Revisiting the Halo Effect Within a Multitrait, Multimethod Framework

Annual meeting of the American Educational Research Association
San Francisco, CA

Emily R. Lai
Edward W. Wolfe
Daisy H. Vickers

April, 2013

Abstract

This report summarizes an empirical study on halo effect. Specifically, we investigate the extent to which multiple analytic scores assigned by a single rater display evidence of a halo effect. We analyzed scored student responses to an expository writing prompt that were scored by five groups of raters—four groups assigned single analytic scores and one group assigned multiple analytic scores—using structural equation modeling. Our results suggest that there is evidence of a halo effect when raters assign multiple analytic scores to a single student response, which tends to inflate observed between-trait correlations.

Keywords: halo effect, trait scores, analytic rubrics, writing assessment, performance assessment

Revisiting the Halo Effect Within a Multitrait, Multimethod Framework

Student responses to writing assessments are commonly evaluated using analytic rubrics that assign a set of analytic scores to each of several traits. Under an analytic scoring model, raters are trained to evaluate each component of the writing separately and assign a separate score to each trait (e.g., mechanics, organization, voice, and development). Although analytic scores require relatively more time of the scorer, they provide potentially useful diagnostic information to the student. However, analytic scores may be subject to halo effects when a single rater assigns all analytic scores to a particular essay. The purpose of this study is to addresses this issue directly, focusing on the following research question: *To what degree do multiple analytic scores assigned by a single rater contain evidence of a halo effect?*

Theoretical Framework

For performance-based or constructed-response tasks that require human scoring, one potential source of bias is referred to as the "halo effect." Originally defined as "suffusing ratings of special features with a halo belonging to the individual as a whole," this effect is generally hypothesized to occur when examinees are rated along multiple dimensions by the same person (Thorndike, 1920, p. 25). More contemporary definitions of the halo effect characterize it as the effect of a rater's overall or general impression of an examinee on specific dimensions of performance (Murphy, Jako, & Anhalt, 1993; Solomonson & Lance, 1997). Bechger, Maris, and Hsiao (2010) explain that halo occurs "when judgments of one rated characteristic influence judgments of other characteristics in a positive or negative direction" (p. 607). The halo effect is typically manifested via the following "symptoms:" inflated correlations between observed scores on different

dimensions; high rater-by-examinee interactions; low within-examinee dimensional variance; and a decrease in the number of independent opportunities for the examinee to demonstrate his or her proficiency (Bechger et al., 2010; Viswesvaran, Schmidt, & Ones, 2005).

Researchers have used several approaches to study halo effect, including examining correlations within and between raters and dimensions (Viswesvaran et al., 2005), conducting generalizability studies (Hoyt, 2000), and building structural equation models (SEM) to capture this effect (Cheung, 1999; Conway, 1999; Marsh & Butler, 1984; Marsh & Yeung, 1997). Researchers tend to invoke the relationship between observed, true, and error components of between-dimension correlations. In particular, observed correlations are a function of "true" correlations, which represent the correlations between the constructs or latent traits rather than the observed measures, and an error component. In this case, the error component represents a combination of halo bias and measurement error (Solomonson & Lance, 1997). Previous approaches to quantifying halo effect typically attempt to "purify" observed correlations by isolating true and error components, an approach that has been criticized due to the difficulty of obtaining good measures of "true" correlation (Kozlowski & Kirsch, 1987; Murphy et al., 1993). To date, no studies that we are aware of have quantified halo 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

The purpose of this paper is to investigate potential halo using a multitrait-multimethod (MTMM) framework. We employ a unique design that allows us to

compare ratings obtained when raters score examinees on *all* dimensions to ratings obtained when each rater scores examinees on only a *single* dimension. We analyze the data via traditional MTMM methods, as well as structural equation modeling approaches.

Method

Data Sources

Five groups of raters (N = 40 per group) participated in this study, with each group assigning scores to student responses to an expository writing prompt (depicted in Figure 1) using one or more scoring rubrics: (a) analytic-idea development (*idea*), (b) analytic-organization (*organization*), (c) analytic-voice (*voice*), and (d) analytic-conventions (*conventions*). Each rubric contained four score points (0 to 3). Raters were selected from two locations that house a Pearson scoring center (one in the midwestern United States and one in the southwestern United States). Raters were assigned to one of five scoring conditions, with conditions nested within location[1]: Group 1 scored only the development trait, Group 2 scored only organization, Group 3 scored only voice, Group 4 scored only conventions, and Group 5 scored all four traits. Hence, Groups 1 through 4 produced trait scores that were assigned independently, and Group 5 produced trait scores that were all assigned by a single rater.

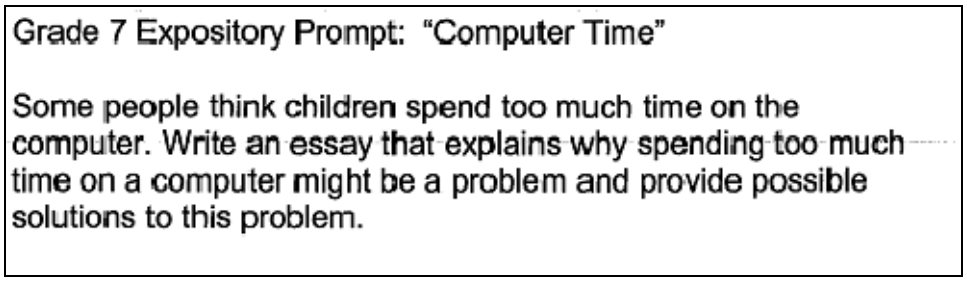


Figure 1. Seventh-grade expository writing prompt

Because raters were not randomly assigned to groups, we conducted preliminary comparisons of the scoring groups before training and qualification 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 agreement on the qualifying sets and with respect to interrater agreement during operational scoring. These comparisons are summarized in the Results section.

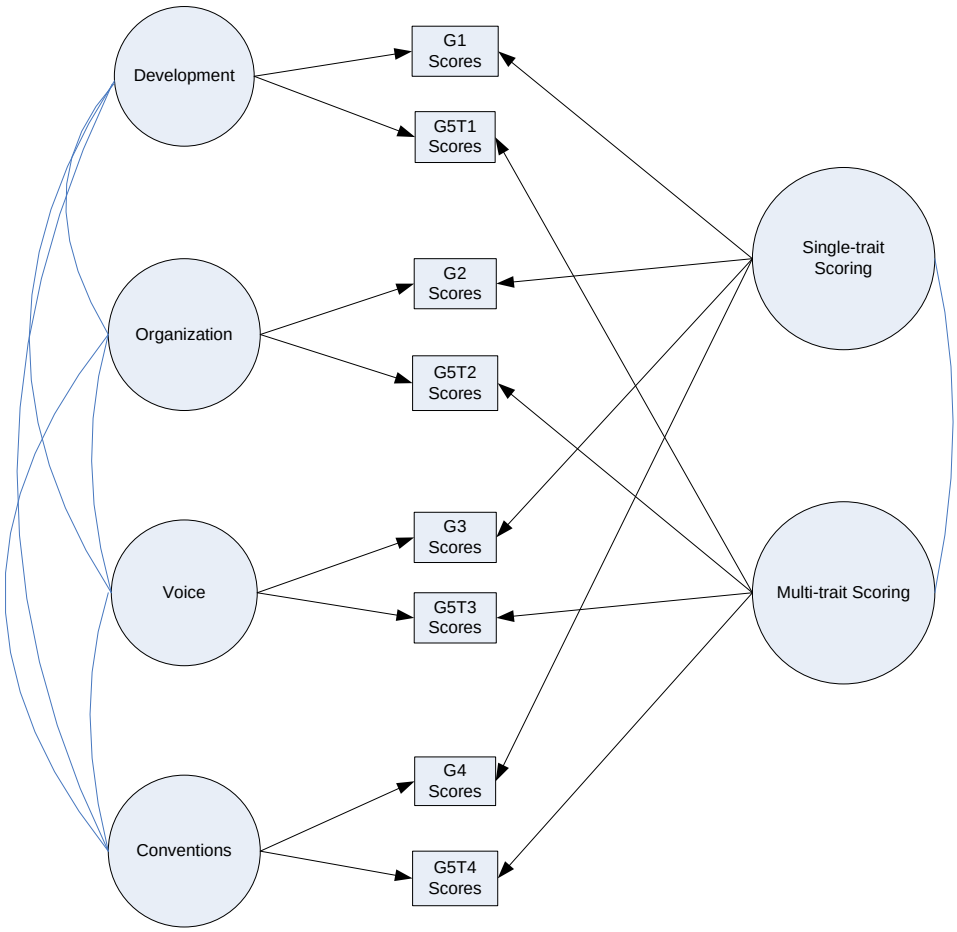
Raters were trained to use the scoring rubric, and each rater attained a qualification standard of at least 80% perfect agreement with expert raters on two of three sets of validity essays before participating in the study. Each group was trained and qualified using the same sets of student responses, and training, qualification, 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. Although training for Group 5 encouraged raters to assign scores for the four traits in a particular order (development, then organization, then voice, and finally conventions), raters in that group could score the traits in any sequence during operational scoring. Following qualification, randomly chosen pairs of raters assigned scores to each of 2,000 student responses. For each student response, a pair of raters was randomly selected from each group, and the

assigned scores were summed across the pair of raters to produce a single score for each student response.

Analysis

In our evaluation of halo effects, we examined bivariate correlation patterns within traditional MTMM correlation methods, where the two scoring approaches constituted the multiple methods. The MTMM correlation matrix is organized to enable identification of convergent validity, discriminant validity, reliability, and potential method effects (Campbell & Fiske, 1959).

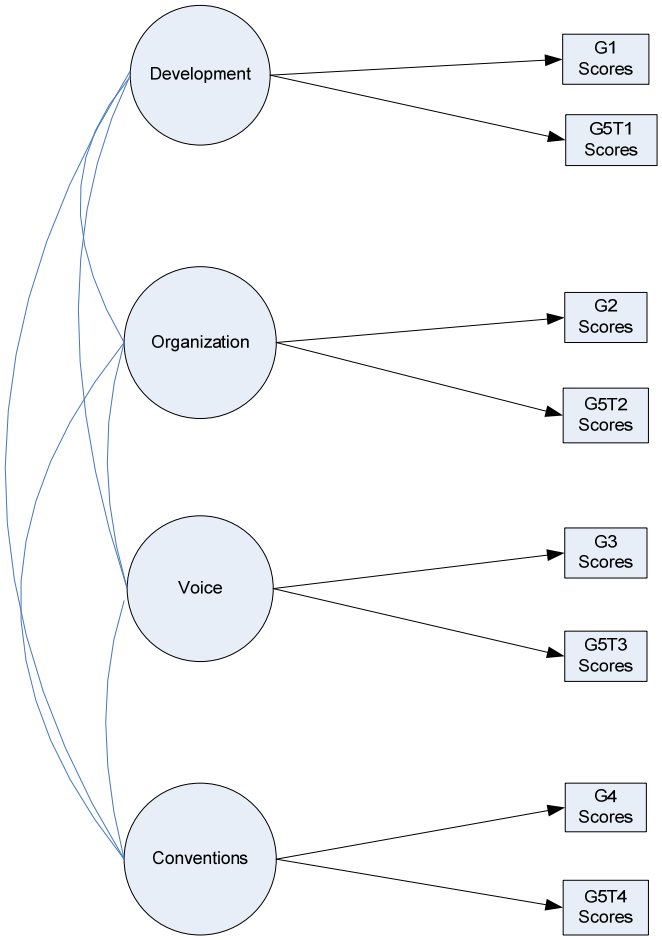
We also used SEMs that included separate latent factors for traits and methods. In these models, each observed indicator loaded on both a trait and a method factor. Such an approach allowed us to decompose observed variance into separate components: variance due to trait ("true" variance), variance due to method (halo), and measurement error. Such models are referred to as correlated traits-correlated methods (CT-CM) models (Figure 2) 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 multi-trait scoring method). Trait factors were allowed to covary, as were the method factors. However, trait and method factors did not covary with one another.



Note. G=group, T=trait

Figure 2. Correlated traits-correlated methods (CT-CM) model

To test for a potential 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 (Figure 3). Instead, summed ratings for each trait produced by the single-scoring method and the multi-scoring method were stipulated to load together on the same trait factor.



Note. G=group, T=trait

We used SAS software to compute descriptive statistics and Mplus (Muthén & Muthén, 1998-2007) to conduct all SEM analyses. Due to the ordinal nature of the scores assigned to each trait, we estimated the models using a robust maximum likelihood (MLR) 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 to1 to establish a scale and allow the model to be identified. Within each model, we treated errors as uncorrelated. To determine the best-fitting model within each study, we compared several fit indices across models, such as AIC and BIC and the Satorra-Bentler (SB) scaled chi-square difference test for nested models (Satorra & Bentler,

1999). 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.

Results

Scoring Group Comparability

We report descriptive statistics in this section that summarize the similarity of scoring groups with respect to demographics, qualifying, and operational agreement rates. Table 1 summarizes demographic indicators by scoring group, including age, gender, and ethnicity. The age of scorers varies slightly across the groups. In particular, the percentage of scorers below the age of 60 ranges from a high of 85% in Group 1 to a low of 40% for Group 5, with these differences being statistically significant, $\chi^2_{(10)\ \text{Age}} = 23.87$, $p = .01$. Thus, group 1 appears to have a larger share of relatively young scorers than the other groups. Although scorer gender and race also vary slightly across the groups, neither of these differences is 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.

Table 1

Rater Demographics by Scoring Group

Group	Demographic Statistic		
	Percent Under 60	Percent Female	Percent White
1	85.0	42.5	97.5
2	60.0	60.0	87.5
3	67.5	50.0	90.0
4	47.5	50.0	80.0
5	40.0	57.5	92.5

Table 2 Table 4 reports qualifying and operational rates of exact agreement by scoring group. For each trait, both qualifying and operational agreement rates for the single-scoring and the multi-scoring conditions were extremely similar. Z-test results demonstrate that there were no statistically significant differences across the single- and multi-scoring groups for any of the four traits

Table 2

Qualifying and Operational Agreement Rates by Scoring Group

Trait	Group	Percent Exact Agreement	
		Qualifying	Operational
Development	1	71	61
	5	66	60
Organization	2	60	61
	5	64	60
Voice	3	54	57
	5	61	54
Conventions	4	50	56
	5	57	55

Traditional MTMM Results

One set of analyses that we utilized involved examining the MTMM correlation matrix, which is shown in Table 3. This matrix is organized to highlight reliability coefficients, validity coefficients, heterotrait-monomethod correlations, and heterotrait-heteromethod correlations. We interpret each of these in turn.

Table 3

Multitrait-Multimethod Matrix

Method		Single-trait				Multi-trait			
	Trait	Dev	Org	Voice	Con	Dev	Org	Voice	Con
Single-trait	Dev	(.50)							
	Org	.56	(.50)						
	Voice	.59	0.63	(.47)					
	Con	.42	0.52	.50	(.55)				
Multi-trait	Dev	*.61*	.55	.57	.48	(.43)			
	Org	.51	*.65*	.58	.53	.73	(.48)		
	Voice	.51	.56	*.61*	.51	.71	.75	(.39)	
	Con	.39	.51	.49	*.67*	.55	.65	.62	(.49)

Note. Dev=Development, Org=Organization, Con=Conventions. Italicized values represent validity coefficients. Values in parentheses represent reliability coefficients, or the correlations between scores from randomly selected raters from the same group. Heterotrait-monomethod blocks are enclosed by solid lines. Heterotrait-heteromethod blocks are highlighted in gray.

Reliability coefficients (or monotrait-monomethod correlations), reported in parentheses on the diagonal, represent the correlation between scores on the same trait from two randomly selected raters in each group. As can be seen, reliability coefficients are rather low, ranging from .47 to .55 for the single-trait groups and from .39 to .49 for the multi-trait group. For each trait, the correlation between scores assigned by two randomly selected raters from the multi-trait scoring group is lower than that between two randomly selected raters from the single-trait scoring group. This pattern suggests that it is more difficult to achieve consistent ratings when scorers are responsible for multiple aspects of examinee performance than when scorers are rating only a single trait. Overall, the low reliability coefficients imply that there is a substantial amount of measurement error in these scores.

Validity coefficients are reported in italics and represent the correlation between scores assigned by single-trait raters and those assigned by multi-trait raters for the same trait. These coefficients depict convergent validity and range from .60 to .67, suggesting a reasonable degree of agreement between trait scores assigned by those in the single-trait and multi-trait groups.

Heterotrait-monomethod coefficients appear in solid blocks and represent correlations between different trait scores assigned using the same scoring method. These coefficients depict discriminant validity, or the extent to which distinct traits actually capture meaningful differences in aspects of examinee performance. As can be seen, correlations among traits for the single-trait scoring group range from .42 to .62, suggesting that scores are tapping meaningfully different constructs. However, corresponding correlations for the multi-trait scoring group are slightly higher, ranging

from .55 to .75. The difference between these two sets of correlations represents potential method variance or bias due to the halo effect. Thus, multi-trait scoring group correlations may be inflated by as much as 16–24% to the extent that scores on one trait were influenced by scores on other traits.

Finally, heterotrait-heteromethod coefficients can be found in the shaded boxes. These coefficients represent correlations between scores on different traits obtained using different scoring methods. According to Campbell and Fiske (1959), validity coefficients ought to be larger than corresponding correlations in both the heterotrait-monomethod block and the heterotrait-heteromethod block. It is evident that each validity coefficient is larger than values from its corresponding row and column within the heterotrait-heteromethod block, and this criterion is nearly satisfied for the heterotrait-monomethod block of the single-trait scoring group. However, several correlations within the heterotrait-monomethod block of the multi-trait scoring group are larger than corresponding validity coefficients. Again, this suggests the presence of potential method variance or halo effects.

SEM Results

To further explore potential halo effects, we constructed two different structural equation models: one that included both trait and method factors (the CT-CM model) and one that included trait factors only. Table 4 presents fit indices for both the traits-only model and the CT-CM model. Comparing the two models, results suggest that fit is significantly improved by incorporating specific 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, which implies the presence of method-specific effects.

Table 4

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	NA	242.18^{***}
(df)		(9)

Note. χ^2 difference test is the Satorra-Bentler scaled chi-square difference test using the log-likelihood (Satorra & Bentler, 1999).
$^{***}p < .0001$

Table 5 reports estimated parameters for the traits-only model, which we review here because results provide a telling comparison with those from the CT-CM model. First, estimated factor loadings are positive and highly significant for all observed indicators, which suggest strong relationships between ratings and their respective traits. Interestingly, factor loadings for the multi-trait scoring group are uniformly larger than those for the single-trait scoring group, suggesting these scores are weighted more heavily within (and contribute more to) the composite trait factor. The proportion of variance explained for each observed indicator in the traits-only model ranges from .50 to .61 for the single-trait scores and from .75 to .81 for the multi-trait scores.

Table 5

Estimated Parameters for Traits Only Model

Observed Indicator	Factor Loading	Standard Error	R^2
Development			
Group 1 score	1.00	0.00	.50
Group 5 score	2.02***	0.14	.81
Organization			
Group 2 score	1.00	0.00	.56
Group 5 score	1.82***	0.10	.81
Voice			
Group 3 score	1.00	0.00	.55
Group 5 score	1.57***	0.08	.75
Conventions			
Group 4 score	1.00	0.00	.61
Group 5 score	1.51***	0.11	.78

Note. All factor loadings are unstandardized estimates.
***$p < .0001$

Table 6 reports the variance-covariance matrix for the latent factors (the psi matrix) in the traits-only model. Factor variances, reported in bold on the diagonal, are large in relation to their standard errors and are significantly different from zero. However, estimated between-trait correlations (corrected for measurement error) are excessively high, ranging from .75 to .99. Such high correlations suggest two possibilities: 1) different traits do not actually tap meaningfully different constructs and/or 2) estimated between-factor correlations are inflated by halo error affecting the multi-trait scoring group.

Table 6

Estimated Psi Matrix for Traits Only Model

Trait	Development	Organization	Voice	Conventions
Development	**3.33**	3.51	3.51	3.06
Organization	*.94*	**4.18**	4.04	3.88
Voice	*.97*	*1.00*	**3.95**	3.77
Conventions	*.75*	*.84*	*.84*	**5.07**

Note. Latent trait correlations are reported in italics below the diagonal. Factor variances are reported in bold on the diagonal. Latent trait covariances are reported above the diagonal.

Tables 7 and 8 report corresponding results for the CT-CM model. First, estimated factor loadings (reported in Table 7) are all positive and statistically significant for both trait factors and method factors. This confirms the viability of separate method factors, even when controlling for the underlying trait. In contrast to the traits-only model, trait factor loadings for the multi-trait scores are all *smaller* than those for the single-trait scores, with the exception of the Conventions factor, where this pattern is reversed. This result implies that the single-trait scores are weighted more heavily in the composite trait factors than the multi-trait scores for all traits except Conventions of Writing.

Table 7

Estimated Parameters for CT-CM Model

Trait	Observed Indicator	Factor Loading	Standard Error	R^2
Development	Group 1 score	1.00	0.00	.69
	Group 5 score	0.70***	0.14	.80
Organization	Group 2 score	1.00	0.00	.76
	Group 5 score	0.59***	0.12	.86
Voice	Group 3 score	1.00	0.00	.73
	Group 5 score	0.44***	0.07	.79
Conventions	Group 4 score	1.00	0.00	.69
	Group 5 score	1.15***	0.22	.78
Single-trait method	Group 1 score	1.00	0.00	.69
	Group 2 score	1.26***	0.09	.76
	Group 3 score	1.16***	0.09	.73
	Group 4 score	1.06***	0.09	.69
Multi-trait method	Group 5, trait 1	1.00	0.00	.80
	Group 5, trait 2	1.25***	0.09	.86
	Group 5, trait 3	0.98***	0.09	.79
	Group 5, trait 4	0.84***	0.06	.78

Note. All factor loadings are unstandardized estimates.
All R^2 indices account for both variance due to trait and variance due to method.
*** $p < .0001$

Table 8

Estimated Psi Matrix for CT-CM Model

Trait	Dev	Org	Voice	Con	Single	Multi
Dev	**2.67**	1.18	1.74	-0.27	0.00	0.00
Org	*.42*	**2.91**	1.74	0.49	0.00	0.00
Voice	*.66*	*.63*	**2.66**	0.46	0.00	0.00
Con	*-.11*	*.20*	*.19*	**2.15**	0.00	0.00
Single	*.00*	*.00*	*.00*	*.00*	**4.74**	6.93
Multi	*.00*	*.00*	*.00*	*.00*	*.91*	**12.21**

Note. Dev=Development, Org=Organization, Con=Conventions, Single=Single-trait, Multi=Multi-trait. Latent trait correlations are reported in italics below the diagonal. Factor variances are reported in bold on the diagonal. Latent trait covariances are reported above the diagonal.

Comparing the relative magnitude of factor loadings for the methods factors, it is evident that loadings for the single-trait method factor are all relatively similar. In contrast, factor loadings for the multi-trait method factor are relatively variable, with organization loading the strongest and conventions loading the weakest. Because the multi-trait method factor represents variance due to halo bias, this result suggests that the organization trait contributes most to halo, whereas the conventions trait contributes the least.

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 scores and from .78 to .86 for the multi-trait scores. The large disparity in R-squared indices for single-trait scores versus multi-trait scores observed in the traits-only model shrinks to some extent in the CT-CM model. This is

mainly due to the fact that we appear to be explaining more of the variance in the single-trait scores when we include separate methods factors than when we model only traits. Notably, variance explained for the multi-trait scores remains virtually the same when we include separate methods factors. This result is consistent with the argument that halo effect frequently masquerades as valid trait variance. However, the fact that we are explaining more of the variance in the single-trait scores with the CT-CM model is puzzling and suggests the presence of some method-specific variance arising from the single-trait scoring process.

Examining estimates from the psi matrix for the CT-CM model (reported in Table 8), one can see that all factor variances are large in relation to their standard errors and all are significantly different from zero. Interestingly, the variance of the single-trait scoring method factor is roughly twice the size of each of the trait factor variances, and the multi-trait scoring factor variance is between 4 and 6 times larger than each of the trait factor variances. This suggests that method-specific variance is contributing more to observed-score variance than is trait-specific variance. Comparing the magnitude of the estimated trait factor variances across the two models, one can see that these variances shrink when specific method factors are included in the model. This result is consistent with the halo effect: When specific method factors are not included in the model, method-specific variance masquerades as valid trait variance.

Finally, estimated between-trait correlations (corrected for both measurement error and halo bias) for the CT-CM model range from -.11 (for development – conventions) to .65 (for development – voice). These correlations are uniformly smaller than corresponding correlations from the traits-only model. Thus, between-trait

correlations appear to decrease when specific method factors are included in the model. This result is also consistent with the presence of a halo effect. When specific method factors are not included in the model, estimated between-trait correlations are inflated by halo error to the extent that scores on one trait affect scores on the other traits for raters who score multiple aspects of examinee performance. Interestingly, the two method factors are correlated very highly (.91), which suggests that method-specific variance in both scoring groups manifests itself in similar ways.

Discussion

This study sought to answer the question: *To what extent do dimensional analytic writing scores assigned by the same raters exhibit a halo effect?* Collectively, our results imply the presence of method-specific variance, and the pattern of parameter estimates across models strongly suggests a halo effect for those in the multi-group scoring model. The model incorporating both trait and method factors exhibits significantly better fit to the data than a model with trait factors only. Patterns of factor loadings, R-squared indices, factor variances, and between-trait correlations are all consistent with a halo effect for scorers rating multiple aspects of examinee performance. Moreover, results suggest that the halo effect does not affect all four traits equally. In particular, organization scores appear to contribute the most and conventions scores the least to the halo effect. This result is consistent with previous studies that have found inconsistent method effects across traits (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). It is possible that certain types of traits are inherently more influential on rater behavior than others.

It is also possible that variable factor loadings simply reflect the relative order in which scorers tend to evaluate the traits. If examinee performance on the first or second trait evaluated impacts ratings on subsequent traits, one might expect that the 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. More 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.

We find it somewhat odd that our results also identify method-specific variance for scorers in the single-trait scoring group. That is, when different raters assigned the trait ratings to a student, those scores were not locally independent. Moreover, the high correlation between the two method factors implies that the method-specific variance is impacting scores for both groups in similar ways. It is unclear what the cause of method-specific variance in single-trait scores could be. It could be due to the fact that all single-trait scoring groups utilized the same training and qualifying papers, although each group was ostensibly focused on scoring a different aspect of examinee performance. Or perhaps there is a more subtle halo effect operating even on scorers trained to evaluate only a single trait. Although they were not asked to explicitly rate other aspects of

examinee writing, perhaps they implicitly paid attention to these factors and allowed them to distort their ratings on their assigned dimension.

We speculate that this effect is similar to the violations of local independence assumptions that have been cited in applications of item response theory models to the analysis of ratings. Specifically, several researchers have identified violations of local independence when multiple raters assign scores to a single student response, and each of those researchers has proposed an alternative model that takes into account the covariance between raters (DeCarlo, Kim, & Johnson, 2011; Patz, Junker, Johnson, & Mariano, 2002; Wilson & Hoskens, 2001). Regardless, additional research is needed to investigate raters' cognitive processes when they are involved in scoring only a single trait to determine whether other aspects of an examinee's writing could be affecting scores.

The use of constructed-response (CR) item formats and performance-based assessments (PBAs) will most certainly increase in coming years as the Race to the Top Assessment consortia roll out the next generation of large-scale assessments. Assessment plans released by the consortia indicate an increased number of complex, multi-stage or "integrative" performance tasks than have been used on large-scale assessments in the past. Such tasks tend to create scores on multiple traits or multiple dimensions of performance for the same response. Although a single-trait-per-rater scoring design may minimize the risk of the halo effect, that design also significantly increases the costs associated with scoring. Hence, additional research is needed to determine whether rater training and monitoring can also reduce the halo effect. In addition, test developers should carefully consider how to best depict the traits for which scores are assigned to

writing assessments and other performance assessments because previous research raises compelling questions about the number of traits that raters can differentiate. This is especially important in light of the fact that assigning fewer scores to a student response would lower scoring costs. Furthermore, although automated scoring represents one potential strategy for avoiding halo, not all task types can be automatically scored using current technology, and scoring engines must be calibrated using human-assigned scores. Thus, human scoring will continue to be a requirement for assessment programs employing CR item types or PBAs.

References

Bechger, T.M., Maris, G., & Hsiao, Y.P. (2010). Detecting halo effects in performance based examinations. *Applied Psychological Measurement, 34*(607-619).

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cheung, G.W. (1999). Multifaceted conceptions of self-other ratings disagreement. *Personnel Psychology, 52*, 1-36.

Conway, J.M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology, 84*, 3-13.

DeCarlo, L.T., Kim, Y.K., & Johnson, M.S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*, 333-356.

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.

Hoyt, W.T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64-86.

Kozlowski, S.W., & Kirsch, M.P. (1987). The systematic distortion hypothesis, halo, and accuracy: An individual-level analysis. *Journal of Applied Psychology, 72*, 252-261.

Marsh, H.W., & Butler, S. (1984). Evaluating reading diagnostic tests: An application of confirmatory factor analysis to multitrait-multimethod data. *Applied Psychological Measurement, 8*, 307-320.

Marsh, H.W., & Yeung, A.S. (1997). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology, 89,* 41-54.

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.

Patz, R.J., Junker, B.W., Johnson, M.S., & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384.

Satorra, A., & Bentler, P.M. (1999). *A scaled difference chi-square test statistic for moment structure analysis*. Technical Report. Department of Statistics, University of California. Los Angeles.

Solomonson, A.L., & Lance, C.E. (1997). Examination of the relationship between true halo and halo error in performance ratings. *Journal of Applied Psychology, 82*(5), 665-674.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25-29.

Viswesvaran, C., Schmidt, F.L., & Ones, D.S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*(1), 108-131.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26,* 283-306.

Endnotes

[1] Groups 1 and 2 were located at the midwestern scoring center, and Groups 3 through 6 were located at the southwestern scoring center.