically assess three different OLS models. Model 1 assesses the relationship between the ratio of discrimination complaints filed and *clarity*, *voice*, and *training* only. These three independent variables reflect key attributes of the appraisal system before ratings are assigned, and as a result, any relationship with the dependent variable could easily be described as a procedural justice outcome. Model 2 adds the global distinctions measure, *consequences*, to assess the added impact of appraisal system outcomes on discrimination complaints. Finally, Model 3 drops *consequences* and instead includes the more detailed three measures of *rewards*, *ratings*, and *consequences for poor performers*. A Wald test is used to evaluate whether these three distinction variables are jointly equal to zero, i.e. whether they improve upon Model 1. We additionally perform two likelihood ratio tests to determine whether Model 2 is better than Model 1, and whether the more detailed Model 3 is better than Model 1.

Table 3. OLS regression models.

	Model 1 OLS	Model 2 OLS	Model 3 OLS
Clarity, ln	.000 [.000]	.000 [.000]	.000 [.000]
Voice, ln	.000 [.000]	.000 [.000]	.000 [.000]
Training, ln	.001** [.000]	.001** [.000]	.001** [.000]
Distinctions, ln		.000 [.000]	
Awards, ln			-.001* [.000]
Rating distrib, ln			.000 [.000]
Conseq for low perf, ln			.000 [.000]
Adjusted R^2	.05	.04	.03
n	79	79	79
F -Statistic	3.38	3.16	3.51
Probab > F	.022	.019	.004

Notes: Huber/White robust standard errors presented in brackets.

* $p < .10$; ** $p < .05$.

The variable *training* is the only variable that is significant across all three models, while the availability of rewards is significant only in the third model. In Model 2, *training* remains significant even when *consequences* are accounted for in the model. The results of Model 3 are similar to Model 1 even when measured against the distribution of performance ratings, awards, and consequences for poor performers. In all three models, as training increases, the probability of an employee filing an appraisal complaint increases by nearly one percent. In Model 3, a one percent increase in performance-related rewards is associated with a one percent decrease in appraisal-related discrimination complaints. Despite the small adjusted *R*-squares of each of the three models, the F-statistics indicate the independent variables are better at predicting appraisal-related discrimination complaints than random chance. Post-estimation tests indicate multicollinearity is not a problem in any of the three models (Model 1 VIF = 2.12, Model 2 VIF = 2.02, Model 3 VIF = 2.31). The models exhibit a normal distribution, with a kurtosis of 0. The two likelihood ratio tests compared Model 1 to Model 2, and Model 1 to Model 3. The test results indicate that no one model is better than the others (Model 3 vs. Model 1: $p > \chi^2 = .560$) (Model 2 vs. Model 1: $p > \chi^2 = .9490$).

As an alternative way to evaluate the relationship between appraisal discrimination complaint rates and appraisal system design, we recoded the dependent variable into three categories and use a multinomial logit analysis to consider the probability of an agency experiencing different levels of appraisal-related discrimination complaints. The base category is zero, so the model assesses the probability of observing high or low complaint rates versus no complaints. As noted previously, the dependent variable was recoded using the median, with an agency receiving a score of zero if it experienced no appraisal complaints, a score of one as a low complaint rate, defined as greater than zero but less than or equal to the median, and category two as a high complaint rate, with complaint rates above the median. From the recoding based on the median, 27% of agencies in the sample have zero complaints, the percent of agencies with low complaints is 33, and agencies with high complaints is 41%. For the sake of consistency, the multinomial logit model included all the variables used in Model 3 to streamline the discussion of results (see Models 4a and 4b in Table 4). The multinomial logit model exhibits a Wald statistic of 26.94 and an adjusted count *R*-squared of .26.

Predicted probabilities for multinomial logit models allow us to determine the probability of observing a particular outcome (zero, low, or high complaints) when the independent variables are set at specific values (see Table 5). An agency with an average appraisal system has a .561 probability of experiencing high appraisal-related discrimination complaints, and a .209 probability of experiencing zero appraisal-related discrimination complaints. In comparison, an agency with a strong appraisal system has a .072 probability of experiencing zero complaints, compared to a .644 probability of experiencing high complaints. When comparing probabilities an agency is more likely to receive high complaints, rather than zero or low complaints, for all three scenarios tested. Additionally, an agency has

Table 4. Multinomial logit and logit models.

	Model 4a	Model 4b	Model 5
	Multinomial logit	Multinomial logit	Logit
	Zero vs. low compl.	Zero vs. high compl.	
Clarity	.013 [.063]	-.043 [.048]	-.033 [.046]
Voice	.124 [.191]	.133 [.170]	.165 [.156]
Training	.293 [.206]	.363** [.143]	.350** [.144]
Awards	-1.073* [.624]	-.826** [.358]	-.842** [.386]
Rating distrib	.344 [.263]	.374* [.206]	.362* [.201]
Conseq for low perf	-.834 [.769]	-.581 [.691]	-.522 [.660]
n	79	79	79

Notes: Huber/White robust standard errors presented in brackets.

* $p < .10$; ** $p < .05$.

Table 5. Predicted probabilities: appraisal systems.

	MNLogit	MNLogit	MNLogit	Logit
	Prob. of zero complaints	Prob. of low complaints	Prob. of high complaints	Prob. of any complaints
Average appraisal system (variables set at means)	.209	.230	.561	.787
Strong appraisal system (variables set at maximum values)	ns	.284	.644	.933
Weak appraisal system (variables set at minimum values)	.286	.349	.365	.558

Note: ns indicates the predicted probability is not statistically significant at the $p < .05$ level.

Table 6. Predicted probabilities: individual appraisal characteristics.

		MNLogit	MNLogit	MNLogit	Logit
		Prob. of zero complaints	Prob. of low complaints	Prob. of high complaints	Prob. of any complaints
Clarity	High	ns	.336	.406	.710
	Low	ns	ns	.754	.870
Voice	High	.162	.240	.598	.846
	Low	ns	ns	ns	ns
Training	High	ns	ns	.709	.933
	Low	.637	ns	ns	.345
Awards	High	.359	ns	.481	.646
	Low	.014	ns	ns	.982
Rating distrib	High	ns	.249	.652	.899
	Low	.671	ns	ns	ns
Consequences	High	.239	.211	.550	.764
	Low	ns	ns	.564	.902

Notes: ns indicates the predicted probability is not statistically significant at the $p < .05$ level.

High indicates the independent variable is set at its maximum value; low is set at its minimum value.

a greater probability of receiving zero complaints with a weak appraisal system. Conversely, an agency has a greater probability of receiving a high rate of complaints with a strong appraisal system.

The predicted probabilities for individual independent variables provide additional detail (see Table 6). To develop the statistics presented in table 6, each variable was set at its highest or lowest value, holding all other variables at their means. When reviewing these findings in table 6, it is first worth noting that individual independent variables set at their minimum values rarely exhibit statistically significant probabilities. High rates for voice, training, awards, and rating distributions all are associated with higher probabilities of experiencing high complaint rates. This is contrary to the hypotheses. On the other hand agencies experiencing low clarity and consequences have a greater probability of experiencing high complaint rates, as compared to having high levels of clarity and consequences. This is consistent with the hypotheses.

For a third approach to analyzing the relationship between appraisal-related discrimination complaints and the design of appraisal systems, we recoded the dependent variable into a dummy variable, with a score of zero being assigned to agencies with zero discrimination complaints, and a score of 1 assigned to agencies with any appraisal-related discrimination complaints. Seventy-three percent of agencies received some complaints, whereas 27% of agencies received zero complaints, as noted previously. The relationship between discrimination complaints and appraisal system design is assessed with a logit model that includes all the variables used in the OLS Model 3 for the sake of streamlined interpretation. The logit model exhibits a Wald statistic of 25.41 and an adjusted count *R*-squared of .24. The sign and significance in the logit model (Model 5 in Table 4) is consistent with the results in the zero vs. high complaints multinomial logit model (Model 4b in Table 4).

Again, predicted probabilities indicate that stronger appraisal systems generally will experience appraisal-related discrimination complaints (see the last column in Table 5). Specifically an agency with an average appraisal system has a .79 probability of receiving some complaints, whereas an agency with a weak appraisal system has a probability of .56 of receiving some appraisal-related discrimination complaints. In comparison, an agency with a strong appraisal system has a .93 probability of receiving some discrimination complaints. The predicted probabilities of the individual independent variables present additional information (see Table 6). Agencies with appraisal systems rated as exhibiting low clarity, low awards, and/or low consequences each exhibit a higher probability of experiencing some discrimination complaints, compared to high rates of clarity, awards, and/or consequences. This is consistent with the hypotheses, and the clarity and consequences findings are consistent with the results in the multinomial logit model. High voice is associated with a significant probability of an agency receiving any discrimination complaints, which is not consistent with the hypothesis, but matches the multinomial logit findings. Agencies that extensively implement training in their appraisal systems realize a significant probability of an agency receiving any discrimination complaints, which is not consistent with the hypothesis, but matches both the multinomial logit and OLS findings. Finally, agencies with a high ratings distribution have a greater probability of receiving

appraisal-related discrimination complaints, which is consistent with the hypothesis, and generally consistent with the multinomial logit findings.

Discussion and conclusion

Existing scholarship suggests that well-designed performance appraisal systems will increase employee perceptions of fairness. If we accept that appraisal-related discrimination complaints are indicators of fairness perceptions, the results in this study suggest that organizations with well-designed appraisal systems will experience more discrimination complaints, contrary to existing literature. Across all three models, increases in training are associated with higher complaint rates, which contradicts existing research and the proposed hypotheses. Furthermore, the multinomial logit and logit model results generally suggest that stronger appraisal systems experience higher appraisal-related discrimination complaints. Voice, the most frequently tested attribute of appraisal systems, was not significant in the OLS model, but in both the multinomial logit and logit models higher voice opportunities was associated with greater complaint rates, contrary to existing research. Clarity, which assesses the degree to which employees understand performance expectations, was not significant in the OLS model, but in both the multinomial logit and logit models low clarity was associated with higher probability of experiencing complaints, which is consistent with the hypothesis.

The effects of making distinctions in levels of performance on discrimination complaints was inconsistent across the models. Awards exhibited different effects on complaints between the OLS (negative), multinomial logit (positive), and logit (negative) models. Agencies with broader distributions in ratings were more likely to experience complaints, generally, and higher complaint in the logit and multinomial logit models, respectively, which is consistent with the hypotheses. Finally a low rate of consequences for poor performers is associated with greater probabilities of high complaints and experiencing any complaints, which is consistent with the hypothesis.

Psychology scholars have shifted research on appraisals to focus more on the interpersonal and contextual settings in which performance appraisal is implemented. These studies suggest that the relationships between raters and ratees are significantly important for perceptions of the appraisal process, but this shift in the research does not suggest that the structure of the appraisal system is completely disconnected from these perceptions. While we did not expect large relationships, the weakness of the associations was unexpected.

As far as we have been able to determine in our review of existing scholarship, there is no standard expectation for how much the appraisal system design should explain effectiveness of the system, versus psychological factors. The results may be due to measurement problems and the small sample. Seventy-nine observations is more than adequate for OLS regression models, but is low for multinomial logit and logit models. Additionally, bias in sample selection may influence the

results. It is difficult to determine how many performance appraisal programs exist in the US federal government, so it is currently not possible to predict why one agency would submit a completed PAAT. Despite these problems, we should be cautious in dismissing these results too quickly. The sample includes agencies with a variety of missions, occupations, and a broad distribution in employee numbers. Two-fifths of the sample agencies met or exceed OPM's standard for a sophisticated appraisal system.

Moving forward, the results present a number of interesting challenges for scholars interested in understanding how performance appraisal systems impact agency culture, especially the culture of fairness in an agency. First, scholars have identified multiple measures for assessing the effectiveness of appraisal systems. One of these many measures are perceptions of fairness of the system, and the filing of discrimination complaints is a narrow behavioral outcome of those fairness perceptions. We might learn more by going into an agency with multiple appraisal systems and evaluating a broader collection of effectiveness measures. It should not be difficult to find a US federal agency with multiple appraisal systems; appraisal systems are included in collective bargaining agreements, so an agency with multiple bargaining units is likely to have interesting variations in system design. Second, if we are to believe the results that structure is minimally related to outcomes, it would be interesting to assess appraisal effectiveness in two agencies that appear to have the same appraisal system design. If the system is the same, but outcomes are different, we can infer that implementation, context, and interpersonal relationships matter more than design. Finally, as agencies implement new appraisal systems, it could be instructive to know whether a new design improves motivation, fairness, citizenship behaviors, commitment, etc.

Under the PAAT scheme, OPM is sending signals that US federal agencies should make larger differentiations when rating employees, and that ratings should have more frequent and stronger consequences for employees. These efforts are leading to significant frustrations in agencies that have explicitly adopted pay-for-performance systems which require differentiation in ratings. For example, the US Congress has exerted significant pressure on the Senior Executive Service to ensure that very few senior managers are assigned a rating of excellence. Widespread dissatisfaction with the appraisal and pay-for-performance system, and early signals that the ratings distribution was discriminatory, played a large part in the repeal of the US Department of Defense civilian personnel reforms. Finally, attempts to implement a pay-for-performance scheme in the US Securities and Exchange Commission and the US Consumer Financial Protection Bureau also failed due to systematic discrimination against subgroups of employees. As agencies respond to pressure from OPM, Congress, and others to revise appraisal systems, the results here suggest we might benefit from asking the question: should we focus on reforming the design of the appraisal systems, or should we focus on how the systems are implemented instead? The results here suggest we might be better off focusing on the latter and not the former.

Notes

1. We requested more recent data but OPM has thus far not responded to our request.
2. For an example of a No FEAR Act report please visit: US Department of Treasury (2017). *No FEAR Act*. Accessed from <https://www.treasury.gov/No-Fear-Act/Pages/default.aspx>

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, 49(1), 141–168.
- Ayers, R. S. (2013). Building goal alignment in federal agencies' performance appraisal programs. *Public Personnel Management*, 42(4), 495–520.
- Battaglio, R. P., & Condrey, S. E. (2009). Reforming public management: Analyzing the impact of public service reform on organizational and managerial trust. *Journal of Public Administration Research and Theory*, 19(4), 689–707.
- Blume, B. D., Baldwin, T. T., & Rubin, R. S. (2009). Reactions to different types of forced distribution performance evaluation systems. *Journal of Business and Psychology*, 24(1), 77–91.
- Cawley, B. D., Keeping, L. M., & Levy, P. E. (1998). Participation in the performance appraisal process and employee reactions: A meta-analytic review of field investigations. *Journal of Applied Psychology*, 83(4), 615–633.
- Chun, Y., & Rainey, H. (2005). Goal ambiguity and organizational performance in U.S. federal agencies. *Journal of Public Administration Research and Theory*, 15(4), 529–557.
- Cox, T., & Nkomo, S. (1986). Differential performance appraisal criteria: A field study of black and white managers. *Group & Organization Management*, 11(1–2), 101–119.
- Daley, R. (2004). Designing effective performance appraisal systems. In S. Condrey (Ed.), *Handbook of human resource management in government* (2nd ed., pp. 499–527). San Francisco, CA: Jossey-Bass.
- Decramer, A., Smolders, C., & van der Straeten, A. (2013). Employee performance management culture and system features in higher education: Relationship with employee performance management satisfaction. *The International Journal of Human Resource Management*, 24(2), 352–371.
- Dickinson, T. L. (1993). Attitudes about performance appraisal. In S. Schuler, J. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 141–161). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Dipboye, R. (1985). Some neglected variables in research on discrimination in appraisals. *Academy of Management Review*, 10(1), 116–127.
- Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: A laboratory and field study. *Journal of Applied Psychology*, 73(3), 551.
- Dominick, P. G. (2009). Forced rankings: Pros, cons, and practices. In J. Smither & M. London (Eds.), *Performance management: Putting research into action* (pp. 411–443). San Francisco, CA: Jossey Bass.

- Erdogan, B., Kraimer, M. L., & Liden, R. C. (2001). Procedural justice as a two-dimensional construct: An examination in the performance appraisal context. *The Journal of Applied Behavioral Science*, 37(2), 205–222.
- Feldman, Jack M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66(2), 127–148.
- Folger, R., Konovsky, M. A., & Cropanzano, R. (1992). A due process metaphor for performance appraisal. *Research in Organizational Behavior*, 14, 129–129.
- Greenberg, J. (1986). Determinants of perceived fairness of performance evaluations. *Journal of Applied Psychology*, 71, 340–342.
- Heilman, M. E., & Eagly, A. H. (2008). Gender stereotypes are alive, well, and busy producing workplace discrimination. *Industrial and Organizational Psychology*, 1, 393–398.
- Hitt, M., Bierman, L., Shimizu, K., & Kochhar, R. (2001). Direct and moderating effects of human capital on strategy and performance in professional service firms: A resource-based perspective. *Academy of Management Journal*, 44(1), 13–28.
- Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, 50(4), 905–925.
- Kellough, J. E., & Nigro, L. G. (2002). Pay for performance in Georgia state government employee perspectives on GeorgiaGain after 5 years. *Review of Public Personnel Administration*, 22(2), 146–166.
- Kellough, J. E., & Selden, S. (2003). The reinvention of public personnel administration: An analysis of the diffusion of personnel management reforms in the states. *Public Administration Review*, 63(2), 165–176.
- Landy, F. J. (2008). Stereotypes, bias, and personnel decisions: Strange and stranger. *Industrial and Organizational Psychology*, 1(4), 379–392.
- Landy, F. J., Barnes, J. L., & Murphy, K. R. (1978). Correlates of perceived fairness and accuracy of performance evaluation. *Journal of Applied Psychology*, 65, 751–754.
- Landy, Frank J., Barnes-Farrell, J. L., & Cleveland, J. N. (1980). Perceived fairness and accuracy of performance evaluation: A follow-up. *Journal of Applied Psychology*, 65(3), 355–356.
- Latham, G. (2012). *Work motivation: History, theory, research, and practice* (2nd ed.). Thousand Oaks, CA: Sage.
- Leslie, L. M., King, E. B., Bradley, J. C., & Hebl, M. R. (2008). Triangulation across methodologies: All signs point to persistent stereotyping and discrimination in organizations. *Industrial and Organizational Psychology*, 1, 399–404.
- Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, 30(6), 881–905.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9), 705.
- Moon, S. H., Scullen, S. E., & Latham, G. P. (2016). Precarious curve ahead: The effects of forced distribution rating systems on job performance. *Human Resource Management Review*, 26(2), 166.
- Murphy, K., & Cleveland, J. (1995). *Understanding performance appraisal: Social organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage Publications.
- Pandy, S., & Wright, B. (2006). Connecting the dots in public management: Political environment, organizational goal ambiguity, and the public manager's role ambiguity. *Journal of Public Administration Research and Theory*, 16(4), 511–532.
- Park, S. (2014). Motivation of public managers as raters in performance appraisal developing a model of rater motivation. *Public Personnel Management*, 43(4), 387–414.
- Perry, J. L., Engbers, T. A., & Jun, S. Y. (2009). Back to the future? Performance-related pay, empirical research, and the perils of persistence. *Public Administration Review*, 69(1), 39–51.

- Roberts, G. E. (2003). Employee performance appraisal system participation: A technique that works. *Public Personnel Management*, 32(1), 89–98.
- Roberts, G. E., & Reed, T. (1996). Performance appraisal participation, goal setting and feedback the influence of supervisory style. *Review of Public Personnel Administration*, 16(4), 29–60.
- Roch, S. G., Sternburgh, A., & Caputo, P. (2007). Absolute versus relative performance rating formats: Implications for fairness and organizational justice. *International Journal of Selection and Assessment*, 15, 302–316.
- Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczyńska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, 85(2), 370–395.
- Rubin, E. (2015). Holding employees accountable for the accomplishment of organizational goals in the U.S. federal government. *International Journal of Organization Theory and Behavior*, 18(1), 76–105.
- Rubin, E., & Kellough, J. (2012). Does civil service reform affect behavior? Linking complaints, alternative personnel systems, and procedural justice perceptions. *Journal of Public Administration Research and Theory*, 22(1), 121–141.
- Schleicher, D. J., Bull, R. A., & Green, S. G. (2008). Rater reactions to forced distribution rating systems. *Journal of Management*, 35(4), 899–927.
- Silverman, S. B., & Wexley, K. N. (1984). Reaction of employees to performance appraisal interviews as a function of their participation in rating scale development. *Personnel Psychology*, 37(4), 703–710.
- Taylor, M. S., Tracy, K. B., Renard, M. K., Harrison, J. K., & Carroll, S. J. (1995). Due process in performance appraisal: A quasi-experiment in procedural justice. *Administrative Science Quarterly*, 40(3), 495–523.
- Tubre, T., & Collins, J. (2000). Jackson and Schuler (1985) revisited: A meta-analysis of the relationships between role ambiguity, role conflict, and job performance. *Journal of Management*, 26(1), 155–169.
- U.S. Office of Personnel Management. (2006). *System Audit tool: Performance appraisal assessment tool*. Washington, DC Retrieved April 2010 from <http://www.opm.gov/perform/PAAT/GS-PAAT-Instructions.pdf>
- U.S. Office of Personnel Management. (2016). *Performance management guidance for end-of-fiscal-year 2015 and beginning-of-fiscal year 2016 activities*. Washington, DC Retrieved February 2016 from <https://www.chcoc.gov/content/performance-management-guidance-end-fiscal-year-2015-and-beginning-fiscal-year-2016>
- Wang, X. M., Wong, K. E., & Kwong, J. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology*, 95(3), 546–561.