Differentiation of Illusory and True Halo in Writing Scores

Educational and Psychological
Measurement

2015, Vol. 75(1) 102–125

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414530990

epm.sagepub.com

Emily R. Lai¹, Edward W. Wolfe¹, and
Daisy Vickers²

Abstract

This report summarizes an empirical study that addresses two related topics within the context of writing assessment—illusory halo and how much unique information is provided by multiple analytic scores. Specifically, we address the issue of whether unique information is provided by analytic scores assigned to student writing, beyond what is depicted by holistic scores, and to what degree multiple analytic scores assigned by a single rater display evidence of illusory halo. To that end, we analyze student responses to an expository writing prompt that were scored by six groups of raters—four groups assigned single analytic scores, one group assigned multiple analytic scores, and one group assigned holistic scores—using structural equation modeling. Our results suggest that there is evidence of illusory halo when raters assign multiple analytic scores to a single student response and that, at best, only two factors seem to be distinguishable in analytic writing scores assigned to expository essays.

Keywords

illusory halo, trait scores, analytic scores, holistic scores, writing assessment, performance assessment

Student responses to writing assessment prompts are commonly scored using rubrics that assign either a single holistic score of writing quality or a set of analytic scores that depict the quality of each of several traits. In many contexts, raters who are not

¹Pearson, Iowa City, IA, USA

²Pearson, Carrboro, NC, USA

Corresponding Author:

Emily R. Lai, Pearson, 2510 North Dodge Street, Iowa City, IA 52245, USA.

Email: emily.lai@pearson.com

the student's classroom teacher are trained to use a scoring rubric to score student responses to writing prompts. After training and perhaps a qualification process, those raters are monitored to confirm that they apply the rubric as intended. Under a holistic scoring model, raters evaluate the writing according to its overall quality by jointly considering multiple facets of the student response. A key advantage of holistic scores is the fact that they can be assigned relatively quickly (Klein et al., 1998). However, holistic scores may exhibit inconsistencies because individual raters may weight the multiple facets differently. Another shortcoming of holistic scores is the fact that those scores provide relatively little diagnostic information to students concerning the basis of the score or how to improve the writing.

Under an analytic scoring model, raters are trained to evaluate different facets of the writing and to assign a separate score to each trait (e.g., mechanics, organization, voice, and development). These multiple decisions require more time for analytic scoring relative to holistic scoring. However, the focus on a separate facet of writing in each analytic score should, in theory, result in better agreement between raters because of the reduction of subjective weighting of those facets on the part of each rater. In addition, analytic scores should provide more useful diagnostic information to students when compared with holistic scores. However, both of these benefits—focused scoring and diagnostically useful information—can only be realized if raters accurately apply the distinct analytic scoring rubrics to student responses. Our study focuses on this issue.

Raters may incorrectly apply analytic rubrics for two reasons. First, raters may allow irrelevant features of the essay (e.g., handwriting) to influence each analytic score. As a result, the assigned scores are not pure indicators of each trait, and the errors introduced into the scores will be correlated due to the common irrelevant feature influencing all of those scores. Second, raters may be unable to distinguish the characteristics of the essay that are indicators of the qualities depicted by the multiple analytic rubrics. For example, a rater may not understand the difference between the writer's use of vocabulary and tone (i.e., voice) and the use of detail and perspective (i.e., development) so that scores assigned to a rubric intended to measure the writer's use of voice become confused with scores assigned based on a storytelling rubric. In either of these cases, scores on each trait do not purely represent the intended construct. That is, each assigned score depicts something other than the unique trait characterized by the distinct rubrics, and this error introduces unintended overlap between the multiple analytic scores that are assigned to a particular student. These rating errors are commonly referred to as a *halo effect*, and the purpose of this study is to address the following research questions: To what degree do multiple analytic scores assigned by a single rater contain evidence of a halo effect, and what unique information do analytic scores assigned to student writing provide beyond that depicted by holistic scores?

Holistic and Analytic Scoring

Previous research concerning holistic and analytic scoring has focused on three issues: (a) determining the impact of score type on score quality, (b) identifying the

implications of scoring models on data collection design and rater selection, and (c) determining the uniqueness of multiple analytic scores relative to holistic scores. Research that has compared the quality of holistic and analytic scores has produced mixed results. While some studies support the notion that holistic scores produce higher interrater agreement and reliability (Barkaoui, 2007; Johnson, McDaniel, & Willeke, 2000; Schoonen, 2005), others suggest that analytic scores may be more reliable (Klein et al., 1998). Although analytic scores have been shown to be less prone to some types of rater effects, such as rater severity (Chi, 2001), those scores may be tainted by halo effects when the same rater assigns multiple analytic scores (Singer & LeMahieu, 2011). In addition, prior research has made it clear that holistic and analytic scoring models engage raters in different cognitive processes and that these differences have implications for data collection designs and rater selection. A study by Singer and LeMahieu (2011) employed think aloud protocols to reveal the thought processes of raters who assigned both holistic and analytic scores. Their results indicated that raters have an easier time differentiating analytic scores when they are asked to first assign holistic scores and then assign analytic scores. Doing so creates more independence between the analytic and holistic scores (i.e., less possibility of a halo effect).

Finally, concerning the uniqueness of multiple analytic scores compared to holistic scores, most studies have indicated that analytic scores may provide a limited amount of information beyond what is provided by holistic scores. Although Carr (2000) concluded that holistic and analytic scores are qualitatively distinct among English as Second Language students, most other studies have indicated that the correlations among analytic scores are too high to support the intended distinctions (Aryadoust, 2010; Bacha, 2001; Lee, Gentile, & Kantor, 2008). Other research, which has directly compared the information in analytic scores with that in holistic scores, indicates that these two types of scores are very highly correlated (Klein et al., 1998; Lee et al., 2008). More detailed analyses suggest that it may only be useful to distinguish mechanics from a composite score that depicts other qualities of writing (e.g., organization, vocabulary, language, and development; Bacha, 2001; Lee et al., 2008). This fact suggests several possible factorial models for writing: (a) a unidimensional model in which the analytic traits constitute multiple measures of a single dimension, (b) a two-dimensional model in which mechanics constitutes one dimension while the remaining traits constitute the other, or (c) a model that maintains a unique dimension for each of the traits tapped by the multiple analytic rubrics.

From the existing literature, it is not clear that either of the supposed benefits attributed to analytic scores are realized in operational settings. Specifically, it is not apparent whether raters who assign analytic scores are better able to focus on distinct and relevant essay characteristics, thus assigning scores of higher quality. Neither is it evident that analytic scores are statistically distinguishable from their holistic counterparts. Unfortunately, shortcomings in data collection designs in research regarding these topics may preclude determining whether analytic scores are too highly correlated to warrant their added expense. Specifically, much of the extant research that

compares analytic and holistic scores that are intended for uses external to the classroom confounds trait correlations with halo effects because of the fact that the same raters assigned holistic and multiple analytic scores. That is, most studies that examine the relationship between multiple trait scores do not control for the possibility of halo effects because the same rater assigns all of the trait scores to a particular student's essay. In addition, most previous studies have focused on only a limited number of potential dimensional structures, typically limiting attention to unidimensional or two-dimensional structures, even though analytic rubrics commonly depict four or more traits (e.g., conventions, organization, development, and voice).

Halo Effect

Thorndike (1920) coined the phrase *halo error* (also known as general impression halo; Balzer & Sulsky, 1992) to refer to the tendency to think of a person in general and to base one's judgment of that person on that general feeling. Guilford (1963) introduced an analogous term, *logical error* (also known as *dimensional similarity halo*; Balzer & Sulsky, 1992), to refer to a rater's tendency to infer relationships among performance dimensions that may not exist in truth. Both of these tendencies on the part of raters result in spuriously high correlations between multiple analytic scores. Differentiating halo and logical errors would be difficult, and we do not attempt to do so within this study. Rather, we refer to them collectively as the halo effect. More specifically, to maintain consistency we adopt the terminology that has been used by several other researchers (Balzer & Sulsky, 1992; Cooper, 1981; Murphy & Balzer, 1986) to differentiate spurious correlations that are introduced by halo error and logical error, *illusory halo*, from the true correlations between performance dimensions, *true halo*.

Illusory halo may be introduced when (a) the rater's general impression of response quality distorts his or her ratings of specific traits; (b) the rater's perception of one or more salient traits distorts his or her ratings on other traits; or (c) the rater is unable to discriminate among different categories of response quality, because of ambiguities in the rubric and/or insufficient training. Fisicaro and Lance (1990) refer to these three causal models as the General Impressions model, the Salient Dimensions model, and the Inadequate Discrimination model, respectively. In multi-trait scoring designs (i.e., each rater assigns multiple trait scores), all three of these causes of illusory halo may be concerning because they may inflate the correlations between the multiple trait scores assigned by each rater. In single-trait scoring designs (i.e., each rater scores one and only one trait), on the other hand, it is more difficult to argue that each potential cause of illusory halo will necessarily inflate observed between-trait correlations. For that to happen, raters would need to be influenced in the same manner by the same features. Otherwise, if individual raters are influenced by different features and/or are influenced by the same features in different ways, then it is difficult to argue that observed correlations would necessarily be inflated.

Regardless, it seems clear that multi-trait scoring designs *maximize* opportunities to observe illusory halo, whereas single-trait designs *minimize* such opportunities.

Halo effects have been operationalized in several ways, and Balzer and Sulsky (1992) and Cooper (1981) categorize these definitions into five categories. *Intraratee variance* measures focus on the variance of ratings within a ratee; smaller within rater variances suggest halo effects. *Intercategory correlations* focus on the similarity of ratings between traits; correlations closer to 1.00 suggest an overall halo effect (i.e., exhibited by raters, in general). *Rater-by-ratee interactions* depict halo effect within an analysis of variance framework for a fully crossed rater-by-ratee-by-dimension design. A statistically significant interaction suggests the existence of a halo effect. *Intercategory factor structure*, and the associated multitrait-multimethod and statistical partialing approaches, employ factor analytic and structural equation modeling (SEM) methods to determine whether separate analytic dimensions are statistically distinct. A single dimension suggests an overall halo effect.

Cooper (1981) criticizes these four methods because they fail to differentiate true and illusory halo. Murphy, Jako, and Anhalt (1993) explain,

Except in those rare circumstances in which the dimensions being rated are truly orthogonal, the observed correlation between dimensional ratings represents a composite of the true correlation and the net result of the cognitive distortions, errors in observation and judgment, and rating tendencies of the individual rater (i.e., illusory halo). This suggests that the distinction between true and illusory halo is fundamental; unless the true part can be removed from measures of observed intercorrelation among performance dimensions, it will never be possible to tell the extent to which observed correlations represent rater errors versus accurate depictions of the relationships among dimensions. (p. 221).

To remedy this problem, Balzer and Sulsky (1992) introduce a fifth operational definition of halo effect, *true halo*, in which halo is estimated by comparing operational ratings with expert ratings, which are assumed to be free of halo. Illusory halo is suspected when the between trait correlations of operational ratings is closer to 1.00 than is the comparable value for expert ratings or when the average within-ratee variance of operational ratings is closer to 0.00 than is the comparable value for expert ratings.

We agree with Cooper (1981) that none of the first four operational definitions of illusory halo are interpretable without a measure of true halo (i.e., ratings that can be assumed to be free from illusory halo). Unfortunately, expert ratings, which are suggested by Balzer and Sulsky (1992), may not be easy to collect, and there is no guarantee that experts are less likely to exhibit illusory halo than are standard raters. Hence, we offer an alternative approach which relies only on operational ratings. Specifically, we propose comparing ratings in which a single rater assigns all trait scores (using a design that maximizes illusory halo potential) to ratings in which different raters assign each trait score (using a design that minimizes illusory halo potential). Clearly, this is not an optimal solution because illusory halo can still exist in single-trait designs. For example, raters who rate a unique aspect of an essay may still be influenced by their overall impression of the response or by other characteristics

of the essay, and these factors could, potentially, influence the ratings of assigned traits, provided the multiple raters are influenced in the same manner by the features that cause the general impression error. Regardless, at the very least, having different raters rate different traits for a particular examinee is likely to minimize illusory halo in comparison to having a single rater rate all traits. We are aware of at least one other study that adopts a similar approach to implementing Balzer and Sulsky's (1992) concept of true halo (Bechger, Maris, & Hsiao, 2010) by comparing ratings obtained when raters score examinees on *all* dimensions to ratings obtained when each rater scores examinees on only a *single* dimension.

Purpose

This article attempts to differentiate illusory and true halo in analytic scoring of writing assessments and to determine the incremental information provided by analytic scores beyond what is captured by holistic scores. This distinction is important because illusory halo effects, when they exist, will artificially inflate observed between-trait correlations, causing researchers to erroneously conclude that fewer traits can be differentiated by raters than would be possible if the illusory halo effect could be eliminated or minimized. To evaluate the prevalence of illusory halo effects in trait scores, we employ a data collection design in which potential illusory halo is minimized and compare, using several common operationalizations of illusory halo, scores assigned under that design to scores assigned using a multitrait design (i.e., where potential illusory halo is maximized). We also use SEM methods to evaluate the uniqueness of the multiple trait scores in comparison to holistic scores. Our research addresses the following questions:

1. To what extent do analytic trait scores of writing assigned by the same raters exhibit an illusory halo effect compared to analytic trait scores assigned by different raters?
2. How highly correlated are holistic and analytic trait scores of writing when illusory halo is minimized?
3. What dimensional configuration best captures the structure of holistic and analytic trait scores of writing when illusory halo is minimized?

Method

Essays and Students

Our study employs the responses of 2,000 middle school students to an expository essay prompt that asked students to explain why spending too much time on a computer might be a problem and to propose possible solutions to that problem. The essay assessment is part of a high-stakes summative state-wide assessment. The responses are a representative sample of those from public school students from a single large state that has a diverse population.

Rubrics and Raters

Six groups of raters ($n = 40$ per group; 240 raters in all) participated in this study. Scores were assigned on a 4-point scale (0 to 3) based on one or more scoring rubrics: (a) analytic-idea development (*development*), meaning the development of content through the use of relevant details; (b) analytic-organization (*organization*), meaning the use of structure to support writing purpose and effectiveness; (c) analytic-voice (*voice*), meaning the appropriate and precise use of language to communicate directly to the audience in a way that is informative, compelling, and engaging; (d) analytic-conventions (*conventions*), meaning the correct use of mechanics, including spelling, punctuation, capitalization, and grammar; and (e) a holistic rubric designed to jointly consider these four traits (included in the appendix).

Raters were selected from two locations that house commercial scoring centers (one in the Midwestern United States and one in the Southwestern United States). Raters were assigned to one of six scoring conditions, with conditions nested within location. Groups 1 and 2 were located at the Midwestern scoring center, and Groups 3 through 6 were located at the Southwestern scoring center. Because raters were not randomly assigned to groups, we conducted preliminary comparisons of the scoring groups before training to ascertain their comparability with respect to several demographic characteristics (education, gender, age, and scoring background). Following training, we also compared the six groups with respect to interrater agreement during operational scoring. These comparisons are summarized in the Results section.

Raters were trained to use one or more of the scoring rubrics. To ensure adequate variation in rater performance, we did not require raters to qualify in order to participate in the study. Each group was trained using the same sets of student responses, and training and scoring procedures were the same for all groups with the exception of the rubrics and rationales provided for the student examples in the training sets.

Following training, randomly selected pairs of raters assigned scores to each of 2,000 student responses within each of the six groups of raters. Group 1 scored only the development trait, Group 2 scored only organization, Group 3 scored only voice, Group 4 scored only conventions, Group 5 scored all four traits, and Group 6 applied the holistic rubric. Hence, Groups 1 through 4 jointly produced trait scores that were assigned independently, Group 5 produced trait scores that were all assigned by a single rater, and Group 6 produced holistic scores that were designed to be a composite of the four traits. In all, each student response was assigned a pair of trait scores for each of the four traits under a single-trait design (Groups 1 through 4), a pair of trait scores for each of the four traits under a multitrait design (Group 5), and a pair of holistic scores (Group 6).

Analysis

Research Question 1: Illusory Halo. To address the first research question, we compared the values of several illusory halo indicators that were computed for Groups 1

through 4 (the single-trait design that minimizes illusory halo) and Group 5 (the multitrait design that maximizes illusory halo). Specifically, we compared the values of the intrarater variance and the intercategory correlations. We also employed a strategy described by Bechger et al. (2010), which compares alternate forms reliability coefficients for pairings of raters in which illusory halo might exist to reliability coefficients for pairings of raters in which illusory halo should not exist. Specifically, we generated separate correlation matrices for Groups 1 through 4 and Group 5 by computing “total scores” for “alternate forms” by creating four subtest total scores: the sum of (a) Rater 1’s scores on Traits 1 and 2, (b) Rater 1’s scores on Traits 3 and 4, (c) Rater 2’s scores on Traits 1 and 2, and (d) Rater 2’s scores on Traits 3 and 4. These total scores were computed for each possible pairing of traits, and a correlation matrix was composed for each of these pairings. When illusory halo arises from multitrait scoring designs, one would expect ratings assigned to different half-tests by the same rater—for example, (a) compared with (b) and (c) compared with (d)—to be more highly correlated than would be ratings assigned to different half-tests by different raters—for example, (a) compared with (d) and (b) compared with (c). These correlations can be converted to an effect size index, via a transformation of the Spearman–Brown prophecy formula (Equation 1), as described by Bechger et al. (2010), which indicates the effective decrease in test length incurred because of the existence of illusory halo.

$$k^{-1} = \frac{\rho_{XX'}(1 - \rho_{XX'}^*)}{\rho_{XX'}^*(1 - \rho_{XX'})}, \quad (1)$$

where $\rho_{XX'}$ is the Spearman–Brown corrected correlation between scores on different half-tests for the same rater and, $\rho_{XX'}^*$ is the Spearman–Brown corrected correlation between scores on different half-tests for different raters.

We also investigated illusory halo using an SEM that included separate latent factors for traits and methods. In this model, each observed indicator loaded on the appropriate trait factor (development, organization, voice, or conventions of writing) and the relevant method factor (either single-trait scoring or multitrait scoring), depending on which group and rubric were used. Such an approach allowed us to decompose observed variance into separate components: variance due to trait (roughly equivalent to true halo), variance due to method (roughly equivalent to illusory halo), and measurement error. Method factors were constrained to be orthogonal to trait factors, which means they captured shared variance arising from a particular scoring method that was unrelated to trait variance. Models such as these are referred to as correlated traits–correlated methods (CT-CM) models (see right panel of Figure 1) when traits are allowed to covary and methods are allowed to covary. Accordingly, our model had four trait factors (development, organization, voice, and conventions of writing) and two method factors (single-trait scoring method and multitrait scoring method). Trait factors were allowed to covary with one another, which is consistent with the expectation that construct-relevant features of a written response will be systematically related to one another. Method factors were also allowed to covary with

Figure 1. Correlated traits-correlated methods (right) model and traits-only model (left).

Note. G = group; T = trait.

one another, which is consistent with the expectation that illusory halo may arise from either the single-trait or the multitrait scoring design. To test for the existence of an illusory halo effect, we compared the fit of the CT-CM model to the fit of a model in which there were four trait factors, but no separate methods factors (see left panel of Figure 1). Instead, ratings for each trait produced by the single-trait method and the multi-trait method were stipulated to load together on the same trait factor.

This additional SEM analysis complements the more traditional illusory halo indicators in several ways. First, by comparing the fit of the two SEM models, this analysis provides a statistical significance test for the presence of illusory halo that can be supplemented by estimated effect sizes. Second, by creating latent factors for traits and methods, we are able to isolate and control for measurement error in the trait scores that is independent of illusory halo. Finally, using an SEM framework allows us to make more specific interpretations concerning the contribution of each trait to illusory halo by examining factor loadings. In contrast, the traditional illusory halo indicators involve taking averages across individual traits, which precludes this type of detailed interpretation.

Research Question 2: Trait/Holistic Score Correlations. To address the second research question, we computed observed-score correlations between the individual trait scores assigned by Groups 1 through 4, where illusory halo was minimized through scoring design, and the holistic scores assigned by Group 6.

Research Question 3: Trait/Holistic Score Dimensionality. To address the third research question, we employed confirmatory factor analysis methods. Specifically, we compared the fit of five models, which are shown in Figure 2. In the *unidimensional* model (UD), all scores (analytic and holistic) loaded on a single factor, representing general writing ability. Although this model has not been suggested in previous research, most studies specify unidimensionality as the null model. The *two-dimensional* model (2D) mapped holistic, development, organization, and voice scores onto one factor and conventions scores onto a separate factor because previous research has suggested that some writing assessments only differentiate a communicative component from a mechanics component (Aryadoust, 2010). The *four-dimensional, holistic* model (4D-H) mapped each analytic score onto its own unique factor along with the holistic score, which was inspired by the results of research by Lee et al. (2008), which revealed high correlations between holistic and analytic scores. The *four-dimensional, conventions* model (4D-C) mapped each analytic score onto its own unique factor and loaded the holistic score onto the factors for development, organization, and voice. Finally, the *five-dimensional* model (5D) mapped each analytic score and the holistic score onto unique factors. These final two models replicate research by Bacha (2001), which found that holistic scores may not reveal underlying differences of students across analytic trait scores.

Figure 2. Confirmatory factor analysis models.
Note. G =group; 2D = two-dimensional model; 4D-H, four-dimensional, holistic model; 4D-C, four-dimensional, conventions model; 5D, five-dimensional model.

Estimation and Model Fitting. We used SAS software to compute descriptive statistics and correlations and Mplus software (Muthén & Muthén, 1998-2007) to conduct all factor and SEM analyses. For each examinee, we had two scores on each trait—one score from each member of a pair of raters randomly selected from each group. For all SEM and factor analysis, we summed the two ratings to produce a single total score. Because of the ordinal nature of the scores assigned to each trait, we estimated the models using a robust maximum likelihood estimator that employs a numerical integration method to handle categorical data. For all models tested, we fixed the factor loading of the first observed indicator for each factor to 1.00 to establish a scale and allow the model to be identified. Within each model, we specified uncorrelated errors. To determine the best-fitting model for research questions 1 and 3, we compared several fit indices across models, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and the Satorra–Bentler scaled chi-square difference test for nested models (Satorra & Bentler, 2001). We also examined the magnitude of estimated parameters (e.g., factor loadings, factor variances, observed-indicator R^2) relative to their standard errors and estimated latent factor correlations for the best-fitting models.

Results

Group Comparability

In this subsection, we summarize the results of ancillary analyses confirming the similarity of scoring groups with respect to demographics and operational agreement rates. The age of scorers varied slightly across the groups. In particular, the percentage of scorers below the age of 60 ranged from a high of 85% in Group 1 to a low of 40% for Group 5 (i.e., raters in Group 1 were more than twice as likely to report being younger than 60 years than those in Group 5), with these differences being statistically significant, $\chi^2_{(10) \text{ Age}} = 23.87, p = .01$. Although scorer gender and race also varied slightly across the groups, neither of these differences was statistically significant: $\chi^2_{(5) \text{ Gender}} = 3.21, p = .67$ and $\chi^2_{(5) \text{ Race}} = 7.31, p = .19$. Hence, although the age distribution differed across scoring groups, the two groups were comparable with respect to gender and race. We know of no research to suggest that age of rater is related to quality of assigned scores. Finally, for each analytic trait, as shown in Table 1, agreement rates for raters who assigned single scores and raters who assigned multiple scores were similar (ranging from 45% to 56% perfect agreement and from 94% to 97% perfect + adjacent agreement). Although perfect agreement percentages are generally slightly lower for the multitrait raters (Group 5), none of the pairwise comparisons between single- and multi-trait scoring groups were statistically significant. The level of perfect agreement achieved by raters who assigned holistic scores (52%) was comparable to that observed for raters who assigned analytic scores, with no statistically significant difference between groups, $Z = 0.55, p = .58$. Quadratically weighted coefficient kappa values ranged from .26 to .40 and were

Table 1. Operational Agreement Rates by Scoring Group.

Trait	Group	% Perfect/adjacent agreement	Weighted κ
Development	1	56/96	.40
	5	51/95	.33
Organization	2	52/97	.35
	5	52/96	.35
Voice	3	52/95	.35
	5	45/94	.26
Conventions	4	50/94	.40
	5	48/94	.35
Holistic	6	52/97	.39

Note. Kappa (κ) was computed using quadratic weighting.

comparable across traits for the single-trait groups and slightly lower for the multi-trait group.

Illusory Halo

Table 2 summarizes the intrarater variance, the intercategory correlations, and the within- and between-rater correlations and associated effect sizes (Bechger et al., 2010) separately for raters in Groups 1 to 4 (single-trait scoring) versus Group 5 (multitrait scoring). Because the intrarater variance is computed for each rater, we report the average of those values across raters. As shown in the first row of Table 2, the average intrarater variance for Group 5 (0.15) is 32% smaller than the corresponding average for Groups 1 to 4 (0.22). Because the intercategory total score correlations are computed for each pairing of traits, we report the average of those values (via a Pearson transformation) across all pairs of traits. As displayed in the second row of Table 2, the average intercategory correlation for Group 5 (0.67) is 1.24 times larger than the corresponding averaged correlation for Groups 1 to 4 (0.54). Again, because there were four traits, we computed the within and between rater correlations and associated effect sizes for all three pairings of dyads of traits, and we report the averaged correlations and effect sizes across those pairings of dyads in the third row of Table 2. The average within-rater correlation for Group 5 (0.85) is 18% larger than that for Groups 1 to 4 (0.72). On the other hand, the average between-rater correlation for Group 5 (0.64) is 12% smaller than that for Groups 1 to 4 (0.73).

The illusory halo effect size for Group 5 is much smaller than 1.00 (0.32), whereas this index is slightly larger than 1.00 for Groups 1 to 4 (1.05). As explained in Bechger et al. (2010), because the index represents a ratio of two reliability coefficients, positive or negative departures from a value of 1.00 represent changes in the number of independent ratings in the multitrait scoring design compared with the single-trait scoring design, with values less than 1.00 signaling an effective decrease

Table 2. Within- and Between-Rater Correlations.

Indicator	Statistic	Groups 1-4	Group 5
Intraratee variance ^a	V(X)	0.22	0.15
Intercategory correlations ^b	<i>r</i>	.54	.67
Within-rater correlation ^c	ρ	.72	.85
Between-rater correlation ^c	ρ^*	.73	.64
Illusory halo effect size ^c	k^{-1}	1.05	0.32

- a. These are averaged values across all raters.
- b. These averaged correlations were obtained by computing total scores within each trait, correlating those total scores, and then averaging the correlations across traits via a Pearson transformation.
- c. These are averaged values across all pairings of trait score dyads.

Table 3. Illusory Halo Model Fit.

Fit index	Traits-only model	CT-CM model
AIC	44419.54	43980.01
BIC	44766.79	44377.67
Log-likelihood	-22147.80	-21919.00
χ^2 difference test ^a (<i>df</i>)	NA	242.18 ^b (9)

Note. CT-CM = correlated traits–correlated methods; AIC, Akaike information criterion; BIC, Bayesian information criterion; NA, not applicable.

- a. χ^2 difference test is the Satorra–Bentler scaled chi-square difference test (Satorra & Bentler, 2001).
- b. $p < .0001$.

in test length due to illusory halo and values greater than 1.00 signaling an effective increase in test length.

Structural Equation Modeling Results

Table 3 presents fit indices for the two structural equation models (the CT-CM model and the traits-only model), which suggest that fit is significantly improved by incorporating method factors into the model. Namely, AIC and BIC decrease whereas the log-likelihood increases. The Satorra–Bentler scaled chi-square difference test suggests that the CT-CM model fits significantly better than a model including only traits.

Table 4 reports the estimated parameters for the CT-CM model. Factor loadings are all positive and statistically significant for both trait factors and method factors. Comparing the relative magnitude of factor loadings for the trait factors, one can see that for each trait except conventions, loadings from the single-trait scoring group are greater than those from the multitrait scoring group. For conventions, this trend

Table 4. Estimated Parameters for Correlated Traits–Correlated Methods Model.

Factor	Observed indicator	Factor loading ^a	Standard error	R ^{2b}
Development	Group 1 score	1.00	0.00	.69
	Group 5 score	0.70 ^c	0.14	.80
Organization	Group 2 score	1.00	0.00	.76
	Group 5 score	0.59 ^c	0.12	.86
Voice	Group 3 score	1.00	0.00	.73
	Group 5 score	0.44 ^c	0.07	.79
Conventions	Group 4 score	1.00	0.00	.69
	Group 5 score	1.15 ^c	0.22	.78
Single-trait method	Group 1 score	1.00	0.00	.69
	Group 2 score	1.26 ^c	0.09	.76
	Group 3 score	1.16 ^c	0.09	.73
	Group 4 score	1.06 ^c	0.09	.69
Multitrait method	Group 5, Trait 1	1.00	0.00	.80
	Group 5, Trait 2	1.25 ^c	0.09	.86
	Group 5, Trait 3	0.98 ^c	0.09	.79
	Group 5, Trait 4	0.84 ^c	0.06	.78

a. All factor loadings are unstandardized estimates.

b. All R² indices account for both variance due to trait and variance due to method.

c. $p < .0001$ (two-tailed test).

is reversed. Comparing the relative magnitude of factor loadings for the methods factors (which represent illusory halo in the different scoring groups), it is evident that trait loadings for the single-trait scoring method factor are all relatively similar. In contrast, trait loadings for the multitrait scoring method factor are more variable, with organization loading the strongest and conventions loading the weakest. For the CT-CM model, the proportion of variance explained in the observed indicators (which encompasses variance due to both trait and method factors) ranges from .69 to .76 for the single-trait group scores and from .78 to .86 for the multitrait group scores.

Estimates from the psi matrix for the CT-CM model (reported in Table 5) indicate that all factor variances are significantly different from zero. In substantive terms, the variance of the single-trait scoring method factor (representing illusory halo in Groups 1-4) is roughly twice the size of each of the trait factor variances, and the multitrait scoring factor variance (representing illusory halo in Group 5) is between four and five times larger than each of the trait factor variances. Estimated true halo correlations (corrected for both measurement error and illusory halo) for the CT-CM model range from $-.11$ (for development and conventions) to $.66$ (for development and voice). However, only two of these correlation coefficients are significantly different from zero—that between development and voice and that between organization and voice. Finally, the two method factors are correlated very highly ($r = .91$).

Table 5. Estimated Psi Matrix for Correlated Traits–Correlated Methods Model.

Factor ^a	Dev	Org	Voice	Con	Single	Multi
Dev	2.67^{be} (1.02)	1.18 ^c (0.86)	1.74 ^e (0.54)	–0.27 (0.50)	0.00	0.00
Org	.42 ^d	2.91^e (1.48)	1.74 ^e (0.63)	0.49 (0.71)	0.00	0.00
Voice	.66	.63	2.66^e (0.67)	0.46 (0.35)	0.00	0.00
Con	–.11	.20	.19	2.15^e (0.72)	0.00	0.00
Single	.00	.00	.00	.00	4.74^e (0.69)	6.93 (0.80)
Multi	.00	.00	.00	.00	.91	12.21^e (1.62)

a. Factors: Dev = development; Org = organization; Con = conventions; Single = single-trait; Multi = multitrait.

b. Factor variances are reported in boldface on the diagonal with standard errors in parentheses.

c. Latent trait covariances are reported above the diagonal with standard errors in parentheses.

d. Latent trait correlations are reported in italics below the diagonal.

e. $p < .05$ (two-tailed test).

Table 6. Intercategory Correlations.^a

Trait ^b	Dev	Org	Voice	Con	Holistic
Dev	1.00	.56	.59	.42	.61
Org	.73	1.00	.62	.52	.63
Voice	.71	.75	1.00	.50	.62
Con	.55	.65	.62	1.00	.53
Holistic	.63	.63	.60	.53	1.00

a. Groups 1 to 4 are reported in the upper off-diagonal for Dev, Org, Voice, and Con, and Group 5 is reported in the lower off-diagonal. Holistic scores are from Group 6.

b. Trait: Dev = development; Org = organization; Con = conventions.

Table 7. Dimensionality Model Fit Statistics.

Model	AIC	BIC	χ^2 difference test ^a (df)	Model for χ^2 comparison
UD	29445.03	29641.06	NA	NA
2D	29444.37	29646.01	14.28 ^b (1)	UD
4D-C	29410.03	29650.87	36.44 ^c (7)	2D
4D-H	29411.26	29657.70	−0.25 (1)	4D-C
5D	29412.40	29664.44	38.56 ^c (9)	2D

Note. UD, unidimensional model; 2D, two-dimensional model; 4D-C, four-dimensional, conventions model; 4D-H, four-dimensional, holistic model; 5D, five-dimensional model.

a. χ^2 difference test is the Satorra–Bentler scaled chi-square difference test (Satorra & Bentler, 2001).

b. $p < .001$.

c. $p < .0001$.

Trait/Holistic Score Correlations

Table 6 summarizes the observed-score correlations between trait scores assigned by raters in Groups 1 to 4 and holistic scores assigned by raters in Group 6. As can be seen, when illusory halo is minimized through scoring design, correlations between individual trait scores assigned using an analytic rubric and scores assigned using a holistic rubric range from 0.53 (for conventions) to 0.63 (for organization). Scores on individual analytic traits are also moderately correlated with one another, with correlations ranging from .42 (for conventions and development) to .62 (for voice and organization).

Trait/Holistic Score Dimensionality

Table 7 presents the model fit indices (AIC, BIC, and Satorra–Bentler chi-square difference test) for each model. The results are somewhat mixed, with the AIC suggesting that the 4D-C model is the best fitting, while the BIC identifies the UD or 2D models as being equal in terms of fit. Taken together, these indices suggest that neither the 4D-H nor the 5D models provide an adequate depiction of the structure of

Table 8. Confirmatory Factor Analysis Model Between-Factor Correlations.

Model	Traits	Correlation
2D	Writing ability vs. conventions	.85
4D-C	Development vs. organization	.82
	Development vs. voice	.87
	Development vs. conventions	.62
	Organization vs. voice	.99
	Organization vs. conventions	.83
	Voice vs. conventions	.81

Note. 2D = two-dimensional model; 4D-C, four-dimensional, conventions model.

the observed data. The Satorra–Bentler chi-square difference test statistic for nested models tells a slightly different story, suggesting that the 2D model is a significant improvement over the 1D model; and in turn, model 4D-C is a significant improvement over the 2D model. When compared with the 2D model, model 5D also fits significantly better, but cannot be compared with either of the 4D models using the Satorra–Bentler index because there is no nesting relationship between them. Finally, model 4D-H, when compared with model 4D-C, produced a negative Satorra–Bentler value, which cannot be used to evaluate relative fit. Thus, we are left with, at best, three potential models: UD, 2D, and 4D-C.

Inspection of the latent factor correlation estimates provides a potential justification for differentiating between the usefulness of these models. Table 8 summarizes those correlations, which indicate that model 4D-C may make unnecessary distinctions between voice and the development and organization traits, with estimated latent factor correlations equal to .87 and .99, respectively. In addition, the correlation between the development and organization latent factors is not altogether low, estimated to equal .82. In all cases, these correlations support the notion that even though these three measures of writing ability are not necessarily distinguishable from one another, they are distinguishable from the conventions latent factor, with all correlations being less than .85. Hence, our results support a 2D model that distinguishes conventions from the remaining scores.

Table 9 reports estimated parameters for this 2D model. First, factor loadings are all positive and large in relation to their standard errors. The score contributing the most to the variance of general writing ability is the holistic score, whereas the score contributing the least is the development score. R^2 indicators suggest that the model accounts for between 56% of the variance (for development scores) and 68% of the variance (for holistic scores) in the observed indicators.

Discussion and Conclusions

This study sought to answer three research questions relating to analytic and trait scores both in the presence and absence of opportunities for raters to introduce

Table 9. Estimated Parameters for Two-Dimensional Model.

Factor	Observed Indicator	Factor Loading ^a	Standard Error	R ^{2b}
Writing ability	Group 1 score	1.00	0.00	.56
	Group 2 score	1.21 ^b	0.07	.65
	Group 3 score	1.23 ^b	0.06	.66
	Group 6 score	1.311 ^b	0.07	.68
Conventions	Group 4 score	1.00	0.00	.58

a. All factor loadings are unstandardized estimates.

b. $p < .0001$ (two-tailed test).

illusory halo into scores. In response to the first research question—*To what extent do analytic trait scores of writing assigned by the same raters exhibit illusory halo?*—collectively, our results strongly imply the presence of illusory halo for raters assigning multiple analytic scores. First, the results of traditional illusory halo indices consistently point to the presence of illusory halo. Scores for raters in Group 5 who evaluated multiple analytic traits exhibited less variability across traits than did scores for raters in Groups 1 to 4 who only rated single traits. Further, trait scores for the former group are more highly correlated than are trait scores for the latter group. Effect sizes for the different scoring groups suggest that having raters score all four traits reduces effective test length to around 33% of what it would be if each rater scored only a single trait, resulting in substantially less information about student performance.

Results from the SEM analyses also support the conclusion that illusory halo exists in the scores assigned by raters in the multitrait scoring group. Including separate method factors significantly improves model fit relative to a model with only trait factors, which is confirmed by the magnitude of estimated parameters for the better-fitting model. Estimated true halo correlations (controlling for illusory halo) are less extreme than are observed-score correlations for raters scoring all four traits, even though true halo correlations have been disattenuated for measurement error. SEM results are also consistent with traditional indices regarding the severity of illusory halo, as the estimated factor variance for illusory halo in the multitrait scoring group was four to five times larger than corresponding variances for the trait factors.

The SEM results also support inferences about the relative contributions to illusory halo of different traits. Because loadings on the multitrait scoring group illusory halo factor varied, this suggests that not all traits are equally likely to elicit illusory halo when a single rater assigns all trait scores. In particular, organization appears to be most subject to illusory halo, whereas conventions is least influenced when raters score all four traits. This result is consistent with previous studies that have found inconsistent method effects across traits (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003).

It is also possible that variable factor loadings simply reflect the relative order in which scorers tend to evaluate the traits. If student performance on the first or second trait evaluated impacts ratings on subsequent traits, one might expect that the multitrait scoring method factor loadings for those initial traits would be larger than those

for traits rated subsequently. In this study, no data were collected concerning the sequence in which scorers rated the traits. Although scorers were trained to evaluate the traits in a certain order (development, organization, voice, and conventions), scorers were free to evaluate the traits in any order during operational scoring. There is no reason to assume that scorers would continue to evaluate the traits in the same order they were trained. Additional research is needed to determine whether variable factor loadings suggest that some traits are more influential than others, or whether they merely reflect the order in which traits happen to be evaluated.

Finally, both the SEM and traditional illusory halo indices suggest the presence of some amount of illusory halo in Groups 1 to 4. The effect size analysis suggests that this illusory halo is small and negative. In other words, as a result of the illusory halo among Groups 1 to 4, effective test length is slightly increased relative to what it would be if raters scored all four traits. Bechger et al. (2010) found a similar negative halo when they assigned different scorers to rate different traits. Based on the magnitude and significance of estimated parameters associated with the single-trait scoring method factor, the SEM analysis also points to illusory halo in Groups 1-4. Furthermore, the very high estimated correlation between the single-trait scoring method factor (representing illusory halo in Groups 1-4) and the multitrait scoring method factor (representing illusory halo in Group 5) implies that illusory halo may be operating in similar ways across the two types of scoring groups. Although raters in Groups 1 to 4 were only tasked with scoring a single trait, it is possible that other aspects of the students' writing, handwriting quality, for example, could have introduced illusory halo into the multiple scores that they assigned. This is why it may be possible to minimize illusory halo through particular scoring designs but not eliminate it completely.

Future research should attempt to confirm the results produced in this study using different types of statistical models for detecting illusory halo. For example, we could have conducted multigroup invariance testing, creating separate latent factors for each rubric trait (which are allowed to covary), and then testing to see whether this factor structure is invariant across scoring groups (single-trait scoring groups vs. the multitrait scoring group). In particular, illusory halo would lead us to expect factor correlations to be larger for the multi-trait scoring group than for the single-trait scoring group. Given the results uncovered in this study, factor loadings and factor variances might also be expected to vary.

Our analyses focusing on the remaining two research questions—*How highly correlated are holistic and analytic scores of writing when illusory halo is minimized?* and *What dimensional configuration best captures the structure of holistic and analytic trait scores of writing when illusory halo is minimized?*—produced slightly mixed results. First, when illusory halo is minimized through scoring design, observed-score correlations between scores assigned using holistic and those assigned using analytic rubrics are only moderate in size, which provides support for the incremental information provided by analytic scores over holistic ones. This result is consistent with the magnitude of observed-score correlations between analytic and holistic scores found in previous studies (Klein et al., 1998). In addition, observed-score correlations

between individual analytic traits are only moderate in size, which lends support to the argument that these scores are capturing distinct aspects of student writing. The magnitude of these correlations fits well within the range of those found in previous studies. For example, Klein et al. (1998) found observed-score correlations between analytic traits ranging from 0.16 to 0.68, whereas Lee et al. (2008) found higher correlations, ranging from 0.66 to 0.89.

However, the confirmatory factor analysis models suggest a slightly different answer to these questions by identifying a different dimensional structure for these data. Namely, when illusory halo is minimized, analytic scores relating to features of the voice, development, and organization may be too highly correlated to provide useful distinctions, at least at the grade level and mode of writing investigated in this study. For example, in the best fitting alternative to the two-dimensional model, the estimated latent factor correlations between voice and the factors defined by development and organization were greater than .87.

Hence, we suggest that, at best, we can distinguish two dimensions of expository writing for middle school students and, even then, the correlation between what we refer to as “writing ability” and conventions is still quite high (i.e., $r = .85$ in our analyses). We believe that this is an important result because prior attempts to depict these correlations relied on data collection procedures that may have confounded between-factor correlations with illusory halo—a fact that may have inflated estimates of the latent factor correlations. Our results suggest that these correlations are indeed as high as those observed in prior studies, even in scoring designs that minimize the influence of illusory halo.

Our results have implications for rubric development and scoring models in direct writing assessments. First, those who develop scoring rubrics should carefully consider whether, and if so how, to best differentiate the potentially multiple traits that they consider important to assess. Our results suggest that it may be difficult to construct scoring rubrics that allow raters to assign scores that clearly distinguish between voice, organization, and development. While it may be desirable on the part of educators to provide those multiple scores, we question whether the added cost of developing the multiple rubrics, training raters to employ those multiple rubrics, and having raters assign the multiple scores is worth the cost, particularly if the within student variability is due, primarily, to measurement error rather than true differences in student performance across those traits. If multitrait rubrics are employed, raters should be trained to clearly distinguish between those traits and to guard against illusory halo in scoring designs in which a single rater assigns multiple trait scores.

Second, concerning the design of operational scoring, those who plan and conduct direct writing assessment scoring projects should carefully consider the design of data collection and how that design may impact the quality of the scores when multiple trait scores are assigned to each response. Our research suggests that a design in which multiple raters (each rating a different trait) rate each response may be less subject to illusory halo than would be a design in which a single rater scores all traits. However, there are a couple of important points to consider before concluding that

the multiple rater design is the best option. Multiple rater designs will be more expensive to implement than will be single rater designs. Multiple rater designs require more raters to be hired at a given time in order to achieve the same turnaround rate (i.e., a separate group for each trait to be scored), and the increase in cost of training and monitoring those additional raters would be proportional to the number of traits being scored. In addition, our study, and no other study that we are aware of, has sought to determine the degree to which raters in a single rater design can be trained in a way that minimizes the risk of illusory halo. Prior research in other domains (e.g., employee evaluation) has suggested that training raters to recognize and avoid specific rating errors in addition to training them to recognize and base decisions on relevant features of a performance may help minimize the occurrence of rater effects (Woehr & Huffcutt, 1994). Hence, those who design and conduct operational rating projects should carefully consider the relationship between scoring design, rater training, and project costs when determining how to best minimize the possibility of illusory halo in trait scores.

Appendix

Holistic Rubric

Score Point 3

The writer presents a clear problem and logical solution and develops ideas through the use of relevant descriptive details. An effective structure supports the purpose and effectiveness of the writing. The writer consistently uses a variety of appropriate, precise language to communicate directly to the audience in a way that is informative, compelling, and engaging. The writer consistently develops the mechanical correctness of the piece, including spelling, capitalization, punctuation, grammar, and usage.

Score Point 2

The writer presents a clear problem and logical solution and develops ideas through the use of relevant details, OR the writer presents only a clear problem or only a logical solution and develops ideas through the use of relevant descriptive details. The structure supports the purpose of the writing. The writer uses appropriate, precise language to communicate to the audience. The writer develops the mechanical correctness of the piece, including spelling, capitalization, punctuation, grammar, and usage.

Score Point 1

The writer attempts to present a problem or solution and develops ideas through the limited use of relevant details. The attempted structure provides limited support to the purpose of the writing. The writer uses little variation in word choice to communicate to the audience. The writer develops the mechanical correctness of the piece, although some grade appropriate words are misspelled and there is limited control of capitalization, punctuation, grammar, and usage.

Score Point 0

The writer fails to respond to the topic and/or the details are consistently irrelevant. The structure fails to support to the purpose of the writing. The writer uses an extremely limited word choice to communicate to the audience. The writer fails to develop the mechanical correctness of the piece, including spelling, capitalization, punctuation, grammar, and usage.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Pearson, Inc. funded the research study. There is no specific product associated with the research; however Pearson has a general interest in researching different methods for scoring.

References

- Aryadoust, V. (2010). Investigating writing sub-skill in testing English as a foreign language: A structural equation modeling study. *TESL-EJ*, 13, 1-20.
- Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, 29, 371-383.
- Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology*, 77, 975-985.
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12, 86-107.
- Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance based examinations. *Applied Psychological Measurement*, 34, 607-619.
- Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale type in the context of composition tests. *Issues in Applied Linguistics*, 11, 207-241.
- Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement*, 2, 379-388.
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-224.
- Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple indicator CT-C(M-1) model. *Psychological Methods*, 8, 38-60.
- Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, 14, 419-429.
- Guilford, J. P. (1936). *Psychometric methods*. New York, NY: McGraw-Hill.
- Johnson, R. L., McDaniel, F., & Willeke, M. J. (2000). Using portfolios in program evaluation: An investigation of interrater reliability. *American Journal of Evaluation*, 21, 65-80.
- Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., & . . . Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, 11, 121-137.
- Lee, Y. W., Gentile, C., & Kantor, R. (2008). Analytic scoring of TOEFL CBT essays: Scores from humans and E-rater. Princeton, NJ: ETS.
- Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology*, 71, 39-44.
- Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, 78, 218-225.

- Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507-514.
- Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22, 1-30.
- Singer, N. R., & LeMahieu, P. (2011). The effect of scoring order on the independence of holistic and analytic scores. *Journal of Writing Assessment*, 4, 1-13.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25-29.
- Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67, 189-205.