Metacognition and Learning (2019) 14:131–166
https://doi.org/10.1007/s11409-019-09201-y

Calibration in multiple text use

Ying Wang[1] · Alexandra List[1]

Received: 2 November 2018 / Accepted: 13 August 2019 / Published online: 2 November 2019

Abstract

The literature on calibration suggests that students consider a multitude of factors when they self-evaluate task performance. Nevertheless, few studies have focused on calibration within a complex task enviornment, such as when students are asked to compose written responses based on multiple texts. In this study, we examined the criteria that undergraduate students considered when they were asked to self-evaluate their written responses, composed based on multiple texts. Moreover, we considered the extent to which these criteria had an effect on students' objective response quality, calibration accuracy, and confidence bias. Findings revealed that students indeed cited a variety of criteria in justifying their self-evaluations including task-, context-, and person-related factors, consistent with prior research. Further, our study indicated that high quality written responses were associated with accurate calibration and with students' relative under-confidence. We further found that low-performing students demonstrated less accurate calibration and greater over-confidence. Implications for improving students' metacognitive awareness during complex task completion are discussed.

Keywords Self-evaluation · Calibration · Confidence bias · Writing composition · Multiple texts

Introduction

Multiple text use is a key competency necessary for learning in the twenty-first century (List and Alexander 2017; Goldman and Scardamalia 2013). At the same time, students, even at the undergraduate level, have been found to struggle with multiple text use, including with composing written responses based on multiple texts (Anmarkrud et al. 2014; Authors, 2019c; Mateos and Solé 2009). Specifically, students

✉ Alexandra List
azl261@psu.edu

Ying Wang
yqw5386@psu.edu

[1] Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, State College, PA, USA

have been found to experience challenges with both the superficial (e.g., mechanics, like spelling and grammar, Cavaleri and Dianati 2016; Fallahi et al. 2006) and deeper-level (e.g., evaluating and integrating information) aspects of response composition (Du & List, under review; List et al. 2017; Britt and Aglinskas 2002; Strømsø et al. 2013; Wiley and Voss 1999; Wiley et al. 2009). This may be explained in a variety of ways. For one, students may become overwhelmed by the informational demands of writing based on multiple texts (Cerdán and Vidal-Abarca 2008). For another, students may not have the skills necessary to engage in multiple text integration (Britt and Sommer 2004). At the same time, another possibility is that, while students are able to somewhat effectively integrate multiple texts during processing, this integration fails to transfer to the written responses that they compose (Wolfe and Goldman 2005). In part, this failure of transfer may be due to students' deficits in conceptualizing what may constitute a quality written response based on multiple texts (List, Du, & Wang (under review); List and Alexander 2015). In this study we examine this possibility. Specifically, we identify the criteria that students draw on in evaluating the quality of their written responses and examine the association between the self-evaluation criteria that students use to determine response quality and their actual task performance.

Multiple text use

While a variety of models have been proposed to conceptualize students' multiple text use, the most prominent of these is the Multiple Documents Task-Based Relevance Assessment and Content Extraction (MD-TRACE) model of multiple texts (Rouet and Britt 2011; Rouet 2006). The MD-TRACE suggests that students' process of multiple text use unfolds through a series of five, recursive steps. In Step 1, students form an initial cognitive representation of the assignment based on the task demands provided (i.e., construct a task model), at the same time conceptualizing the external (e.g., texts) and internal resources (e.g., prior knowledge) available to them to aid in task completion. In Step 2, students determine that they have an information need, requiring them to access information, beyond their prior knowledge. This information need serves to initiate the process of multiple text use. Step 3, therefore, is devoted to students' engagement with multiple texts, including information selection, processing, and integration. Specifically, in integrating multiple texts, students may be expected to construct a *documents model* or a cognitive representation of content presented across texts and the associations among them (Britt et al. 1999; Perfetti et al. 1999). In Step 4 of the MD-TRACE, students are expected to produce a task product, typically a written response based on multiple texts. Finally, in Step 5, students evaluate the quality of their task product, relative to the task model developed in Step 1, and determine either that the response they composed is sufficient to meet task demands or decide to iteratively return to earlier stages in the model.

Step 5 of the MD-TRACE, or students' self-evaluations of their task products, is the focus of the present paper. We consider Step 5 to be a critical, albeit under-examined, step of the multiple text use process, related to students' judgments of their own performance and to students' potential re-engagement in processes associated with selecting, evaluating, and integrating earlier accessed texts. While students' self-evaluations of written responses composed based on multiple texts have received limited consideration in the literature (List and Alexander 2015), students' judgments of performance have been emphasized in the literatures on self-regulation and calibration.

Task type in multiple texts

Work on multiple text use has long considered differences in task assignment to be associated with differences in multiple text processing and performance (Gil et al. 2010; Le Bigot and Rouet 2007; Wiley and Voss 1996, 1999). Specifically, Wiley and Voss (1996, 1999) suggest that, in part, multiple text tasks differ in asking students to engage in either “knowledge-telling” or “knowledge-transforming,” with these two task types tapping different levels of understanding on the part of learners. For instance, tasks asking students to write a narrative response (i.e., to engage in knowledge-telling) may be considered to only require the restating of text-based information and to produce only superficial learning. As a contrast, writing an argument (i.e., knowledge-transforming) is considered to require reorganizing and reformulating text-based information in support of a claim, producing deep-level understanding. Indeed, in a seminal study, Wiley and Voss (1999) found that college students who were asked to write an argument based on multiple texts produced more integrated responses, with a greater number of causal and transformed statements included, as compared to students who were asked to write narrative, summary, or opinion-based essays. Bråten and Strømsø (2009) similarly found that students who were asked to construct an argument based on multiple texts developed deeper-level and more integrated understanding than students asked to report their global understanding. These studies echo the preponderance of empirical evidence indicating that students’ understandings of multiple text use and writing quality differ in association with task assignment (Cerdán and Vidal-Abarca 2008; Gil et al. 2010).

In a recent study, List, Du, and Wang (under review) examined college students’ conceptions of five common academic tasks. In particular, students were asked to report what composing an argument, essay, opinion, summary or research report entailed. Results showed that students had different conceptions regarding each of these five tasks. In particular, students’ conceptions varied in the degree to which they considered different writing assignments to require that they defend a personal opinion or be “objective” and that they provide text-based evidence or not. Of particular interest to us in this study are the differences in students’ conceptualizations of argument tasks vis-à-vis research reports. Although argument tasks and research reports may be considered types of “knowledge-transforming” tasks in Wiley and Voss (1999) parlance, students in List et al.'s (under review) study conceptualized argument tasks as requiring them to take a stance, supported with evidence. This was reflected in responses such as, "I think this is asking me to pick one side of the topic I choose and defend it and provide support as to why it is the right choice." At the same time, students conceptualized writing a research report to require research, source use, and citation; multiple text use behaviors typically considered to be associated with argument composition in prior work. This was reflected in students reporting that writing a research report required them to: *use outside articles, normally peer reviewed, to come to a conclusion based on my topic.* In this study, we build on these prior results to directly compare students’ response composition and self-perceptions of response quality when asked to write an argument vis-à-vis a research report. To the extent that List et al. (under review) found students to have different conceptions of argument tasks vis-à-vis research reports, we extend this prior work to consider whether differences in task assignment also result in differential task performance and in students' differential self-evaluations of response quality.

Calibration

Across a variety of domains, *calibration*, or the association between students' subjective assessments of their performance and their objective performance on corresponding tasks, has been strongly associated with learning (Keren 1991; Nietfeld et al. 2006b; Pieschl 2009). In particular, students who are better calibrated, or more accurate in predicting their task performance, tend to have improved academic outcomes (e.g., Bol et al. 2012; Dole et al. 1991; Hadwin and Webster 2013). This may be the case for a number of reasons. For one, calibration may be associated with students' monitoring of their performance during the course of task completion, allowing students to better identify potential gaps in understanding and to deploy strategies, aimed at improving comprehension, accordingly (Baker 1989; Flavell 1979; Glenberg and Epstein 1985; Huff and Nietfeld 2009). Indeed, calibration and students' improved monitoring during task completion have been found to be associated with more purposeful and effective strategy selection and use (Pressley and Ghatala 1990; Pieschl 2009; Winne and Jamieson-Noel 2002). For another, improved calibration may indicate more strategic deployment of motivational and cognitive resources (Sperling et al. 2004; Stone 2000; Winne and Jamieson-Noel 2002). Finally, better calibration may be associated with students' better understanding of task demands, improving performance (Dinsmore and Parkinson 2013; Kulhavy and Stock 1989; Lin and Zabrucky 1998; Pieschl 2009; Winne and Hadwin 1998). This particular aspect of calibration is examined in this study. Specifically, we seek to identify the criteria that students use to judge their performance on a multiple text writing task and to determine the extent to which these are associated with objective measures of task performance. More generally, we seek to understand whether students more accurately gauging their performance on a multiple text writing task, do, indeed, perform better.

Nevertheless, the literature on calibration has been limited in two ways. First, students have been asked to calibrate, or formulate confidence judgments, of their performance on discrete or fairly low-level tasks (e.g., multiple-choice questions, Higham 2013), rather than on more complex tasks (e.g., writing papers), typical of those commonly assigned in academic contexts (List and Alexander 2015). Second, studies of calibration have focused on computing the discrepancy between measures of students' subjective and objective task performance (e.g., Lichtenstein and Fischhoff 1977; Schraw et al. 2013), neglecting the cognitive and metacognitive processes underlying students' judgments of calibration (Dinsmore and Parkinson 2013). In this study we aim to address these limitations by examining the accuracy of students' judgments of performance on a complex task (i.e., composing a written response based on multiple texts) and by considering the criteria that students use to form these judgments. This represents a critical area of investigation since students may experience particular difficulties in forming judgments of performance on complex tasks. Indeed, forming such judgments requires students to identify, interpret, and evaluate their performance on a variety of task sub-components and to synthesize these in formulating an overall self-evaluation or judgment of performance (Pieschl 2009). These difficulties in self-evaluation demand not only the investigation of the overall accuracy of students' judgments of performance but also their basis.

Formation of judgments of performance

Students have been found to consider a variety of criteria when forming judgments of performance or self-evaluating their task products. Lin and Zabrucky (1998) in a review of the literature on students' calibration of reading comprehension, found students' judgments of performance to be based on factors associated with the task, text, and the individual. Task-related factors reflected students' judgments of performance that were based on features of the assignment that students were asked to complete (e.g., discrete or open-ended questions) or students' analysis of task difficulty. Text-related factors included calibration judgments based on characteristics of the texts that students were asked to read, including text genre and reading difficulty. Finally, individual-related factors were reflected in calibration judgments rendered based on students' assessment of their personal characteristics, like prior knowledge or reading ability. Although identified as somewhat distinct factors contributing to the formation of calibration judgments, Lin and Zabrucky (1998) emphasized that task, text, and individual-related factors all interact with one another in students' formation of judgments of performance.

Dinsmore and Parkinson (2013) identified similar factors as contributing to students' judgments of calibration. Specifically, they identified five categories in students' justifications for confidence judgments. These were justifications based on prior knowledge, characteristics of the text, item characteristics, guessing, and an "other" category. More importantly, they demonstrated that students considered multiple factors when rendering confidence judgments. Indeed, evaluations jointly rendered based on multiple characteristics may be expected to be more accurate than judgement that are based only on one factor, alone (Pieschl 2009).

Present study

The dimensions identified by Lin and Zabrucky (1998) and Dinsmore and Parkinson (2013) provide initial insights into the factors that students may consider in rendering judgments of performance. Nevertheless, these dimensions were only examined within the context of students judging their performance on fairly simple reading comprehension measures (e.g., multiple choice questions, single inference verification items). More needs to be understood regarding the extent to which these same dimensions are invoked and comparably conceptualized when students judge their performance on more complex or open-ended tasks.

There are a number of reasons why students' judgments of performance, formed in response to open-ended tasks, should be considered separately from judgments of performance rendered in response to multiple choice and other discreet items. Mosenthal (1998) suggests that open-ended tasks (e.g., writing assignments) are distinguishable from discrete tasks (e.g., multiple choice questions) in a number of ways. To start, open-ended tasks present students with more ambiguity regarding what task demands may be, whereas discrete tasks are defined by their presentation of a more limited problem space for learners. As such, open-ended questions demand that students drawn on their prior knowledge and make a variety of inferences to self-determine what task parameters may be. Moreover, open-ended tasks are more complex because they require students to complete and coordinate more aspects of the task than is necessary to correctly respond to a discrete question. For instance, within the context of multiple text use, composing a quality open-ended response requires the consideration and integration of information presented across texts and its synthesis into a written response, whereas responding to a multiple choice item has typically been conceptualized as a more limited information-location task (Cerdán and Vidal-Abarca 2008). In other words, in composing an open-ended response, students may need to generate a greater number of inferences to connect information, as compared to when completing a discrete task. Relatedly, open-ended tasks are scored comprehensively, based on a variety of criteria (e.g., organization, mechanics, evidence quality), rather than dichotomously, as correct or incorrect. While students have well-defined schema for responding to discrete questions (e.g., multiple choice items), open-ended questions vary more widely, limiting the response schema that students may be able to draw on in selecting criteria for self-evaluation. Collectively, the complexity of open-ended tasks and the variation in their scoring (Firetto (forthcoming)) may make judgments of performance more difficult for students to render, such as when students are asked to self-evaluate their writing performance based on multiple texts.

Work on multiple text use per se suggests a number of reasons why it is considered to be a quite challenge academic task for many students. Processing data, gathered through think-alouds and log data, have found students' multiple text use to simultaneously require constructive-integrative, critical-analytic, and metacognitive processing (Afflerbach and Cho 2009; Anmarkrud et al. 2014; Authors, 2017; 2019a; Cho and Afflerbach 2017). This includes determining information relevance, monitoring comprehension, forming connections across disparate texts, and considering source information in guiding the interpretation of texts. Scoring rubrics, used to rate students' written responses composed based on multiple texts, have implicitly suggested that multiple text use further requires students to draw a conclusion or formulate an argument based on information provided across texts, explain and provide evidence for their chosen argument in an inter-connected fashion, and consider counter-arguments and conflicting perspectives, and rebut these to varying extents (Anmarkrud et al. 2014).

In this study, we are particularly interested in examining the task criteria that students attend to when making judgments about their performance on a complex writing task addressing a controversial topic (i.e., the threats of and solutions to overpopulation). In a recent addition to the MD-TRACE, Rouet et al. (2017) introduced the RESOLV Model of purposeful reading, as a way of emphasizing the contextual factors that students may have available to them and may attend to when completing multiple texts tasks. Specifically, Rouet et al. (2017) provide a taxonomy of the contextual factors that may have a bearing on students' engagement with multiple texts. These factors include the request (e.g., the task for reading), the requester (e.g., teacher), and the audience benefitting from task completion, as well as the supports and obstacles (e.g., texts provided, time limits) available in the task environment. All-together these factors are represented by learners in their cognitive model of the task context situating multiple text task completion (i.e., the context model). In this study, we are, in part, interested in determining which contextual factors shaping students' multiple text use manifest in the justifications for self-evaluations of response quality that students provide. For instance, the number of texts that students have available to them during task completion may be thought of as both a support to be represented in students' context models, prior to task completion, and as a justification for students to consider when self-evaluating their task products in the final stage of multiple text use. Moreover, we examine how students' calibration of complex task performance (i.e., the correspondence between students' subjective and objective performance) is associated with their actual performance on a multiple text writing task. We also consider how the justifications that students cite for their self-evaluations of response quality are associated both with overall task performance and with calibration.

We were further interested in determining whether the criteria that students use to evaluate their writing performance and their calibration accuracy differed as a result of task assignment. Bråten and Strømsø (2009) found that students who were asked to construct an argument or to summarize information from across multiple texts were able to develop more deep-level and integrated understanding than students asked to produce a general overview. Thus, in this study, we were interested in determining the extent to which different task conditions may differ in the criteria that students reported considering when evaluating response quality and differences in calibration.

We have the following research questions and hypotheses:

1. How do students evaluate the quality of their written responses, composed based on multiple texts? What criteria do students use to justify their evaluations of writing quality?

   Based on prior work, we expected students to rate their written responses fairly highly, potentially demonstrating a degree of over-confidence (e.g., Dunlosky and Rawson 2012; Miller and Geraci 2011). We expected that students would consider a variety of factors when self-evaluating their written responses. Specifically, corresponding to prior work (Dinsmore and Parkinson 2013; Lin and Zabrucky 1998), students were expected to consider person-related, context-related, and task-related factors in rendering their self-evaluations.
2. To what extent do students' writing quality, calibration, confidence bias, and criteria for evaluating task performance differ by task?

   We expected students' writing quality, calibration, and confidence bias to significantly differ by task condition. This hypothesis is drawn from a recent study which found that students had different perceptions of common academic tasks, including arguments and research reports List et al. (under review). We expected these different conceptions to particularly manifest in differences in the criteria that students considered when self-evaluating task performance. In particular, based on their conceptions of these tasks, we expected students to justify their self-evaluations of the research report according to their inclusion of text-based evidence; conversely, we expected students to justify their self-evaluations of the argument task according to their assumption of a personal stance or their provision of a personal opinion.
3. What is the association between students' self-evaluations of writing quality and actual writing performance? What are the associations among students' calibration accuracy, confidence bias, and writing performance? To what extent does the association between calibration accuracy, confidence, bias, and writing performance differ for high-performing versus low-performing students?

   We expected students' self-evaluations of writing quality and actual writing performance to be positively correlated, but only to a moderate extent. At the same time, we expected students' calibration and confidence bias to be negatively associated with objective writing performance. That is, we expected students who were less accurate in their self-evaluations to also have a lower degree of writing performance, overall. Finally, we expected this negative association between calibration and confidence bias and objective performance to be particularly pronounced for lower-performing students, reflecting their relative over-confidence.
4. What is the association between the criteria that students use to justify their self-evaluations of performance and their actual writing quality, calibration accuracy, and confidence bias?

   We expected that writing quality would differ in association with the criteria that students cited to justify their self-evaluations. In particular, we expected that students who considered criteria related to text use (e.g., reporting that they evaluated library texts or cited sources) would demonstrate a higher level of task performance. Moreover, while we could not generate specific hypotheses regarding which justification categories would be associated with calibration accuracy and confidence bias, due to the dearth of prior work; we did expect the number and variety of criteria that students cited to be associated with these two outcomes.

Method

Participants

A total of 143 undergraduate students (age: $M = 20.20$, $SD = 1.67$) enrolled in a university in the Mid-Western United States participated in this study. The sample consisted of 102 female students (71.33%) and 39 male students (27.27%). Students reported White (67.83%, $n = 97$), African American/Black (18.88%, $n = 27$), Hispanic/Latino (4.90%, $n = 7$), Asian (2.10%, $n = 3$), and biracial/multiracial (1.40%, $n = 2$) ethnicity. Three students reported their ethnicity as other (2.10%). Students represented a variety of majors, primarily in the social and natural sciences. They also represented various class standings: 24.48% ($n = 35$) of participants were freshmen, 33.57% ($n = 48$) were sophomores, 25.17% ($n = 36$) were juniors, and 15.38% ($n = 22$) were seniors. Two students did not report demographic information.

Procedures

The study included three phases. First, students were asked to complete a variety of individual difference measures, including an assessment of prior knowledge. Second, students completed a multiple text task, wherein they were asked to research a complex and controversial topic (i.e., threats of and solutions to overpopulation) using a library of six digital texts and to compose a written response based on the texts provided. Specifically, students could choose how many texts to read and could revisit as many texts as needed, prior to indicating that their multiple text use was complete and proceeding to writing composition.

Students were randomly assigned to one of two task conditions, differing in the type of written response assigned. Specifically, students were either asked to compose an argument or a research report on the threats of and solution to overpopulation. Third, after composing their written responses, students were asked to complete several post-task measures. In particular, students were asked to self-evaluate their written responses with letter grades as well as to compose written justifications for their self-evaluations. Students could provide as many justifications as they wanted for their self-assigned letter grades.

This study is part of a larger project examining students' interactions with multiple texts (List et al. 2019). However, students' self-evaluations of writing performance and calibration accuracy are uniquely explored in this paper.

Measures

Prior knowledge Prior knowledge was assessed using a term identification measure. Specifically, students were asked to define seven terms (e.g., population bomb, high-yield crops) related to the topic of the task (i.e., overpopulation). Students' definitions were scored dichotomously as correct (1) or incorrect (0). Students' total prior knowledge scores ranged from zero to seven. Cohen's kappa inter-rater reliability, based on 29 (20.28%) student responses, was .76, with 88.18% exact agreement. Students' prior knowledge scores indicated that this was a low knowledge sample. See Appendix 1 for all prior knowledge items.

Multiple text task The multiple text task required students to first research overpopulation using a library of six digital texts and then to compose a written response. Prior to text use, students were randomly assigned to one of two task conditions and asked to use the texts either to compose an argument or a research report on the threats of and solutions to overpopulation. The task was as follow: *Write an argument/a research report for policy makers about the threats of overpopulation and how these may be the most effectively addressed.* After receiving this task assignment, students were presented with a library of six digital texts related to overpopulation. No time limit was imposed for this task.

Texts Students were provided with six digital texts relevant to the topic of the task (i.e., overpopulation). Texts were selected from the Room for Debate segment of the New York Times (https://www.nytimes.com/roomfordebate/2015/06/08/is-overpopulation-a-legitimate-threat-to-humanity-and-the-planet and https://www.nytimes.com/roomfordebate/2011/05/04/can-the-planet-support-10-billion-people). These were then adapted and modified for inclusion in this study. The six texts provided partially complementary and partially conflicting information. For instance, while one text argued that women should have easy access to contraception as a way of reducing birth rates, another text argued that birth rates were not the issue as fertility rates were falling and, rather, that overconsumption was the primary threat associated with overpopulation. All of the texts were attributed to trustworthy authors and reputable publishers. Texts ranged in length from 222 to 274 words and had Flesch-Kincaid grade levels from 10.5 to 17.4, Flesch reading ease from 23.8 to 44.8. While using the digital library, students were able to access as many of the six texts as they wanted to as well as to revisit texts, as needed. The summaries of each text are presented in Table 1.

Texts were presented in a random order, via a digital library. Each text was presented only by title. Students were free to access none, some, or all of the texts available in the digital library and could elect to revisit texts, as needed. Appendix 2 presents a screenshot of the digital library used to introduce the six texts to students.

Response coding Students' written responses, composed based on multiple texts, were scored using a five-point rubric, ranging from zero to four, with half points. The scoring rubric was established based on three indices, reflecting the number of arguments (i.e., claims and evidence) included in students' written responses and their degree of elaboration and integration. A score of zero corresponded to no relevant arguments provided. A score of one indicated students' provision of a single argument (i.e., claim with evidence) regarding a threat of or a solution to overpopulation. A score of two corresponded to students providing multiple arguments, or several claims with evidence, about overpopulation. A score of three was assigned when students were both providing multiple arguments in their responses and elaborating their arguments with the provision of added evidence, examples, or explanations, beyond a single justification provided. Finally, students received a score of four if their responses included multiple arguments about overpopulation, that were elaborated, and integrated, or in some way reflected the linking of information presented across texts. Half points were assigned when students attempted a response of some degree of quality, but were not entirely successful in its execution. For instance, students attempting to present arguments integrating information from multiple texts, but not doing so entirely successfully, received scores of 3.5, rather than 4. Cohen's kappa inter-rater reliability for response scores was 0.71, based on two raters. Scoring discrepancies were resolved through discussion. Table 2 includes the description of the scoring rubric, descriptives of students' open-ended response scores, and sample written responses.

Table 1 Information for the six texts

Text title	Author	Affiliation	Main point	
			Threat	Solution
Overconsumption is a Grave Threat to Humanity	Jamias Cascio	Institute for the Future	Resources are not enough to support the population due to overconsumption.	People need to reduce daily consumption and use resources more efficiently.
Building a Less Wasteful Economy	Chandran Nair	Global Institute of Tomorrow	Western lifestyles waste the world's resources.	We should build a new economic system to cope with the threats of overpopulation.
The Violent Side Effect of High Fertility Rates	Jack Goldstone	George Mason University	Overpopulation should largely be viewed as a regional issue, with regions with high birth rates susceptible to violence.	Governments must invest in human capital and development to mitigate the harms of overpopulation.
Technology and Population	Brad Allenby	Arizona State University	We do not know what level of population Earth can sustain as technological advancements constantly increase Earth's carrying capacity.	Future technologies may be a way of addressing the threats of overpopulation.
Empower Women for the Health of the Planet	Carmen Barroso	International Planned Parent-hood	Globally, women may not be able to control their reproductive health, increasing birth rates.	Providing women with family planning services, education, and contraception can reduce the threats of overpopulation.
More Efficient Food Production	Jason Clay	World Wildlife Fund	Overpopulation increases food consumption.	Producers and the food industry need to produce food in a more sustainable and responsible way.

Although students were asked to complete two different task assignment (i.e., write an argument and a research report), a common rubric was used to score students' responses. Using a common rubric allowed for maximum comparability in responses. Moreover, this rubric emphasized factors (e.g., elaboration, evidence integration) that we hoped would be featured in students' responses, regardless of whether they were composing arguments or research reports. Using a common rubric to score essays completed in response to varied tasks is common in the literature (Bråten and Strømsø 2009; Gil et al. 2010; Wiley and Voss 1999).

Self-evaluation task After students completed their written responses, they were asked to self-evaluate these. First, students were asked to give themselves a letter grade for their performance. Specifically, students were asked: *If you submitted your response as a class assignment, what grade do you think you would earn,* and asked to select a letter grade from A to F, including plus and minus grades, corresponding to the U.S. grading system. Then, students were asked to justify their self-evaluations or to explain why they thought that they would earn their designated grade. In the U. S., there are two types of grades: numerical grades and letter grades. Oftentimes, students receive grades on a scale from zero to 100. These numerical grades are then categorized into letter grades to represent levels of performance. These general levels of performance, demarcated by letter grades A, B, C, D, and F are further

Table 2 Scoring rubric for written responses

Score	Category	Description	N(%)	Example
0	Speculative	Single claim without evidence	1 (0.70%)	“**Overpopulation is a really important thing in our society (CLAIM)**, and we need to use our methods to solve it whenever you want to deal with it.”
0.5	Marginal	Multiple claims without evidence	0	
1	Basic	Single claim with evidence	3 (2.10%)	**In order to effectively manage /sic/ overpopulation you have to start by reducing overconsumption. (CLAIM)** One source said that the exponential population growth is a problem its overconsumption. By 2100 we would need three planets to sustain our style of living. **(EVIDENCE)** If we change the way we consume, we can change the way overpopulation is looked at.
1.5	Beyond basic	Single claim with evidence, or multiple claims without evidence	6 (4.20%)	Overpopulation is a topic that most people are not very well educated on, including me. Many factors can be affected by overpopulation including poverty, resource replenishment, the economy, and much more. **The source of overpopulation comes from pregnancies.(CLAIM)** Today more and more women are subjected to go to school and further their education than in past years. **(EVIDENCE)** In more developed countries, we have resources and education to be aware becoming pregnant. In other countries, it might not be a concern to raise a child in the best surroundings. **(EVIDENCE)** Overpopulation is a threat and in my opinion the only way to decrease the population would be to start with better education of women.
2	Sequential	Multiple claims with evidence	3 (2.10%)	**One of the most important reasons is efficient food supplement (CLAIM 1).** According to Allenby’s article, experts have predicted the overpopulation phenomenon in 1970, but the result shows it didn’t happen. The reason why it didn’t happen is because they have an unanticipated agricultural evolution that produces more type of high-yielding crops. **(EVIDENCE)** **The second reason why overpopulation happened is unpredictable technology (CLAIM 2)**. The development speed of technology is faster than our imagination. People never know what the future world will be. **(EVIDENCE)** People sometimes give their hope to technology instead of solve the current problems. **Third, there is another important reason causes the overpopulation is that the effect of the economy saving (CLAIM 3).** According to Building a Less Wasteful Economy written by Nair.C from the Global Institute for tomorrow, the core of our current economic model is to build a less wasteful economy. **(EVIDENCE)**
2.5		Multiple claims with evidence, somewhat elaborated	11 (7.70%)	**“The threats of overpopulation alone are numbered, but combined with the current rate of consumption, the threats increase dramatically. (CLAIM 1)** By 2100, we’ll need as many as

Table 2 (continued)

Score	Category	Description	N(%)	Example
	Sequential Partially Expanded			3 Earths to support the current population growth rate and consumption. **(EVIDENCE)** In areas where there is overpopulation now, there is a significantly higher level of violence, particularly in Sub-Saharan Africa and the Middle East. **(EVIDENCE)** How much we are consuming, especially in a western lifestyle, shows that our economy is geared towards eating our way through most of the Earth's natural habitat in 40 years. **(EVIDENCE)** On this path, we will be facing a world-wide social collapse within the next century. **The solutions, however, are even more numerous, but will all need to be employed to avert a social collapse on a global scale. (CLAIM 2)** More efficient and sustainable production of food will greatly relieve the stress on our resources, since we've been operating as though we have a "free ride on planetary resources" **(EVIDENCE)**. Another solution would be shifting the current economy to boost our environment and our resources, rather than expecting our current resources to support a growing economy **(EVIDENCE)**. It would need to be an economy that places more value on the welfare of everyone, rather than meeting the expectations of the individual. **(ELAB)** Another simple solution is access to family planning, which on it's own will provide 29% of the needed carbon emission reductions. Access to family planning will decrease unwanted pregnancy by 70%, while creating a generation of daughter who will likewise have fewer children. **(EVIDENCE)** Furthermore, governments in the Middle East and Sub-Saharan Africa will need to support their countries' infrastructure, technological advances, and skill building to help ensure long term stability and prosperity for all. **(EVIDENCE)**"
3	Sequential Expanded	Multiple claims with evidence, elaboration all expanded upon	48 (33.60%)	"I feel the threat of overpopulation is a great one, for a myriad of reasons**. Not only do we use a massive portion of the Earth for food, but we also give so much of said food away to frivolous things. (CLAIM 1)** Jason Clay, author of 'More Efficient Food Production,' stated that we use one third of our planets *[sic]* surface for food production. **(EVIDENCE)** That is a ton of space. Taking into consideration the amount of empty space that has been filled by subdivisions and new houses I've observed in the last decade it's plain to see we're running out of space. **(ELAB)** In fact, Clay also said that in order to sustain our activities in the year 2100 we would require two extra Earths. **(EVIDENCE)** Guess we better get looking, huh? Walking hand in hand with the issue of elbow room is the way we use this food we spend so much square footage on. **(ELAB)** In the article, "Overconsumption is a Grave Threat to Society," it's stated that one third of our produced grain is fed to our animals, and that one sixth of our grain is used for things like biofuel. **(EVIDENCE)** That only leaves humanity with half of it's grain to

Table 2 (continued)

Score	Category	Description	N(%)	Example
				eat. We spend a third of planets surface so we can only eat half of what we grow? That's absolutely ludicrous. That's the robber baron coming to your farm and taking half of your crop. There needs to be a change. Something has to give. Even if we cut down on the use of grain as an industry tool that still leaves us with more food, and with the growing issue of hunger any grain is good grain. **(ELAB)** **Perhaps our technology can aid us though, as it has in the past. (CLAIM 2)** In 1970 it was predicted that society would collapse due to overpopulation. However, due to advances in agriculture with high yield crops, the world's food production doubled, according to 'Technology and Population." **(EVIDENCE)** From 100% to 200%. That's such an incredible jump, it makes my heart sing. Perhaps if we could make a similar advancement in agricultural technology today, we could save ourselves from the looming doom of malnourishment and death. **(ELAB)** **Lastly, I know I overeat. I'm sure most people do. It's almost impossible these days to teach a child to eat until they're satisfied, not until they're full. (CLAIM 3)** Sometimes the simplest solution is right there, always watching. I think if we learned to eat until satisfied we can make a minimal dent in the issue of crops and their output to the people. However, if paired with the above techniques to avoid the impending crop debacle, we can help to sustain our society. **(ELAB)**"
3.5	Integrated	Multiple claims with evidence, majority expanded on; integration attempted	27 (18.90%)	The threats of overpopulation are mainly in three aspects. **First is overconsumption. (CLAIM 1)** The directly result of overpol pulation is that people will consume more foods and other Earth resources. **(EVIDENCE)** The possible solution to solve this problem is to change resource use and people current behaviors. **(ELAB) The second aspect is that overpopulation will leads to food scarcity which means people need to create new sustainability standards for food production. (CLAIM 2)** *The possible solution is to develop more high-yield crops and create business-to-business demand for sustainably produced food.* **(INTE) The last threat of overpopulation is the reconstruction of economy. (CLAIM 3)** Currently, the economy style is so-called free ride economy that people don't need to take the responsibilities of what they have done. **(EVIDENCE)** To overcome the possible reconstruction, the new economy system should be aware of that natural resources can have huge influence on economical growth and be subservient to maintain the natural resources. **(ELAB) More importantly, no matter people believe they can avoid overpopulation or have huge influence by overpopulation, development of technology is critical to both of the situation. (CLAIM 4)** By developing and

Table 2 (continued)

Score	Category	Description	N(%)	Example
				applying new technologies, people will have chances to avoid social collapse like the baby boom in 1970s and to solve the overconsumption problems efficiently. **(EVIDENCE)**
4	Fully Integrated	Multiple claims with evidence, majority expanded upon; integration fully explained	42 (29.40%)	*Based on the provided articles, there is somewhat of a disagreement among specialists whether or not overpopulation is actually a problem for our world today or not.* **(INTE)** *Of the six sources listed about half stated it is not that big of an issue and the other half argued that overpopulation can be or is an issue.* **(INTE) According to Jamias Cascio of New York Times, overpopulation is nothing for our world to fret over. (CLAIM 1)** Cascio mentions within his article that women now a days *[sic]* have 2.5 children, which is half as many as their grandmother. With this, the fertility rate continues to fall within today's world. **(EVIDENCE)** Cascio moves to state that what we should be worried about is a consumption bomb, meaning our population is depleting water, soil. and other life-supporting resources. **(EVIDENCE)** So what needs to change is how our population is consuming these resources, rather than the actual amount of population itself. **(ELAB)** *Another provided source stated quite the opposite of Cascio.* **(INTE) Jason Clay of Foreign Affairs stated that the United Nations projection shows troubling trend of increased population and rising consumption on a planet with finite resources. (CLAIM 2)** If this trend continues, by 2100 we will need three planet earths to support human activities, and 40 years from now we will have "eaten" nearly all of our natural habitat on planet Earth. **(EVIDENCE)** Clay stated that to address this threat of overpopulation we need to convince global companies to maximize efficiencies in their supply chain producers. **(ELAB)** *Overall, after reading all six articles I come to the conclusion that overpopulation is a problem as well as overconsumption.* **(INTE)** Whether it is our current population consuming too many resources or the fact that we will have to *[sic]* large of a population to support with our resources, something needs to change. **(ELAB)**

ELAB = Elaboration; INTE = Integration; claims are bold; evidence is underlined; and integrative statements are italicized qualified with pluses and minuses (e.g., A+/A-). Although the exact interpretation of letter grades varies across institutions, grades of C- or below are usually considered to demonstrate minimal achievement or failing, at least at the undergraduate level.

Coding Students' justifications for self-evaluation were coded using a four-step process. Based on the literature, we initially considered students' justifications for evaluations of response quality to be based on aspects of text, task, or individual characteristics. However, students seemed to consider a much broader span of features when evaluating their written responses than was reflected in these general categories. To capture this variability in students' responses a bottom-up coding scheme was adopted. In the first phase of coding, two researchers independently read through all of the justifications that students offered for self-evaluations of performance. In this phase, researchers independently segment participants' responses into idea units, typically corresponding to a single phrase containing a verb or a single justification for response quality, and identified an initial set of justification categories, based on the specific terms that students used to justify their evaluations of response quality. Within this phase, researchers also completed an initial independent coding of participants' justifications, by idea unit. In Phase 2, researchers compared the categories that they had independently derived as well as their initial coding of students' responses to ensure a sufficient degree of inter-rater agreement. Based on discussions in Phase 2, researchers reconciled the coding categories they derived and independently re-coded data as needed. In Phase 3, researchers compared their coding of students' responses and discussed any disagreements based on a common set of coding categories. Moreover, researchers looked at their coding scheme to determine which categories potentially needed to be combined, collapsed, or removed from analysis. Finally, in Phase 4, researchers reviewed responses one last time and ensured that these were consistently coded according to updated category definitions. The first and second author scored all students' justifications. Cohen's kappa inter-rater reliability was .74 for 561 justifications coded (exact agreement: 86.81%).

A final set of 12 coding categories, reflecting the variety of criteria that students used to self-evaluate their written responses, were identified. Three coding categories reflected students' self-evaluations of task performance based on composition-related factors. These were self-evaluations based on (1) writing mechanics (e.g., writing format, length, grammar), (2) writing structure (e.g., including an introduction or conclusion), and (3) the degree of elaboration included in students' writing responses. Four coding categories reflected students' self-evaluations considering the degree of source use evidenced in their responses. The (4) superficial source use category included justifications based on students' reports of drawing on texts at a fairly low-level (e.g., simply mentioning the number of texts used), while the (5) deep source use category reflect students' reports of integrating, evaluating, or otherwise deeply engaging with texts. The (6) evidence use category reflected students' reports of using data or factual information in response composition as a justification for writing performance (e.g., *I think that I explained my argument in a factual manner*). Some students further justified response quality based on their provision of (7) citations.

One coding category reflected students' self-evaluations rendered based on (8) contextual factors, related to their completing the task within a laboratory setting (e.g., perceived time limit, needs for more information). Further, students' justifications of response quality based on personal characteristics (e.g., prior knowledge or personal opinion) were coded as (9) personal attributes. We also coded self-evaluation criteria as (10) general statement when students gave themselves an overall positive or negative evaluation of response quality, without further specific criteria provided (e.g., *I think I did an above average job*). Moreover, we coded self-evaluation criteria as (11) task positive when students reported that they completed the task in a way that satisfied task demands (e.g., *I feel I answered the question completely, addressing all parts of the question*). Any responses not able to be otherwise coded were placed into the (12) “other” category. Responses in the “other” category presented unique justifications, not reported by other students, that were often unclear or irrelevant to the task. Table 3 offers descriptive inforomation and a sample response for each coding category.

Results

Research question 1: Students’ self-evaluations of response quality

Our first research question examined the criteria that students cited in self-evaluating response quality. Across the two conditions, students rated their responses by selecting a letter grade from A to F, including plus and minus grades. Overall, the majority of students rated their writing quality as a B (27.27%, $n = 39$) or a B+ (30.07%, $n = 43$). Only four students (2.80%) rated their writing quality as reflecting an A. A histogram of students’ self-assigned grades is presented in Fig. 1.

In justifying their written response quality, students offered a total of 561 justifications, with an average of 3.92 ($SD = 2.39$) justifications provided per student. Overall, students most frequently reported superficial source use, contextual factors, writing mechanics and personal attributes as factors they considered in evaluating the quality of their written responses. Specifically, 41.26% ($n = 59$) of students cited source use that was superficial in nature as one of their evaluative criteria. This included students evaluating their responses based on the number of sources they used, their use of direct quotations, or other superficial indicators.

Table 3 Sample self-evaluation justifications

Category	Example	N(%)
Writing	“…I may have made a few grammatical errors…”	45 (31.47%)
Structure	“... having an introduction, thesis, body(s), and conlsusion...”	26 (18.18%)
Superficial source use	“...I also used three sources that I felt were relevant...”	59 (41.26%)
Deep level source use	“…relate the articles to each other…”	34 (23.78%)
Task context	“…I knew I didn’t have enough time to look through all the resources…”	56 (39.16%)
Personal attributes	“I’m not a writer, and I don’t write a lot in my free time or in class…”	39 (27.27%)
Elaboration	“I believe I was thorough in my explanation…”	28 (19.58%)
Task positive	“…I followed directions…”	19 (13.29%)
Citation	“...I was able to cite the examples back to the articles...”	20 (13.99%)
[a]General statement	“I think that I did a good job showing issues and how to overcome those issues…”	30 (20.98%)
Evidence	“...I feel that I used good data to back up the points I made in the response...”	18 (12.59%)
[b]Other	“…I want to do more research but unfortunately I never find the time...”	20 (13.99%)

The [a] General category reflected students’ global judgment of task performance, without any further elaboration. The [b] Other category reflected students elaborating their criteria for self-evaluation; however, these criteria were unique and often irrelevant to the task and did not share semantic or conceptual overlap with the self-evaluation criteria identified by other students

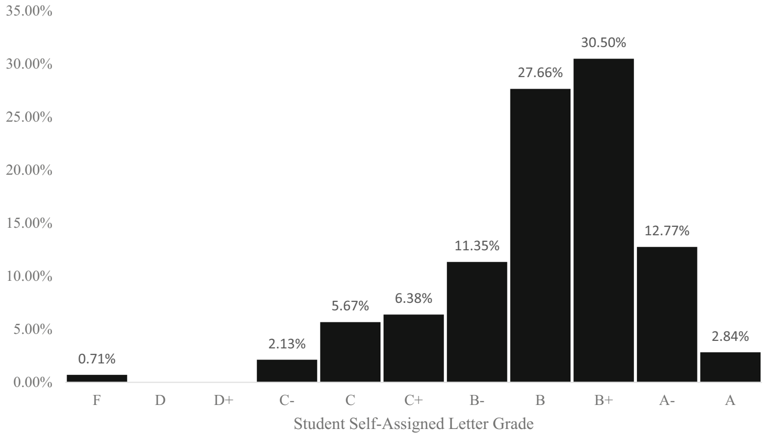


Fig. 1 Student self-assigned letter grade

Further, 39.16% ($n = 56$) of students frequently cited *context*-related factors as one of the evaluative criteria they used. This category included students' consideration of factors in the task context that might have affected their performance (e.g., perceived time limits, limited library texts). Context-based self-evaluations of performance included statements like: *I feel like the source is not enough for me.* Moreover, 31.47% ($n = 45$) of students evaluated their written products based on considerations of *writing mechanics*. These were students' response evaluations rendered based on their use of accurate spelling or grammar or according to response length. For instance, one student justified the grade that they assigned to their written response as: *I rated this a B because my grammar was off.* Notably, 23.78% ($n = 34$) of students' self-evaluation justifications were characterized as reflecting a concern for deep-level source use. These were justifications for self-assigned letter grades based on students having processed texts deeply or integrated information across texts. For example, one student's self-evaluation justification placed into the deep-level source use category was: *I did an ample job of connecting the different sources into one argument.* All justification categories and sample responses are summarized in Table 3.

Research question 2: Writing quality, calibration, confidence bias, and self-evaluation criteria by task

The second research question first examined the extent to which students' writing quality, calibration, and confidence bias differed by task condition (i.e., writing an argument or a research report). We performed three independent sample t-tests to examine the differences between the two task conditions in overall writing performance, calibration accuracy, and confidence bias. Results showed that students' writing quality [$t(139) = -2.29$, $p < .05$] and confidence bias [$t(137) = 2.18$, $p < .05$] significantly differed by task. In particular, students who were asked to write a research report received higher objective scores ($M = 3.36$, $SD = 0.66$) than students who were asked to write an argument ($M = 3.06$, $SD = 0.84$) based on multiple texts, Cohen's $d = 0.40$. Moreover, students were under-confident in their writing performance when responding to the research report task ($M = -0.21$, $SD = 1.02$) as compared to students who were overconfident about their performance when asked to write an argument ($M = 0.26$, $SD = 1.47$), Cohen's $d = -0.37$. Students' calibration accuracy was not significantly different across task conditions ($p = .09$). The results of these t-tests are presented in Table 4.

Table 4 Difference in writing quality and confidence bias by task condition

	Argument		Research report		*t*	*p*	Cohen's *d*
	M	*SD*	*M*	*SD*			
Writing quality	3.06	0.84	3.36	0.66	−2.29	.02	0.40
Confidence bias	0.26	1.47	−0.21	1.02	2.18	.03	−0.37

No significant difference in calibration across conditions were identified

We also examined the extent to which the self-evaluation criteria that students cited differed by task condition. Since relatively few students reported multiple justifications within each category, justification categories were dichotomized in analyses. An independent sample t-test determined that students' self-assigned letter grades did not differ across task conditions ($p = 0.47$). Further, a chi-squared test of association was used to examine whether citing any of the 12 criteria in justifying response quality was associated with task condition. However, no significant associations were found (*p*s > .17). Nevertheless, the prevalence of various justification criteria reported across task conditions are summarized in Table 5.

Research question 3: Self-evaluations of writing quality, objective writing quality, calibration, and confidence bias

For the third research question, we were interested in examining the association between students' self-evaluations of response quality (i.e., self-assigned letter grades) and their objective performance on the multiple text task. Students' self-assigned letter grades were quantified (e.g., A = 11, and F = 1). Both quantified letter grades and the holistic scores assigned to students' written responses were standardized prior to analysis. Three sets of analyses were performed. First, we examined the association between students' objective response quality and subjective letter grades. Additionally, we examined the association between students' objective performance and their *calibration*, computed as the absolute difference between objective and subjective task performance. Finally, we looked at the association between objective response

Table 5 Prevalence of justifications across the two task conditions

	Argument		Research report	
Justifications	n	%	n	%
Writing	19	26.4%	26	36.6%
	10	13.9%	16	22.5%
Superficial source use	33	45.8%	26	36.6%
Deep source use	16	22.2%	18	25.4%
Context	26	36.1%	30	42.3%
Person	16	22.2%	23	32.4%
Elaboration	13	18.1%	15	21.1%
Task	9	12.5%	10	14.1%
Citation	11	15.3%	9	12.7%
General statement	12	16.7%	18	25.4%
Evidence	7	9.7%	11	15.5%
Other	8	11.1%	12	16.9%

A chi-squared test of association found no significant differences across task conditions in the justification criteria cited quality and students' *confidence bias*, corresponding to the difference between students' self-assigned letter grades and objective task performance. See Table 6 for definitions of objective performance, subjective performance, calibration, and confidence bias.

In this study we considered it important to distinguish calibration, or absolute accuracy, from confidence bias. While calibration refers to the absolute difference between students' self-assigned grades (i.e., subjective) and actual (i.e., objective) performance, confidence bias reflects the extent to which students are over- or under-confident in the written responses that they compose (Schraw and Nietfeld 1998; Yates 1990). Examining confidence bias, in addition to calibration, was considered key as students' relative over- or under-confidence in task performance was expected to spur or inhibit strategy use aimed at improving response quality (Griffin et al. 2013; Ramdass and Zimmerman 2008; Winne and Jamieson-Noel 2002). For instance, students who were under-confident in their written responses may have worked to improve these (e.g., including additional evidence or elaboration), while students who were overconfident may have devoted less time and effort to task completion.

We first examined the correlation between students' quantified letter grades and holistic scores, corresponding to objective response quality. However, these were only marginally associated with one another [$r(1) = .16$, $p = .06$].

Calibration We also examined the association between students' objective performance and calibration (i.e., absolute accuracy). We found these to be negatively correlated [$r(1) = -.37$, $p < .01$]. This negative association can be interpreted as students with increased objective performance being more accurate in their subjective self-evaluations, corresponding to lower calibration scores. See Table 7 for correlations among key variables.

We were further interested in examining the extent to which the calibration of high-performing students differed from that of low-performing students. To examine differences in calibration at different levels of achievement, students were split into two groups based on median objective performance (i.e., resulting in the creation of a high-performing and low-performing student sample). We separately examined the association between objective task performance and calibration accuracy for low-performing and high-performing students. While no significant association between objective performance and calibration accuracy was found for high-performing students ($p = .94$), a significant negative association was found between objective performance and calibration for the low-performing students, [$r(1) = -.68$, $p < .001$]. Moreover, as may be expected, a Fisher's z-test found the correlation coefficients between objective performance and calibration accuracy to be significantly different for low-

Table 6 Definitions of key variables

Variable	Definition
Subjective performance scores	Quantified students' self-assigned letter grades (e.g., A = 11, and F = 1).
Objective performance scores	Actual scores students earned on their written responses. Responses were rated based on an analytic scoring rubric (see Table 2).
Calibration or absolute accuracy	Absolute value of the difference between subjective performance scores and objective performance scores.
Confidence bias	Subjective performance scores (quantified self-assigned letter grades) minus objective performance scores.

Quantified self-assigned letter grades and objective scores were both standardized prior to analysis

Table 7 Correlations among key variables

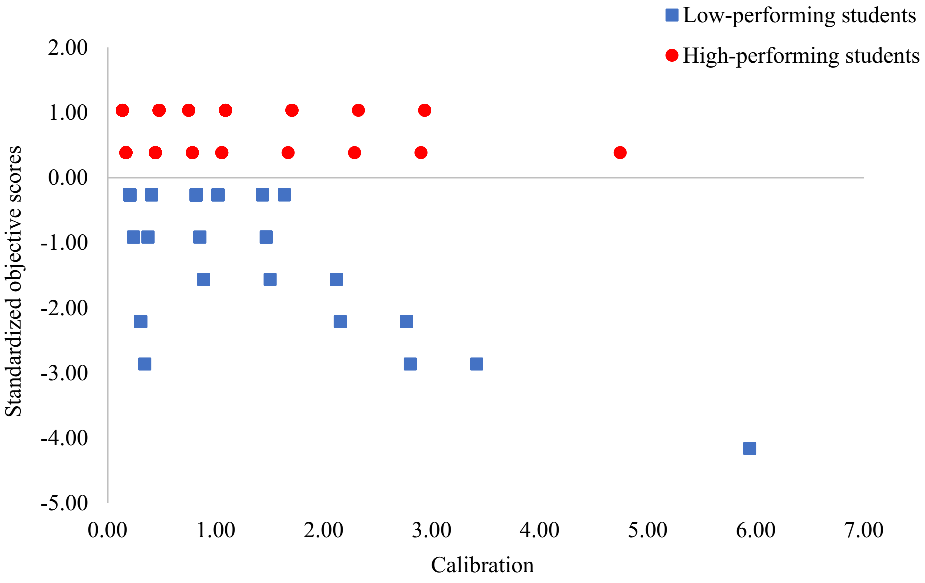
	1. Objective scores	2. Calibration accuracy	3. Confidence bias
1.Objective scores	–		
2.Calibration accuracy	$-.37^{**}$	–	
3.Confidence bias	$-.65^{**}$	.07	–
4. Number of justifications	$.27^{**}$	$-.18^{*}$	−.15

Note. $* \ p < .05$, $** \ p < .01$.

performing students vis-à-vis high-performing students (Fisher's $z = 4.84$, $p < .001$). A scatter plot of these associations is presented in Fig. 2.

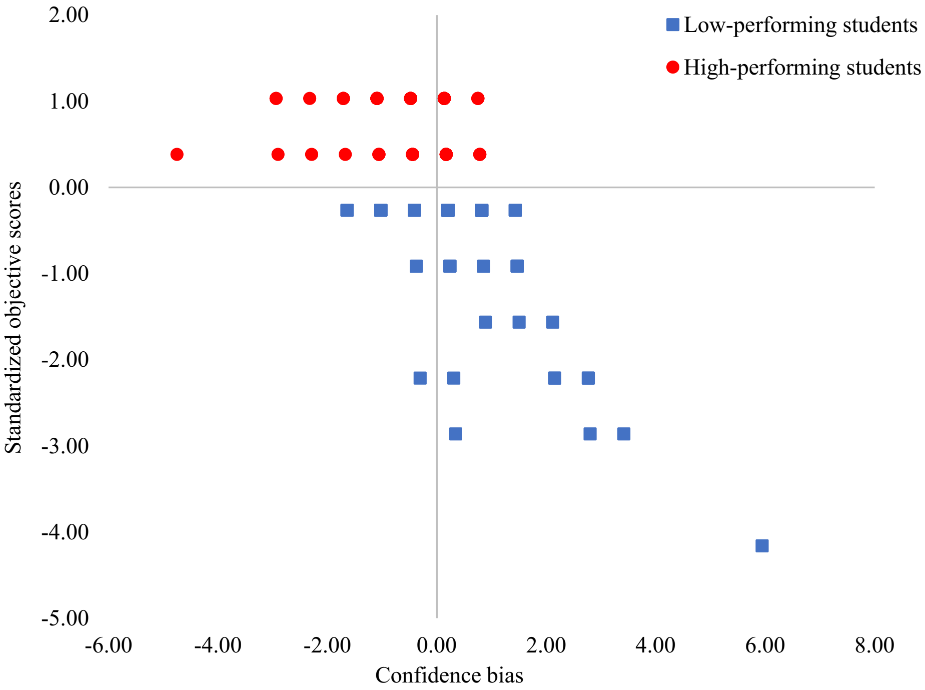
Confidence bias We were further interested in examining the association between students' objective performance and confidence bias (i.e., the difference between students' self-assigned letter grades and objective scores). Again, a significant negative correlation was found [$r(1) = -.65$, $p < .01$]. This shows that students who were under-confident, reflected in negative confidence bias scores, tended to have better responses.

Again, we compared the relation between objective performance and confidence bias for low-performing vis-à-vis high-performing students, as determined via a median split. For high-performing students, there was no significant association between objective performance and confidence bias ($p = .67$); however, there was a significant negative association between objective performance and confidence bias for low-performing students [$r(1) = -.66$, $p < .01$]. A Fisher's z-test found the correlation coefficients for these two performance groups to be significantly different from one another (Fisher's $z = 4.28$, $p < .001$). A scatter plot for the correlation between objective performance and confidence bias for these two groups is presented in Fig. 3.



Note. Quantified self-assigned grades and objective scores were both standardized prior to analyses.

Fig. 2 Correlations between objective scores and calibration by performance group



Note. Quantified self-assigned grades and objective scores were both standardized prior to analyses.

Fig. 3 Correlations between objective scores and confidence bias by performance group

Research question 4: Self-evaluation criteria and writing quality

Our fourth research question used a series of three multiple regression models to examine the extent to which the criteria that students cited in justifying response quality predicted actual performance, calibration, and confidence bias. For each regression model, prior knowledge and task condition (i.e., writing an argument or a research report) were controlled for in Step 1. In Step 2, a sub-set of the criteria that students could have cited in justifying response quality were entered as predictors. Specifically, five criteria (i.e., structure, superficial source use, deep-level source use, citation use,

Table 8 Objective performance, calibration, confidence bias by five justification criteria

	Objective scores		Calibration scores		Confidence bias	
Criteria	Yes *M (SD)*	No *M (SD)*	Yes *M (SD)*	No *M (SD)*	Yes *M (SD)*	No *M (SD)*
Structure	3.46 (0.69)	3.15 (0.78)	0.67 (0.60)	0.98 (0.94)	−0.11 (0.90)	0.07 (1.36)
Superficial source use	3.29 (0.66)	3.15 (0.84)	0.87 (0.88)	0.96 (0.91)	0.05 (1.25)	0.02 (1.33)
Deep-level source use	3.46 (0.57)	3.13 (0.81)	0.74 (0.74)	0.98 (0.94)	−0.11 (1.05)	0.08 (1.36)
Citation	3.40 (0.50)	3.17 (0.80)	0.58 (0.39)	0.98 (0.95)	0.15 (0.69)	0.01 (1.36)
Evidence	3.21 (0.53)	3.21 (0.80)	0.97 (0.93)	0.59 (0.51)	0.41 (0.68)	−0.02 (1.35)

Three sets of twelve t-tests were run to examine the extent to which objective scores, calibration, and confidence bias respectively differed by self-evaluation criteria. None of comparisons were significant, $p > .03$ (objective scores), $p > .07$ (calibration), $p > .20$ (confidence bias), αswere adjusted and evidence evaluation) were used as predictors in the model. These justification criteria were selected for inclusion because they reflected students' understanding of the need to use text-based evidence in composing a quality written response. Moreover, these justification criteria aligned with the scoring rubric used in this study, and more broadly, with the criteria commonly used to evaluate students' written responses in prior work (Anmarkrud et al. 2014; Reznitskaya et al. 2009). Table 8 compares students' objective performance, calibration, and confidence bias according to whether or not they cited particular criteria in justifying response quality.

Holistic score A hierarchical multiple regression model was run to predict students' actual writing performance based on the criteria they cited to justify response quality. Prior knowledge and task condition were entered as control variables in Step 1. Students' citation of five justification criteria (i.e., structure, superficial source use, deep-level source use, citation use, and evidence evaluation) were entered in Step 2. The overall model was significant, $F(7, 133) = 3.32$, $p < .05$, $R^2_{adj} = .10$. Nonetheless, none of the criteria for self-evaluation were individually significant in the model, $ps > .08$. A model summary is presented in Table 9.

Calibration A second hierarchical multiple regression model was run to examine students' calibration accuracy based on the five criteria they cited to justify response quality. Prior knowledge and task condition were entered at Step 1 as control variables, while five variables, corresponding to justifications of response quality based on response structure, superficial source use, deep-level source use, citation, and evidence evaluation, were entered at Step 2. The overall model was not significant, $p = .09$.

Confidence bias A final multiple regression model was run predicting students' confidence bias based on the criteria they cited when self-evaluating response quality. Prior knowledge and task condition were entered at Step 1 as control variables, while the five target criteria used to justify response quality were entered at Step 2. The overall model was significant, $F(7,131) = 2.83$, $p < .05$, $R^2_{adj} = .09$. However, none of the five criteria cited for self-evaluation were uniquely significant predictors in the model, $ps > .23$. A model summary is presented in Table 10.

Follow-up analyses The focal analyses conducted for Research Question 4 used theory-based reasons to select justification criteria to use as predictor variables. However, we also wanted to examine models predicting performance, calibration, and confidence bias using a more data-driven approach. Therefore, we ran a series of complementary hierarchical multiple regression

Table 9 Hierarchical regression model summary predicting objective performance

Step and predictor variable	*B*	*SE B*	β	*p*
Step 1:				
Task condition	.36	.16	.18	.03
Prior knowledge	.14	.05	.22	.01
Step 2:				
Structure	.28	.21	.11	.19
Superficial source use	.22	.17	.11	.18
Deep-level source use	.34	.19	.14	.08
Citation	.29	.24	.10	.23
Evidence	-.04	.26	-.01	.87

Note. $F(7,133) = 3.32$, $R^2 = .15$, adjusted $R^2 = .10$, $p < .01$.

Table 10 Hierarchical regression model summary predicting confidence bias

Step and predictor variable	*B*	*SE B*	β	*p*
Step 1:				
Task condition	−.49	.21	−.19	.02
Prior knowledge	−.24	.07	−.29	.00
Step 2:				
Structure	.01	.28	.00	.98
Superficial source use	−.05	.22	−.02	.80
Deep−level source use	−.08	.25	−.03	.74
Citation	.01	.32	.04	.97
Evidence	.41	.34	.11	.23

Note. $F(7,131) = 2.83$, $\mathrm{R}^2 = .13$, adjusted $R^2 = .09$, $p < .01$.

models predicting each target outcome variable using the four justification criteria that students most frequently cited (i.e., superficial source use, context, writing, and personal attributes). Each of these justification criteria were reported by more than 25% of the sample. Moreover, these four most popular criteria reflected the context- and person-related factors that Dinsmore and Parkinson (2013) suggest are important to the calibration of reading comprehension.

We were interested in examining how these factors affected students' objective performance, calibration, and confidence bias. Prior knowledge and task condition were again controlled for at Step 1, while the four most commonly cited justification criteria were entered at Step 2. The model predicting actual response quality was significant, $F(6, 134) = 3.19$, $p < .05$, $R^2_{\mathrm{adj}} = .09$. However, superficial source use was the only uniquely significant predictor in the model, $\beta = .17$, $p < .01$. See Table 11 for a model summary.

The model predicting calibration accuracy based on prior knowledge and task assignment (Step 1) and the four criteria that students most commonly cited in justifying response quality (Step 2) was not significant ($p = .41$). However, the model predicting confidence bias was significant, $F(6,132) = 3.32$, $p < .05$, $R^2_{\mathrm{adj}} = .09$. However, none of the predictors reflecting criteria for self-evaluation were uniquely significant in the model. See Table 12 for a model summary.

Discussion

This study had three primary goals. First, we were interested in examining students' calibration and confidence bias when completing a complex task (i.e., composing a written response based on multiple texts). Second, we were interested in examining the criteria that students

Table 11 Follow−up hierarchical regression model summary predicting objective performance

Step and predictor variable	*B*	*SE B*	β	*p*
Step 1:				
Task condition	.39	.16	.20	.02
Prior knowledge	.16	.05	.24	.00
Step 2:				
Superficial source use	.35	.17	.17	.05
Context	.21	.17	.10	.23
Writing	.17	.18	.08	.35
Personal attributes	−.22	.19	−.10	.25

Note. $F(6,134) = 3.19$, $R^2 = .13$, adjusted $R^2 = .09$, $p < .01$.

Table 12 Follow−up hierarchical regression summary for confidence bias

Step and predictor variable	*B*	*SE B*	β	*p*
Step 1:				
Task condition	−.45	.21	−.18	.04
Prior knowledge	−.24	.07	−.29	.00
Step 2:				
Superficial source use	−.14	.22	−.06	.52
Context	−.28	.22	−.11	.22
Writing	−.17	.23	−.06	.48
Personal attributes	.05	.25	.02	.84

Note. $F(6,132) = 3.32$, $R^2 = .13$, adjusted $R^2 = .09$, $p < .01$.

used to justify their self-evaluations of written responses composed based on multiple texts. Finally, we investigated the role of justification criteria in students' objective response quality, calibration, and confidence bias.

This paper contributes to the literature on multiple texts use and calibration in at least four ways. First, it is one of the few studies to examine calibration and confidence bias within a multiple-text context where students are asked to produce written responses. While calibration has been considered to be a strong indicator of task performance, the prior literature on calibration has been limited by an over-focus on students' completion of fairly low-level, rather than high-level, academic tasks (e.g., multiple-choice questions, discrete short answer questions; Dinsmore and Parkinson 2013; Griffin et al. 2009; Nietfeld et al. 2006a). The present study builds on this prior work by examining students' calibration within the context of a common, high-level academic task (i.e., composing a written response based on multiple texts).

Second, this study examines not only students' calibration (i.e., absolute accuracy) but also their confidence bias when self-evaluating task performance. Moreover, it considers the multitude of criteria that students cite in justifying their self-evaluations of response quality. This is consistent with prior work which has found students to report idiosyncratic task criteria when judging their performance on complex tasks (List and Alexander 2015; Pieschl 2009; Winne and Hadwin 1998). Indeed, in this study, a variety of criteria, cited as justifications for subjective judgments of response quality, were found to be associated with both objective task performance and with students' confidence bias. This included justifications based on students' perceived depth of source use, citation use, evidence evaluation, and response structure.

In examining students' reported justification criteria, this study sought to explain the discrepancy between students' objective task performance and subjective judgments of calibration and confidence bias. Echoing prior work (Dunlosky and Rawson 2012; Schraw et al. 1995), we demonstrate that students are often inaccurate in self-evaluating their task performance and that this inaccuracy should be conceptualized not in the absolute, but rather as a relative confidence bias. In particular, students' relative over-confidence in the quality of their written responses seems to be associated with a greater inaccuracy in their predictions of performance and with lower objective performance, overall. In this paper, we suggest a reason for why such over-confidence may arise. Specifically, it may be due to students' inaccurate selection of criteria to use in justifying response quality. For instance, if students consider a quality response to include accurate spelling and grammar, they may judge their answers quite favorably, even when these fail to include evidence or sufficient elaboration. Such an interpretation is further supported by students placed into the high performing category, according

Table 13 Justification criteria by high- and low-performing students

Criteria	Low-performing students		High-performing students			
	n	%	n	%	$\chi^2(1)$	*p*
Writing	20	44	25	56	1.40	.24
Structure	9	35	17	65	3.74	.05
Superficial source use	29	50	30	50	0.27	.60
Deep level source use	12	35	22	65	4.84	.03
Task context	28	50	28	50	0.11	.74
Personal attributes	20	51	19	49	0.01	.95
Elaboration	17	61	11	39	1.12	.29
Task satisfaction	9	47	10	53	0.17	.68
Citation	8	40	12	60	1.29	.26
General statements	12	40	18	60	2.10	.15
Evidence use	10	56	8	44	0.12	.73
Other	7	35	13	65	2.61	.11

to a median split, being disproportionately more likely to cite deep-level source use as a criteria in justifying response quality, $\chi^2(1) = 4.84$, $p < .05$, Cramer's $V = .18$. See Table 13.

Finally, our findings reveal the discrepancy between high-performing students and low-performing students in calibration and confidence bias. This corresponds to the existing literature (e.g., Hacker et al. 2000; Snyder et al. 2011; Zabrucky et al. 2009). In our study, we specify that it was low performing students that accounted for the negative associations among objective performance, calibration, and confidence bias.

Research question 1: Students' self-evaluations of response quality

For our first research question, we asked students to self-evaluate their written responses, composed based on multiple-texts, by assigning themselves a letter grade and justifying their letter-grade assignment. Overall, students tended to rate their performance at a moderately high level. This is consistent with prior work which has found students to have generally high levels of confidence in their performance (e.g., Allwood et al. 2006; Dunning et al. 2003; Graham et al. 2005; Koriat 1993; Stone and Opel 2000).

Moreover, students cited a variety of criteria in justifying their self-evaluations of response quality. This suggests that students have a broad span of criteria that they may use in self-evaluating complex task performance. Indeed, the range of criteria that students cited as justifications for self-evaluation in this study reflected the task-, text-, and personally-related criteria previously identified in the literature as contributing to calibration (Dinsmore and Parkinson 2013; Lin and Zabrucky 1998). For instance, justifications based on students' perceptions that they satisfied task demands mapped onto Dinsmore and Parkinson's description of task characteristics as contributing to calibration accuracy. Moreover, the personal attributes category, including justifications based on students' skills as writers or prior knowledge, mapped onto Dinsmore and Parkinson's description of personal characteristics contributing to calibration (2013). However, in comparison to the variety of categories that we identified from students' justifications overall, on average, each student only considered a limited number of criteria ($M = 3.92$, $SD = 2.39$). This may indicate that students' deficits in calibrating their performance on a complex task, such as writing based on multiple texts, may stem from their consideration of a limited number of criteria, or incorrect criteria, when judging performance. At the same time, the number of justifications that students produced in this study dwarfs results from Dinsmore and Parkinson (2013) who found students to commonly consider only a single factor when justifying their self-evaluations of performance.

At the same time, many of the criteria that students cited in justifying response quality were of a superficial nature, reflecting a concern with response length or spelling accuracy, to the exclusion of evaluating responses based on their quality of argumentation and evidence provision. Indeed, even within self-evaluations that attended to the same criteria (i.e., text use) students were found to consider these criteria along a spectrum of sophistication. For instance, justifications placed into both the superficial source use and the deep-level source use categories reflected students' concerns with their use of information from the texts provided in composing their written responses. However, while students placed into the superficial source use category only described using several texts or summarizing information from texts, students' justifications in the deep-level source use category included reports of multiple text evaluation, comparison, and integration. While superficial and deep-level source use may seem to represent polar ends of a common scale, in students' reports they reflected distinct categories of evaluation that sometimes were simultaneously used. For example, one student justified their self-evaluation like: *I took extensive notes…and I put them in conversation with each other, instead of simply summarizing everything that I read.* In this student's response, they first justified their self-evaluation as dependent on their taking of "extensive notes" (superficial source use, reflecting information accumulation), then this student also justified their self-evaluation based on their deep-level source use (*I put them in conversation with each other*) during task completion. Thus, students may consider both superficial and deep-level source use for justifying their self-evaluations. Nevertheless, this difference in source use sophistication reflects the need to articulate for students not only the criteria to attend to in determining response quality (e.g., using texts) but also how these criteria should be interpreted (e.g., integrating or evaluating texts). The need to prompt students to engage in deep-level source use is further emphasized by 41.26% of students citing reasons related to superficial source use in justifying response quality, as compared to only 23.78% of students reporting reasons related to deep-level source use.

An interesting category to arise among justifications for response quality reflected students' contextual concerns. Justifications in this category included students' reports that their task performance was impacted by perceived time limits, the limited information that they had access to, and their completion of the task in a lab setting. Indeed, all of these factors likely contributed to students' task completion, in ways that may be consistent with other studies of multiple text use completed in laboratory settings. At the same time, more work is needed to systematically examine the contribution of each of these factors to response quality. Notably, all four of the contextual factors identified in the RESOLV model (Rouet et al. 2017) were cited by participants in our study, to varying extents. For instance, one student attributed their self-evaluation of performance to perceived time and information-related limitations: "*I could have spent more time, but I felt like I needed more sources to get the bigger picture of the issue*." This student's response reflects his or her identification of obstacles in the task context as potentially hindering task performance. As suggested by this response, the RESOLV model can be used not only to understand the context of students' multiple text use but also to examine how students' perceptions of contextual factors may contribute to self-evaluations of task performance.

Beyond contextual factors, a variety of individual difference factors may play a role in students' self-evaluations during task completion. Chief among these are students' prior domain or topic knowledge and prior experience completing similar tasks. The need to consider such individual difference factors is emphasized by Glenberg and Epstein (1987) who found the accuracy of self-assessments to differ across students more or less expert in various domains. More generally, students' personal experiences with a task have been found to be associated with judgments of performance (Winne and Perry 2000), such that students with previous task experience are more likely to perform better and to have more accurate self-evaluations. Domain or topic knowledge may likewise be expected to contribute to students better understanding of what a particular task may require, and therefore to better performance and calibration. At the same time, we expected students in this study, although limited in prior knowledge, to have quite a bit of task experience, as both argument assignments and research reports constitute common academic tasks. Task experience, in this case, did not seem to have a clear role in improving calibration accuracy. In part this may reflect students' difficulties transferring performance from classroom to lab settings. Alternately, students may need explicit cuing to activate their relevant prior task experience when self-evaluating response quality (Winne 2001).

Research question 2: Self-evaluation criteria by task

Our second research question examined the differences in students' writing quality, calibration, confidence bias, and self-evaluation criteria across task conditions (i.e., writing a research report or an argument). Results showed that students' writing quality and confidence bias significantly differed across task conditions, with students demonstrating higher writing quality and lower confidence bias scores (i.e., under-confidence) when writing a research report. These results can be explained in a number of ways. First, consistent with prior work on calibration, finding task performance to be negatively associated with confidence bias (Bol et al. 2005; Kruger and Dunning 1999), students may have considered the research report to be more difficult to compose. Their evaluations of this task may have resulted in their being comparatively under-confident in their report writing and therefore deploying strategies and investing effort during task completion that ultimately produced improved performance.

More importantly, differences in students' writing quality and confidence bias can also be explained by their differential conceptions of argument tasks vis-à-vis research reports. In a recent study examining students' conceptions of common academic tasks, students were found to differentiate between argument and research report tasks (Authors, under review). Students considered research reports to require the use of external information from texts and to be associated with gathering information, conducting research, and with the demands of writing. In contrast, argument tasks were found to be associated with students' adopting personal stances toward a topic and forming an opinion. Such different task conceptions may have resulted in students' differential conceptions of task complexity. In particular, as the research report was associated with more extensive multiple text use and writing, which students may have perceived as more demanding to complete, the research report may have engendered students' greater under-confidence. This under-confidence may, in turn, have contributed to students' improved writing quality. More generally, students' conceptions of research reports as requiring the explicit use of multiple texts may have explained their inclusion of a greater number of text-based evidence and citations in their written responses, ultimately demonstrating improved task performance. Such an explanation is consistent with Pieschl's (2009) description of the association between task complexity and judgment accuracy when students are asked to self-evaluate performance.

Research question 3: Self-evaluations of writing quality, objective writing quality, calibration, and confidence bias

The third research question examined the associations among students' objective task performance and subjective self-evaluations, as reflected in both calibration accuracy and confidence bias. First, we found that students' actual scores were only marginally associated with their self-assigned grades. Second, students' calibration and confidence bias were both negatively associated with students' actual performance. In particular, consistent with prior research, students' objective performance was associated with higher calibration accuracy (i.e., indicated by low calibration scores) and negative confidence bias, reflecting under-confidence (e.g., Bol and Hacker 2001; Labuhn et al. 2010).

The association between poorer subjective judgments of performance and lower objective performance can be explained in a number of ways. First, students' deficit in accurately judging response quality may correspond to a failure to deploy strategies aimed at improving performance (Butler and Winne 1995). In particular, this may reflect a monitoring failure, or a deficiency in students' online awareness of their strategy use during task completion (Schraw 1998; Thiede et al. 2003).

Second, lower levels of objective performance and inaccurate calibration may both be attributable to deficits in task model construction. Britt and Rouet (2012) define a task model as students' cognitive representation of task demands and how these may be satisfied. If students have poor or incomplete cognitive models of what writing an argument or a research report entails, they may not only produce low quality written responses but also inaccurately calibrate their performance. Deficits in task model construction may be particularly likely to arise when students are presented with a complex task, such as one asking them to compose a written response based on multiple texts (Pieschl 2009; Stahl et al. 2006). As compared to selecting the correct answer when responding to a multiple choice item, constructing a task model for writing an argument or a research report requires students to simultaneously recognize that they need to write organized and elaborated responses, that present and integrate evidence, and reflect a unique perspective on the target issue – resulting in a much more complex task model needing to be formed.

Furthermore, our findings indicate that better writing quality was associated with students' relative under-confidence. This is consistent with prior research which has likewise found that under-confident students tend to perform better than over-confident students on a range of tasks (e.g., Bol et al. 2005). While we may expect high self-efficacy to be positively associated with task performance (e.g., Bandura 1982; Nietfeld and Schraw 2002; Pintrich and De Groot 1990), this was not the case in the present study. It may be that for some tasks, or for some students, a comparative under-confidence is associated with better performance by stimulating strategy use or the further expenditure of time and effort.

To further examine the role of under-confidence in performance, we conducted a median split to examine the associations between objective performance and calibration accuracy and confidence bias for high-performing versus low-performing students, as a supplemental analysis. We found that calibration accuracy and confidence bias were not associated with objective performance for high-performing students but were strongly and negatively associated with objective performance for low-performing students. This discrepancy indicates that low-performing students, in particular, are hindered by their deficits in being able to accurately evaluate the quality of their written responses. This corresponds to prior research that high-performers are more able to accurately evaluate their performance as well as to better render judgments of calibration as compared to low-performers (e.g., Boud et al. 2013; Chiu and Klassen 2010; Hacker et al. 2008). This also suggests that lower-performing students may especially benefit from training and scaffolding in calibration and metacognitive monitoring.

Research question 4: Self-evaluation criteria and writing quality

For the fourth research question, we examined the extent to which the justification criteria that students cited were associated with objective task performance, calibration, and confidence bias. Results showed that students' citations of various justification criteria significantly predicted their objective task performance and confidence bias. In particular, students' citations of structure, superficial source use, deep-level source use, citation use, and evidence evaluation as justification criteria were found to predict objective task performance. This seems logical as these criteria corresponded to the rubric that we used to evaluate students' written responses. Moreover, these criteria have been found to impact written response quality in prior work (Anmarkrud et al. 2014; Reznitskaya et al. 2009). These same criteria were also found to predict students' confidence bias, in addition to objective task performance. This indicates that students' attention to the same criteria that improve task performance also contribute to more accurate judgments of response quality.

Notable is that both objective performance and confidence bias were predicted by students' consideration of a multitude of criteria in judging response quality. The need to coordinate these multiple criteria, directed toward task, text, and learner-related factors, is what may make the calibration of complex task performance particularly difficult for students. Further, these findings point to the need to systematically examine the multitude of dimensions that students may consider in forming judgments of task performance (Dinsmore and Parkinson 2013; Lin and Zabrucky 1998).

In follow-up analyses we examined the extent to which the four most frequently cited criteria for judging task performance (i.e., superficial source use, context, writing, and personal attributes) contributed to students' objective scores, calibration, and confidence bias. Our findings showed that these criteria significantly predicted objective scores and confidence bias. This provides further evidence that students considered both task-related and person-related characteristics in producing a written product and rendering subjective judgements of task performance. Notably, superficial source use was the only individually significant predictor of students' objective task performance in this more data-driven model. Seemingly, this finding seems to be a deviation from prior work which has found students' deep-level source use, or cross-textual elaboration, to contribute to multiple text response quality, to the exclusion of surface-level source use, or information accumulation (Bråten and Strømsø 2011). One caveat is nevertheless important in contextualizing this anomalous finding. Superficial source use was an individually significant predictor only in data-driven models predicting response quality. These models, in including only frequency-based justification criteria, did not include deep-level source use as a possible predictor. Therefore, we take these results to mean that any attendance to source use in self-evaluating response composition is an effective strategy for students to use, even if considering deep-level source use may ultimately prove to be more fruitful. Put another way, results from this study may suggest that supporting students to think about their source use, at all, even in a superficial manner lays the groundwork for further reflection on the extent to which they evaluated, integrated, or otherwise deeply engaged with texts during response composition.

As a final point, while various justification criteria were associated with both objective task performance and with confidence bias, these were not significant in predicting calibration accuracy. We believe this was because computations of calibration reflect only the absolute difference between students' objective and subjective performance, with no consideration of whether students are over- or under-estimating response quality. This may serve to attenuate the variability in calibration scores we are able to model and, in particular, to mask students' who are over- rather than under-confident in their performance. The analyses in this paper suggest the importance of examining multiple indicators of calibration, including both absolute accuracy and confidence bias, in understanding students' metacognitive evaluations.

Limitations and future directions

Despite the strengths of this study, a number of limitations must be acknowledged. First, our study only examined students' postdictions, or self-evaluations of writing following task completion, rather than examining predictions of response quality. This was done because the literature on calibration has suggested that students' postdictions represent more accurate self-evaluations than their predictions of performance (i.e., the post-diction superiority effect; Glenberg and Epstein 1985; Lin et al. 2001; Pierce and Smith 2001). Nevertheless, students' predictions of performance may be important to examine in future work. These may determine students' efficacy for task completion as well as their inclination to deploy various strategies to improve performance. A particularly fruitful investigation may be to compare students' predictions and postdictions of writing task performance to determine how these may change in association with task completion.

Second, in this study, we coded the self-evaluation criteria that students cited when justifying their self-assigned letter grades. However, we did not examine the association between students' consideration of various criteria when self-evaluating task performance and actual behaviors during multiple text task completion. For instance, when students rated their responses favorably, as reflecting multiple text evaluation or integration, we did not verify the extent to which students engaged such strategies during task completion. This disconnect means that even when aware of the criteria that contribute to generating a high quality response, students may fail to act in accordance with these criteria during multiple text use. As such, examining students' justifications for response quality in relation to strategy use during multiple text use constitutes a promising direction for future work. Nevertheless, we were encouraged by the association among students' objective task performance, confidence bias, and criteria cited when self-evaluating response quality. This also draws our attention to the association between individual difference factors and students' justifications for their self-evaluations of writing quality. For instance, students with high prior knowledge or more extensive task experience, may have more sophisticated criteria for justifying their evaluations of writing quality. These criteria may stem from students receiving external feedback on prior work or from students' self-reflections on prior task completion (Glenberg et al. 1987). Moreover, a higher level of self-regulation and metacognition, more generally, may mean that students are better at self-evaluating response quality (i.e., have a broader set of task criteria to use or are more accurate in their self-assessments) and at engaging compensatory strategies, when they find response quality to be wanting. Examining the association between students' individual difference factors and self-evaluations is another direction for future work.

Further, in this study students were assigned to complete one of two writing tasks (i.e., to compose an argument or a research report based on multiple texts). Nevertheless, despite this difference in task assignment, these were scored using the same rubric. While scoring students' written responses in a similar fashion, across task conditions, is commonly done to allow these to be compared to one another (Le Bigot and Rouet 2007; Wiley and Voss 1999), some of the variability in task assignment may not have been captured by the use of a common rubric. An additional goal in future work may be to explicitly link students' task goals, generated prior to multiple text use, with their self-evaluations of response quality, following task completion. We believe that such an investigation would contribute greatly to theoretical calls to further examine the role of self-regulation and metacognition in students' learning from multiple texts (List and Alexander 2019; Rouet et al. 2017). This would also constitute a validity check that could be used as evidence of students' differential task conceptions and as a basis for associating these with calibration performance.

At the same time, capturing task goals presents a number of challenges. In prior work (List et al, under review), we have explicitly asked students to report their perceptions of different task assignments; however this was done in a rather decontextualized fashion (e.g., with no specific topic assigned). Another method to consider in future work is specifically asking students to elaborate on their perceptions of a specific task, prior to processing. However, this may change the course of students' multiple text use, by stimulating a greater degree of self-regulation and metacognition, and may be limited by students' unwillingness to report all of the factors they consider in representing a task (e.g., a desire to put in little time or effort).

Moreover, students were not able to revisit library texts once they had proceeded to the writing task. This constitutes a limitation in study design. In particular, prior work has found revision to be a key part of the writing process (Spivey and King 1989). Allowing students to revisit texts, as they were writing, could have contributed to this revision process, improving students' writing quality and self-evaluation accuracy (Wallace et al. 2007). Moreover, enabling students to revisit texts during writing may have given us unique insights into their metacognitive engagement or strategy use as a direct result of self-evaluation.

Finally, our study was carried out in a lab setting. It is clear from the justification criteria that students cited, that the study context that they were presented with was in some ways atypical of how they usually compose written responses based on multiple texts. For instance, the number of texts that students had access to was limited and students perceived there to be a time limit, even though there was none. Nevertheless, as reflected in their self-evaluations, even in this lab setting, students were fairly engaged during task completion and perceived themselves as composing written responses of a relatively high quality. This points to the need to both examine students' objective performance, calibration, and confidence bias in naturalistic classroom settings and to consider how aspects of the lab environment may facilitate or hinder task completion.

Conclusions and implications

Findings from this study indicate that students consider a broad range of criteria when self-evaluating their written responses composed based on multiple texts. In this study, we examined students' objective task performance, calibration, and confidence bias as associated with one another and with the justification criteria that students reported when self-evaluating response quality. We found that both more accurate calibration and relative under-confidence contributed to more effective task performance. Moreover, low-performing students were particularly found to suffer from an over-confidence bias and less accurate calibration, when compared to their high-performing peers. In combination, these findings suggest that teachers should help students to consider a variety of criteria when self-evaluating complex task performance and should specifically scaffold low-performing students to render more accurate judgments of performance, perhaps as a means of improving writing quality. Our study also suggests that asking students to self-evaluate their responses may be a viable avenue for getting students to be more metacognitively and situationally conscious of the demands of complex task completion.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest and that there is no funding to declare in association with this project.

Appendix 1

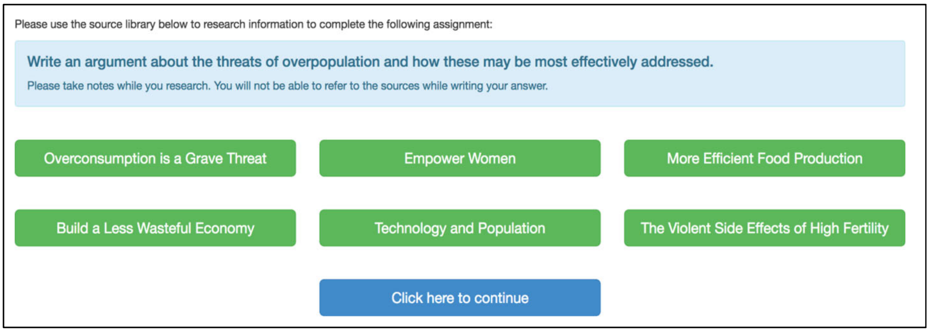
Prior Knowledge Measure

The study will ask you to research and write an argument/research report about overpopulation. To start off, please define each term related to overpopulation. If you don't know the definition of a term, please write N/A.

1. Population bomb
2. Earth's carrying capacity
3. Overconsumption
4. Peak population
5. High-yield crops
6. Overpopulation
7. Fertility rate

Note. Responses to the prior knowledge measure were scored as correct or incorrect, with students' total prior knowledge scores ranging from zero to seven. Specifically, students received a point for any response that uniquely (i.e., differentiated from similar terms) and accurately described a term using non-synonymous language (e.g., defining 'population bomb' as 'a type of bomb' was scored incorrectly)

Appendix 2



Please use the source library below to research information to complete the following assignment:

Write an argument about the threats of overpopulation and how these may be most effectively addressed.

Please take notes while you research. You will not be able to refer to the sources while writing your answer.

Overconsumption is a Grave Threat
Empower Women
More Efficient Food Production
Build a Less Wasteful Economy
Technology and Population
The Violent Side Effects of High Fertility
Click here to continue

Fig. 4 The digital library of the six texts presented to students

References

Afflerbach, P., & Cho, B. Y. (2009). Determining and describing reading strategies: Internet and traditional forms of reading. In H. S. Waters & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 201–225). New York: Guilford.

Allwood, C. M., Granhag, P. A., & Jonsson, A. C. (2006). Child witnesses' metamemory realism. *Scandinavian Journal of Psychology, 47*(6), 461–470.

Anmarkrud, Ø., Bråten, I., & Strømsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multile conflicting documents. *Learning and Individual Differences, 30*, 64–76 https://doi.org/10.1016/j.lindif.2013.01.007.

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1*(1), 3–38.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*(2), 122–147.

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*(2), 133–151. https://doi.org/10.1080/00220970109600653.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269–290.

Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology, 37*(4), 280–287. https://doi.org/10.1016/j.cedpsych.2012.02.004.

Boud, D., Lawson, R., & Thompson, D. (2013). Does students engagement in self-assessment calibrate their judgement over time? *Assessment and Evaluation in Higher Education, 38*(8), 941–956.

Bråten, I., & Strømsø, H. I. (2009). Effects of task instruction and personal epistemology on the understanding of multiple texts about climate change. *Discourse Processes, 47*(1), 1–31. https://doi.org/10.1080/01638530902959646.

Bråten, I., & Strømsø, H. I. (2011). Measuring strategic processing when students read multiple texts. *Metacognition and Learning, 6*(2), 111–130.

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*(4), 485–522. https://doi.org/10.1207/S1532690XCI2004_2.

Britt, M. A., & Rouet, J. F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276–314). New York: Cambridge University Press.

Britt, M. A., & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing task. *Reading Psychology, 25*, 313–339.

Britt, M. A., Perfetti, C. A., Sandak, R. L., & Rouet, J. F. (1999). Content integration and source separation in learning from multiple texts. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of tom Trabasso* (pp. 209–233). Mahwah: Erlbaum.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281.

Cavaleri, M., & Dianati, S. (2016). You want me to check your grammar again? The usefulness of an online grammar checker as perceived by students. *Journal of Academic Language and Learning, 10*(1), A22–A236.

Cerdán, R., & Vidal-Abarca, E. (2008). The effects of tasks on integrating information from multiple documents. *Journal of Educational Psychology, 100*(1), 209–222. https://doi.org/10.1037/0022-0663.100.1.209.

Chiu, M. M., & Klassen, R. M. (2010). Relations of mathematics self-concept and its calibration with mathematics achievement: Cultural differences among fifteen-year-olds in 34 countries. *Learning and Instruction, 20*, 2–17.

Cho, B. -Y., & Afflerbach, P. (2017). An evolving perspective of constructively responsive reading comprehension strategies in multilayered digital text environments. *Handbook of research on reading comprehension*, 109–134.

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgements made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001.

Dole, J. A., Duffy, G. G., Roehler, L. R., & Pearson, P. D. (1991). Moving from the old to new: Research on reading comprehension instruction. *Review of Educational Research, 61*(2), 239–264.

Du, H. & List, A. (2019). Writing based on multiple texts. (Manuscript submitted for publication)

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–86.

Fallahi, C. R., Wood, R. M., Austad, C. S., & Fallahi, H. (2006). A program for improving undergraduate psychology students' basic writing skills. *Teaching of Psychology, 33*(3), 171–175.

Firetto, C. M. (forthcoming). Learning from multiple complementary perspectives: a systematic review. In: Van Meter, P., List, A., Kendeou, P., & Lombardi, D. (Eds.), Handbook of learning from multiple representations and multiple perspectives. New York: Routledge.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911.

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. I. (2010). Summary versus argument tasks when working with multiple documents: Which is better whom? *Contemporary Educational Psychology, 35*(3), 157–173. https://doi.org/10.1016/j.cedpsych.2009.11.002.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 702–718. https://doi.org/10.1037/0278-7393.11.1-4.702.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*(1), 84–93.

Glenberg, A. M., Sanoki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*(2), 119–136.

Goldman, S. R., & Scardamalia, M. (2013). Managing, understanding, applying, and creating knowledge in the information age: Next-generation challenges and opportunities. *Cognition and Instruction, 31*(2), 255–269. https://doi.org/10.1080/10824669.2013.773217.

Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30*(2), 207–241. https://doi.org/10.1016/j.cedpsych.2004.08.001.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*(7), 1001–1013. https://doi.org/10.3758/MC.37.7.1001.

Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 19–34). New York: Springer.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160.

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–456). New York: Taylor & Francis.

Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgements of confidence. *Learning and Instruction, 24*, 37–47. https://doi.org/10.1016/j.learninstruc.2012.10.001.

Higham, P. A. (2013). Regulating accuracy on university tests with plurality option. *Learning and Instruction, 24*, 26–36. https://doi.org/10.1016/j.learninstruc.2012.08.001.

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgements to improve metacognitive monitoring. *Metacognition and Learning, 4*, 161–176. https://doi.org/10.1007/s11409-009-9042-8.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica, 77*, 217–273.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609–639.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*(4), 279–308.

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*, 173–194. https://doi.org/10.1007/s11409-010-9056-2.

Le Bigot, L., & Rouet, J. F. (2007). The impact of presentation format, task assignment, and prior knowledge on students' comprehension of multiple online documents. *Journal of Literacy Research, 39*(4), 445–470.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159–183.

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implication for education and instruction. *Contemporary Educational Psychology, 23*(4), 345–391. https://doi.org/10.1006/ceps.1998.0972.

Lin, L. M., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22*, 111–128.

List, A., & Alexander, P. (2015). Examining response confidence in multiple text tasks. Metacognition and Learning, 10, 407–436. https://doi.org/10.1007/s11409-015-9138-2.

List, A., & Alexander, P. (2017). Analyzing and integrating models of multiple text comprehension. Educational Psychologist, 52(3), 143–147. https://doi.org/10.1080/00461520.2017.1328309.

List, A., Alexander, P. A., & Stephens, L. A. (2017). Trust but verify: Examining the association between students' sourcing behaviors and ratings of text trustworthiness. Discourse Processes, 54(2), 83–104. https://doi.org/10.1080/0163853X.2016.1174654

List, A. (under review). Six questions regarding strategy use when learning from multiple texts. In: D.L. Dinsmore, L.K. Fryer, & M.M. Parkinson (Eds.). Handbook of strategies and strategic processing: conceptualization, intervention, measurement, and analysis. New York: Routledge

List, A., & Alexander, P.A. (2019) Toward an integrated framework of multiple text use, Educational Psychologist, *54*(1), 20–39. https://doi.org/10.1080/00461520.2018.1505514.

List, A., Du, H., & Wang, Y. (2019). Understading students' perceptions of task assignments. (Manuscript submitted for publication)

List, A., Du, H., Wang, Y., & Lee, H. Y. (2019). Toward a typology of integration: Examining the documents model framework. Contemporary Educational Psychology, 58, 228–242. https://doi.org/10.1016/j.cedpsych.2019.03.003.

Mateos, M., & Solé, I. (2009). Synthesising information from various texts: A study of procedures and products at different educational levels. *European Journal of Psychology of Education, 24*, 435–451.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 502–506.

Mosenthal, P. B. (1998). Defining prose task characteristics for use in computer-adaptive testing and instruction. *American Educational Research Journal, 35*(2), 269–307.

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*(3), 131–142. https://doi.org/10.1080/00220670209596583.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006a). The effects of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1*, 159–179.

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006b). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*(2), 258–271.

Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Towards a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Hillsdale, NJ: Erlbaum.

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition, 29*(1), 62–67.

Pieschl, S. (2009). Metacognitive calibration–an extended conceptualization and potential applications. *Metacognition and Learning, 4*(1), 3–31. https://doi.org/10.1007/s11409-008-9030-4.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33–40.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25*(1), 19–33.

Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics, 20*(1), 18–41.

Reznitskaya, A., Kuo, L., Glina, M., & Anderson, R. C. (2009). Measuring argumentative reasoning: What's behind the numbers? *Learning and Individual Differences, 19*, 219–224. https://doi.org/10.1016/j.lindif.2008.11.001.

Rouet, J. F. (2006). *The skills of document use: From text comprehension to web-based learning*. Mahwah, NJ: Erlbaum.

Rouet, J. F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McGrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Charlotte, NC: Information Age.

Rouet, J. F., Britt, M. A., & Durik, A. M. (2017). RESOLV: Readers' representation of reading contexts and tasks. *Educational Psychologist, 52*(3), 200–215. https://doi.org/10.1080/00461520.2017.1329015.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*, 113–125.

Schraw, G., & Nietfeld, J. (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology, 90*(2), 236–248.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87*(3), 433–444.

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibration ten commonly used calibration scores. *Learning and Instruction, 24*, 48–57.

Snyder, K. E., Nietfeld, J. L., & Linnenbrink-Garcia, L. (2011). Giftedness and metacognition: A short-term longitudinal investigation of metacognitive monitoring in the classroom. *Gifted Child Quarterly, 55*(3), 181–193. https://doi.org/10.1177/0016986211412769.

Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2004). Metacognition and self-regulated learning constructs. *Educational Research and Evaluation, 10*(2), 117–139.

Spivey, N. N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Research Quarterly, 24*, 7–26.

Stahl, E., Pieschl, S., & Bromme, R. (2006). Task complexity, epistemological beliefs and metacognitive calibration: An exploratory study. *Journal of Educational Computing Research, 35*(4), 319–338.

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437–475.

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance of environmental feedback. *Organizational Behavior and Human Decision Processes, 83*(2), 282–309.

Strømsø, H. I., Bråten, I., Britt, M. A., & Ferguson, L. E. (2013). Spontaneous sourcing among students reading multiple documents. *Cognition and Instruction, 31*(2), 176–203. https://doi.org/10.1080/07370008.2013.769994.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73.

Wallace, R., Pearman, C., Hail, C., & Hurst, B. (2007). Writing for comprehension. *Reading Horizons, 48*(1), 41–56.

Wiley, J., & Voss, J. F. (1996). The effects of 'playing historian' on learning in history. *Applied Cognitive Psychology, 10*, 63–72.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*(2), 301–311.

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning internet science inquire tasks. *American Educational Research Journal, 46*(4), 1060–1106.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 153–189). Mahwah: Erlbaum.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 279–306). Mahwah: Erlbaum.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology, 27*(4), 551–572. https://doi.org/10.1016/S0361-476X(02)00006-1.

Winne, P., & Perry, N. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). San Diego, CA: Academic Press.

Wolfe, M. B. W., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction, 23*(4), 467–502. https://doi.org/10.1207/s1532690xci2304_2.

Yates, J. F. (1990). *Judgement and decision making*. Englewood Cliffs: Prentice-Hall.

Zabrucky, K. M., Agler, L. L., & Moore, D. (2009). Metacognition in Taiwan: Students' calibration of comprehension and performance. *International Journal of Psychology, 44*(4), 305–312. https://doi.org/10.1080/00207590802315409.