

Does an Automated Source Integration Score Respond to Revision? Extending Evidence of Construct Validity with a Feedback Experiment

Andrew Potter¹, Yu Tian¹, Kaitlin Van Houghton¹, Manmeet Singh¹,
Renu Balyan², Maria Goldshtein¹, Laura K. Allen³, Danielle S. McNamara¹

¹Learning Engineering Institute, Arizona State University, USA

²School of Arts and Sciences, SUNY Old Westbury, USA

³Department of Educational Psychology, University of Minnesota, USA

Abstract

This study examined whether an automated source integration measure responds to revision. In a randomized experiment, undergraduates drafted and revised a source-based essay. Source integration scores increased across drafts, although gains did not differ significantly between students who received source integration feedback and those who did not. Findings support use of the measure for formative evaluation and feedback.

1 Introduction

Source integration (i.e., the coordination, attribution, and synthesis of ideas across multiple documents) is central to academic writing and difficult to assess reliably (McCarthy et al., 2022; Vandermeulen et al., 2023). Automated writing evaluation (AWE) tools have largely not addressed connections between reading and writing (Allen et al., 2024). To address this gap, Potter, Tian, et al. (2026c) modeled source integration as a latent construct estimated from linguistic indices. The present study builds on that evidence by asking whether the measure registers what students change when they revise.

According to Kane (2013, 2016), a score is validated by how well the evidence supports the claims made for it. The claim made for this measure is that it can provide formative feedback to help students revise their source-based writing and improve source integration. Such a claim has two inferences: that the score is accurate and that the feedback derived from it is useful when students decide what to revise (Ranalli et al., 2017). The validation evidence described above addresses the first, and grounding the score in interpretable linguistic features also meets the expectation that automated scores be transparent and construct relevant (Ferrara and Qunbar, 2022). The second calls for a different kind of evidence, drawn from students revising with the score, as when draft-to-draft

improvement aligned with feedback received supported a formative validity argument for an AWE system (Correnti et al., 2022). Thus, the present study extends the validity argument by examining how the score behaves when undergraduates revise a source-based essay with formative feedback delivered under one of four randomly assigned conditions.

2 Prior Validity Evidence

Potter, Tian, et al. (2026c) developed the source integration measure for formative use in the Writing Analytics Toolkit (WAT). The development corpus comprised 2,643 source-based argumentative essays written by undergraduates and adult learners across eight prompts and scored by trained raters on a four-point source use and inferencing rubric, with acceptable interrater agreement (quadratic weighted $\kappa = .74$). Three claims were advanced and supported.

Construct representation. Source integration can be represented as a coherent latent construct rather than individual linguistic features. Of 41 candidate linguistic indices extracted with WAT, 11 were retained across four families, namely citation behavior, quotation behavior, verbatim overlap with the source texts, and semantic similarity with the source texts. A hierarchical confirmatory factor analysis fit with lavaan (Rosseel, 2012) modeled citations, quotations, and overlap as first-order factors; word2vec similarity loaded directly onto a second-order source integration factor. Fit was acceptable, $\chi^2(48) = 900.34$, CFI = .945, RMSEA = .082, and the factor predicted expert human ratings of source use and inferencing strongly, $\beta = .81$, $R^2 = .65$.

Generalizability. The construct functions consistently across prompts and populations. Measurement invariance testing across the development corpus and PERSUADE 2.0 (Crossley et al., 2024)

supported metric but not scalar invariance, indicating that factor loadings held across corpora while intercepts differed. The measure therefore supports the claim that the factor can be applied to novel essays.

Predictive utility. The construct improves prediction of human ratings. The strongest agreement came from concatenating the factor score with BERT embeddings of the essay (Devlin et al., 2019), $\kappa = .43$, quadratic weighted κ (QWK) = .68. That combination outperformed embeddings alone ($\kappa = .39$, QWK = .65), feature-based machine learning (QWK = .45 to .47), and rubric-driven GPT-5.2 scoring (QWK = .58).

In general, these claims establish that the measure represents its construct, can be generalized across tasks, and reflects judgement of expert trained human raters. All three bear on the accuracy inference. The second inference, about usefulness when students revise, is addressed in the present study.

3 The Present Study

The present study extends prior validation work by examining whether automated source integration scores support an inference required for formative evaluation and feedback, that draft-to-draft score change can be interpreted as change in students' source integration. Following Kane (2013, 2016), we treat this as an interpretation and use claim and ask how strongly the available evidence supports it.

Evidence is stronger if scores increase from first to revised draft, if that increase remains after accounting for changes in essay length and overall writing quality, and if gains are larger among students whose feedback included source integration. The covariate adjustment addresses two rival interpretations, that score change reflects essay expansion or general writing improvement. Feedback condition corroborates rather than isolates, because source integration appeared alongside three other metrics.

RQ1. Responsiveness to revision and feedback context. To what extent do source integration scores increase from first to revised draft, does that increase remain after accounting for changes in essay length and overall writing quality, and does it differ between students whose feedback included source integration and students whose feedback did not?

RQ2. Comparison of score versions. Prior work compared the factor score and the hybrid score on single drafts, so this question asks whether the evidence for the claim depends on which is used. Do the two versions differ in responsiveness to revision, sensitivity to feedback context, and resistance to changes in essay length and overall writing quality?

4 Method

4.1 Participants

Participants were 138 undergraduates at a large public university in the southwestern United States who took part for course credit (M age = 20.04, SD = 1.56, $n = 137$ reporting; 57.2% men, 42.8% women; 50.0% White, 21.7% multiracial, 13.0% Hispanic or Latino, 10.9% Asian, 4.4% Black or other; 33.3% grew up speaking a language other than or in addition to English). They were recruited through the university research participation system and randomly assigned to one of four between-subjects feedback conditions. Chi-square tests and one-way analyses of variance confirmed no condition differences in demographic or self-efficacy measures. Two participants lacked complete data for both drafts, leaving 136 participants in the analytic sample.

4.2 Research Design

The study used a randomized between-subjects design with four feedback conditions. All participants wrote and revised in the same platform and completed identical tasks; conditions differed only in the feedback available during revision. Participants were assigned to a control condition ($n = 36$), in which automated metrics were disabled; a WAT-only condition ($n = 33$), in which students saw platform metrics; a chatbot-only condition ($n = 37$), in which students received conversational feedback from a large language model; or a combined condition ($n = 32$), in which students received both forms of feedback. After excluding two participants with incomplete draft data, the analytic sample included 35 control, 33 WAT-only, 36 chatbot-only, and 32 combined-condition participants. Chi-square tests and one-way analyses of variance indicated no condition differences in demographics.

4.3 Materials

Writing task. Students wrote a source-based argumentative essay on teacher retention, taking a

position on why teachers leave the profession and how the problem should be addressed and supporting that position with two sources adapted from online publications, a think-tank blog on national teacher shortages and a newspaper article on teacher burnout. The topic is not represented among the eight prompts of the corpus used to develop and validate the measure, so it is a genuinely unseen task for the scoring model.

WAT-Classroom. WAT-C (McNamara et al., 2026) is a web-based writing platform that provides automated analytics on student writing. Four metrics were enabled for this study, namely source integration, language variety, development of ideas, and academic writing, each shown as a percentile with a short description of what the metric captures and what a writer might do to strengthen it. The panel also displayed a holistic performance score and a word count. The source integration percentile is the rank transform of the same latent factor score evaluated here, so students in the WAT-feedback conditions saw, and could act on, one of the two score versions under comparison. The remaining three metrics are derived by principal component analysis of 72 linguistic features computed over a corpus of source-based essays scored by trained expert human raters (Potter, Crossley, et al., 2026). See Potter, Serhan, et al. (2026) for the design of the panel and the wording of its feedback.

WAT-Chatbot. A pilot conversational agent built with ASU CreateAI Builder on GPT-5.2 (temperature = 0.5) complemented the platform. Its system prompt directed it to offer revision-oriented strategies, to emphasize accurate source use, clarity, and argumentation, and to redirect any request to rewrite text back to the student's own revision, scaffolding revision without off-loading the writing.

4.4 Procedure

Participants completed the study asynchronously in a single online session of approximately 85 minutes. After an orientation to the feedback tools available in their condition, they wrote a first draft with no access to any tool, practiced by revising a peer essay using their condition's feedback, then revised their own essay and completed a survey. Because the first draft was written without feedback and the second was written with it, draft-to-draft change indexes revision under the assigned condition.

4.5 Measures

Source integration. All 272 essays (136 students $\times$ 2 drafts) were processed through WAT and scored twice. The factor score was obtained by applying the published scaler and the confirmatory model described in Section 2. The hybrid score was obtained from the architecture that performed best in the earlier predictive comparison, a BERT model that concatenates essay embeddings with the four factor scores, trained on all eight prompts of the 2,643-essay development corpus and producing a value on the scale of the averaged human rating (Potter, Tian, et al., 2026).

Because teacher retention is a new topic, no human source integration ratings exist for these essays and scoring accuracy on this prompt cannot be observed directly. We estimated it by simulating the same situation inside the development corpus, training models on six prompts and evaluating them on the two prompts they had never seen (see Potter, Tian, et al., 2026, for details). Agreement with human ratings was QWK = .64 on one held-out prompt and .43 on the other (pooled QWK = .61, $r = .72$, $n = 791$), below the QWK = .68 reported for a random split in which every prompt appears during training. These results suggest that cross-prompt transfer is more difficult than random-split estimates imply, so the pooled held-out agreement provides the more conservative benchmark for the present study. The held-out models were used only to estimate expected transfer accuracy, and all study essays were scored with the final model trained on the full eight-prompt development corpus.

The two score versions are computed on different scales; therefore, their raw values are not directly comparable. The hybrid score was trained on the averaged human rating, which ranged from 0.5 to 4.0 in half-point steps across the 2,643 development essays. That target was min-max normalized for training, and predictions were mapped back to the original scale; thus, hybrid values are reported in rubric points. In this sample, the hybrid score ranged from 1.14 to 3.33 ($M = 2.64$, $SD = 0.35$), and a change of 0.10 rubric points is roughly one-third of a within-sample standard deviation. The factor score is a regression factor score estimated from indices standardized against the development corpus. It has no rubric anchor and, like all regression factor scores, is shrunken toward the corpus mean, which is why its dispersion is narrow (range = -0.56 to 1.42 , $M = 0.06$, $SD = 0.20$). Raw

values were analyzed for both versions, and standardized change (d_z) is the metric on which the two can be compared directly.

Covariates. Both drafts were also scored by WAT-C for holistic writing quality and word count. WAT-C holistic scores show strong agreement with expert-trained human raters on source-based writing tasks (McNamara et al., 2026).

Revision behavior. Edit operations between drafts were computed with TRACE (Tian et al., in preparation), which estimates text change at the span, sentence, and token levels and returns counts of insertions, deletions, replacements, and moves. Revision counts were available for 133 participants and were right-skewed (total operations $M = 15.31$, $SD = 20.25$, $Mdn = 8$, range 0 to 134).

4.6 Data Analysis

All analyses were conducted separately for the factor score and the hybrid score. Draft-to-draft change was tested with paired-samples t tests and, because change scores were not normally distributed, with Wilcoxon signed-rank tests on the same pairs. To examine whether score change reflected source integration rather than general essay expansion or improvement, we regressed score change on change in essay length and change in holistic writing quality. The adjusted intercept estimated the mean gain remaining after both covariates were accounted for. Feedback context was tested with an analysis of covariance predicting the revised-draft score from the first-draft score and whether the student received feedback that included source integration.

5 Results

5.1 RQ1. Responsiveness to Revision and Feedback Context

Scores increased from first to revised draft on both versions. On the hybrid score, essays gained $M = 0.076$ rubric points, 95% CI [0.045, 0.107], $t(135) = 4.82$, $p < .001$, $d_z = 0.41$, with the rank test in agreement, $z = -5.23$, $p < .001$. On the factor score, the two tests diverged: the paired-samples t test was not significant, $M = 0.017$ factor-score units, $t(135) = 1.46$, $p = .148$, whereas the Wilcoxon test was significant, $z = -4.25$, $p < .001$.

The divergence between the parametric and rank-based tests was traced to one influential participant. The participant's first draft was 538 words, 469 of

which (87%) appeared in properly attributed quotations, indicating extensive but credited source use rather than undisclosed copying. The revision reduced the essay to 277 words and removed every quotation; the factor score decreased by 1.34 factor-score units, relative to a sample standard deviation of 0.13. That value inflated the standard deviation on which the parametric test depends, whereas the rank-based test was less sensitive to it. The exclusion was not prespecified, and because the case is substantively interpretable rather than an artifact, Table 2 reports each test with and without it. Excluding this participant, both score versions increased under both tests, hybrid $t(134) = 5.42$ and factor $t(134) = 4.70$, both $p < .001$. Results below are reported for the full sample, and the exclusion is noted only where it changes a conclusion.

Score gains were partly, but not wholly, associated with essay expansion and overall writing quality. Revised drafts were longer ($M = +18.4$ words) and received higher holistic scores ($M = +0.06$). Score change correlated moderately with both length and quality change for the hybrid score ($r = .42$ and $.40$) and somewhat less strongly for the factor score ($r = .31$ and $.27$), indicating that part of the gain was shared with general essay improvement. After simultaneous adjustment for change in length and holistic quality, the hybrid gain remained significant, $+0.058$, $t(131) = 3.91$, $p < .001$, roughly three-quarters of the unadjusted effect. The factor gain did not remain significant in the full sample, $+0.007$, $t(131) = 0.62$, $p = .539$, although it did after the influential case was excluded, $+0.024$, $t(130) = 4.06$, $p < .001$.

There were no significant differences in source integration scores by feedback condition. Table 1 reports change by condition. Controlling for the draft-1 score, students whose feedback included source integration did not gain more than students whose feedback did not on the hybrid score, $b = 0.033$, $t(133) = 1.16$, $p = .249$. On the factor score, the coefficient was marginal, $b = 0.035$, $t(133) = 1.84$, $p = .067$, and reached significance only when the influential case was excluded. Score change was also unrelated to how much students revised. Edit operations between drafts, computed with TRACE (Tian et al., in preparation), were right-skewed ($M = 15.31$, $SD = 20.25$, $Mdn = 8$, range = 0 to 134, $n = 133$) and did not predict score change on either version (hybrid $r = -.02$, factor $\rho = .15$).

The feedback result should be interpreted cau-

tiously for three reasons. First, source integration was one of four metrics on the panel, so students were not directed specifically toward source use and any construct-specific effect may have been diluted. Second, with group sizes of 65 and 71, the design had 80% power to detect $d = .48$, whereas the observed adjusted effects were smaller ($d = .20$ and $d = .32$). The study was therefore underpowered for effects of the observed size. Third, the largest gain occurred in the combined condition (+0.128), but the control condition also gained substantially (+0.090), a pattern more consistent with revision itself driving change than with any particular feedback source.

The core evidence supports the claim that score change can be interpreted as revision-related change in source integration. Scores increased when students revised, and, for the primary measure, the increase remained after adjustment for essay length and overall writing quality. The feedback comparison did not provide corroborating evidence. Thus, the evidence supports reading score change as change in source integration, but not attributing that change to the feedback students received.

5.2 RQ2. Comparison of Score Versions

The two score versions were related but not redundant. They correlated moderately on draft-1 scores ($r = .63$), draft-2 scores ($r = .61$), and change scores ($r = .59$), leaving substantial independent variance.

For responsiveness to revision, the hybrid score detected change under both the parametric and rank-based tests in the full sample, whereas the factor score did so only under the rank-based test. For resistance to construct-irrelevant change, the hybrid gain remained significant after adjustment for essay length and holistic quality, whereas the factor gain did not. For sensitivity to feedback context, neither score version showed an effect that held across specifications.

The difference between the score versions was driven by the factor score's sensitivity to one influential case. Removing one participant in 136 halved the factor score's change-score standard deviation from 0.135 to 0.066, increased its paired t from 1.46 to 4.70, changed its adjusted gain from $p = .539$ to $p < .001$, and changed its condition test from $p = .067$ to $p = .005$. In contrast, the hybrid score shifted only from $t = 4.82$ to 5.42, and its condition test remained nonsignificant.

The case illustrates why the factor score was

less stable. The factor score is computed from citation, quotation, and overlap indices. A draft composed of 87% quoted text can therefore score near the top of the distribution on quotation and overlap, even if that material is not well integrated into an argument. This limitation was anticipated in the validation study, where overlap indices loaded positively on the construct despite their theoretical association with weaker integration, because such indices cannot distinguish unproductive copying from legitimate quotation (Potter, Tian, et al., 2026). The hybrid score also incorporates the essay text, and its estimate for this student fell by 0.63 rubric points, compared with a 1.34-unit drop on the factor score.

One asymmetry should be interpreted cautiously. Students saw the factor score on the feedback panel, and only the factor score's condition coefficient approached, then crossed, the conventional significance threshold after the influential case was removed. Because the effect was small and depended on one case, we treat it as a direction for further study rather than as evidence of a feedback effect.

Overall, both score versions provided evidence of responsiveness to revision, but the hybrid score provided more stable evidence across the full set of tests. The evidence for revision-related change did not depend strongly on score version once the influential case was set aside. However, the evidence for resistance to construct-irrelevant change and sensitivity to feedback context did depend on score version, with the factor score's conclusions turning on a single participant.

6 Discussion

Potter, Tian, et al. (2026) established evidence of construct validity for the source integration measure by demonstrating that it represents its construct, functions consistently across corpora, and predicts expert human ratings. However, that evidence concerned single drafts. The present study extends this validity evidence to a second occasion of measurement in which students revised a source-based essay with or without feedback on source integration. Scores increased when students revised, and, for the hybrid score, the increase remained after adjustment for changes in essay length and holistic writing quality. These findings support interpreting draft-to-draft score change as change in students' source integration, an inference required for use in formative evaluation and feedback.

However, gains in source integration scores were not larger among students whose feedback included source integration. Three design features qualify this finding. First, source integration appeared alongside three other metrics, so students were not directed specifically toward source use. Second, the study was underpowered for effects of the observed size. Third, scores increased in every condition, including the control condition, suggesting that revision itself accounted for much of the change. Isolating a construct-specific feedback effect would require a design in which source integration feedback is the only feedback feature that varies. Future experimental studies should isolate source integration feedback from other feedback dimensions, vary its presence systematically, and test whether students who receive targeted source integration feedback show larger score gains than students who receive either no automated feedback or feedback on other writing dimensions.

Furthermore, the method in which the source integration score was calculated diverged in ways that are important for future use in AWE systems. One participant in 136 moved the factor score's adjusted gain across the conventional significance threshold, while leaving the hybrid score largely unchanged. That participant quoted sources extensively in the first draft and removed all quotations in the revision. This pattern is consistent with Potter, Tian, et al. (2026), who noted that overlap indices loaded positively on the construct because they cannot distinguish unproductive copying from legitimate quotation. Indices counted without reference to the surrounding text can register a heavily quoted draft as strong source use, whereas the essay text contains information those counts omit. These results extend the earlier finding that the hybrid architecture best predicted human ratings by showing that its advantage also appears across occasions of measurement.

6.1 Limitations

Several limitations qualify the findings and point to future research. Human source integration ratings were unavailable for the study essays, so scoring accuracy on this prompt was estimated through cross-prompt transfer rather than observed directly. In addition, the covariates used to evaluate rival interpretations, essay length and holistic writing quality, were automated. These constraints limit conclusions about scoring accuracy on the present prompt.

The design also limits generalizability. Students revised once, on one prompt, therefore the findings are related to a single revision occasion rather than to development over time. Future studies should follow writers across multiple source-based tasks and revision cycles to test whether score change accumulates in the direction instruction intends.

7 Conclusion

The validity argument for this measure now includes evidence from both single-draft scoring and revision-based use. Potter, Tian, et al. (2026) established that source integration can be represented as a coherent latent construct, that the construct generalizes across prompts and populations, and that it improves prediction of human ratings, especially when combined with essay embeddings. The present study adds evidence that draft-to-draft change in the score can be interpreted as change in students' source integration. This interpretation is qualified by the absence of corroborating evidence from feedback condition and by the fact that the evidence was stronger and more stable for the hybrid score than for the factor score.

Two implications follow for practice. First, responsiveness to revision is a distinct source of validity evidence for scores intended to support formative evaluation and feedback, and it should be demonstrated rather than inferred from single-draft accuracy. Second, decisions about score deployment should consider how candidate score versions behave across drafts, because formative interpretation depends on score change during revision.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180261 to Arizona State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

- Laura K. Allen, P. Öncel, and L. E. Flynn. 2024. Expanding AWE to incorporate reading and writing evaluation. In M. D. Shermis and J. Wilson, editors, *The Routledge International Handbook of Automated Essay Evaluation*, pages 575–589. Routledge.
- R. Correnti, L. C. Matsumura, E. L. Wang, D. Litman, and H. Zhang. 2022. Building a validity argument for

- an automated writing evaluation system (eRevise) as a formative assessment. *Computers and Education Open*, 3:100084.
- S. A. Crossley, P. Baffour, Y. Tian, A. Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, 61:100865.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics.
- S. Ferrara and S. Qunbar. 2022. Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, 59(3):288–313.
- M. T. Kane. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73.
- M. T. Kane. 2016. Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2):198–211.
- K. S. McCarthy, E. F. Yan, L. K. Allen, A. N. Sonia, J. P. Magliano, and D. S. McNamara. 2022. On the basis of source: Impacts of individual differences on multiple-document integrated reading and writing tasks. *Learning and Instruction*, 79:101599.
- D. S. McNamara, A. Potter, Z. Serhan, M. Singh, N. Patne, T. Arner, R. D. Roscoe, L. K. Allen, and S. A. Crossley. 2026. The writing analytics tool: A learning engineering approach to designing AI-supported writing instruction. In *Proceedings of the Learning Engineering Research Network Convening (LERN 2026)*. EdTech Archives.
- A. H. Potter, S. A. Crossley, P. Öncel, L. A. Allen, X. Xu, R. Balyan, T. Arner, R. D. Roscoe, and D. S. McNamara. 2026a. From words to components: Mapping linguistic features in source-based writing. *Assessing Writing*, 70:101108.
- A. H. Potter, Z. Serhan, N. A. Patne, P. Öncel, I. Ahmed, T. Arner, R. Islam, R. D. Roscoe, L. K. Allen, S. A. Crossley, and D. S. McNamara. 2026b. Human-AI collaboration for qualitative analysis in participatory design: Refining the writing analytics tool. *Journal of Educational Data Mining*, 18:113–155.
- A. H. Potter, Y. Tian, R. Balyan, M. Goldshtein, L. K. Allen, and D. S. McNamara. 2026c. Source integration in source-based writing: From theoretical modeling to predictive analytics. *Assessing Writing*, 69:101081.
- J. Ranalli, S. Link, and E. Chukharev-Hudilainen. 2017. Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37:8–25.
- Y. Rosseel. 2012. lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.
- Y. Tian, K. Christhlf, A. Potter, S. Graham, J. Early, and D. S. McNamara. in preparation. TRACE: Automated multi-granularity analysis of text revision. Manuscript in preparation.
- N. Vandermeulen, E. Van Steendam, and G. Rijlaarsdam. 2023. Introduction to the special issue on synthesis tasks: Where reading and writing meet. *Reading and Writing*, 36(4):747–768.

Table 1

Mean draft-1 to draft-2 change by feedback condition. *Ns* are the analyzed sample.

Condition	<i>n</i>	Hybrid Δ	Factor Δ
Control	35	+0.090	+0.025
WAT-Only	33	+0.053	−0.004
Chatbot-Only	36	+0.037	+0.006
WAT+Chatbot	32	+0.128	+0.041
Overall	136	+0.076	+0.017

Table 2

Change, adjustment, and condition tests for both score versions, on the full sample and excluding the influential case.

Statistic	Hybrid		Factor	
	Full	−Out	Full	−Out
Change <i>M</i>	+.076	+.081	+.017	+.027
Change <i>SD</i>	.184	.174	.135	.066
Paired <i>t</i>	4.82***	5.42***	1.46	4.70***
Wilcoxon <i>z</i>	−5.23***	−5.44***	−4.25***	−4.45***
<i>d_z</i>	.41	.47	.12	.40
Adjusted <i>b</i>	+.058	+.064	+.007	+.024
Adjusted <i>t</i>	3.91***	4.41***	0.62	4.06***
ANCOVA <i>b</i>	+.033	+.042	+.035	+.032
ANCOVA <i>t</i>	1.16	1.53	1.84	2.88**

Note. −Out = excluding the influential case. Hybrid values are in rubric points and factor values are in factor-score units, so only *d_z* is comparable across versions. Adjusted rows give the intercept from regressing score change on change in essay length and change in holistic quality. ANCOVA coefficients are for source integration feedback controlling the draft-1 score. ***p* < .01. ****p* < .001.