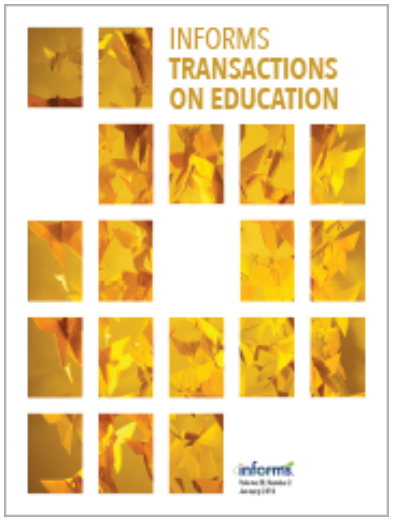


INFORMS Transactions on Education

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

That's Incorrect and Let Me Tell You Why: A Scalable Assessment to Evaluate Higher Order Thinking Skills

Kyle D. S. Maclean, Tiffany Bayley

To cite this article:
Kyle D. S. Maclean, Tiffany Bayley (2024) That's Incorrect and Let Me Tell You Why: A Scalable Assessment to Evaluate Higher Order Thinking Skills. INFORMS Transactions on Education 25(1):23-34. https://doi.org/10.1287/ited.2023.0020

https://pubsonline.informs.org/journal/ited

INFORMS TRANSACTIONS ON EDUCATION
Vol. 25, No. 1, September 2024, pp. 23–34
ISSN 1532-0545 (online)

That's Incorrect and Let Me Tell You Why: A Scalable Assessment to Evaluate Higher Order Thinking Skills

Kyle D. S. Maclean,[a,*] Tiffany Bayley[a]

[a] Ivey Business School, Western University, London, Ontario N6G 0N1, Canada
*Corresponding author
Contact: kmaclean@ivey.ca, https://orcid.org/0000-0001-7216-7633 (KDSM); tbayley@ivey.ca, https://orcid.org/0000-0001-6572-3259 (TB)

Received: March 28, 2023
Revised: July 1, 2023
Accepted: July 24, 2023
Published Online in Articles in Advance: September 1, 2023

https://doi.org/10.1287/ited.2023.0020

Abstract. We introduce a novel type of assessment that allows for efficient grading of higher order thinking skills. In this assessment, a student reviews and corrects a technical memo that has errors in its formulation or process. To overcome the grading challenges imposed by essay-type responses in large undergraduate courses, we provide a Visual Basic for Applications Excel tool for instructors, ensuring efficient grading of student submissions. We report our experience using this type of assessment in a multisection introductory business analytics course over several years and present survey-based evidence indicating that students perceive it to be clear and beneficial for learning.

Supplemental Material: Data is available at https://www.informs.org/Publications/Subscribe/Access-Restricted-Materials

Keywords: high-order thinking • memo • analytics • business • communication • grading • critique • comment • Word • Excel • VBA

1. Introduction

Large undergraduate courses are ubiquitous in higher education, particularly for foundational subjects that lead to advanced topics. In business education, introductory courses in business statistics and business analytics often have hundreds of students enrolled. Depending on the institution, these large undergraduate courses may consist of a single offering with all students in class simultaneously or as multiple sections with lower enrollment (e.g., under 100 students) per section.

With multisection introductory courses (MICs) taught by several faculty members, teaching responsibilities are coordinated so that learning experiences across the sections are similar and address the common learning objectives of the course. In addition, assessments must be carefully developed and evaluated so that grading practices are applied consistently to every section and for every student. Whereas fair grading practices are always considered regardless of enrollment size, they become increasingly challenging to implement consistently when there are several graders involved in the evaluation process. One way to mitigate (or even eliminate) inconsistencies is to pose well-structured questions that have clearly defined answers, for example, multiple choice or algorithmic problems (Jonassen 2000, Miltenburg 2019). Those types of problems test lower order thinking skills in Bloom's taxonomy of educational objectives and are ideal for assessing immediate knowledge recall and basic understanding or application of formulas (Anderson and Krathwohl 2001). They do not, however, adequately test students' ability to analyze, synthesize, or evaluate an ill-structured or open-ended problem.

Ill-structured problems typically require students to demonstrate understanding through essay-style responses or in-depth qualitative and quantitative analyses. For science, technology, engineering, and mathematics fields in particular, educators regularly encourage students to go beyond simply stating a final quantitative result and practice their communication skills. Carrithers and Bean (2008) and Williams et al. (2016) do this through memo-writing assignments for students that require managerial interpretation following numerical analyses. Of course, students may one day find themselves as the recipients of these memos, making it crucial to develop higher order thinking (HOT) skills, such as judgement or evaluation, to discern the analysis that took place and interpret its value in business decision making. For their master of business administration course, Saladin and Shafer (2006) developed a class activity in which students verbally critiqued an assessment presented to them, allowing students to exercise both their technical knowledge and critical thinking skills.

Learning from errors is not new; identifying and correcting mistakes has been an effective pedagogical tool

for self-evaluation in K–12 mathematics education (Bray 2013), college physics (Mason et al. 2016), and healthcare simulations (King et al. 2013). Using a student's evaluation of their own mistakes, educators are able to reframe the negative connotation associated with making errors and instead turn those experiences into positive learning moments (Alvidrez et al. 2022). Furthermore, erroneous examples in assessments provide students with opportunities to reflect upon and demonstrate their own understanding of the material in the hopes of not repeating those errors in the future (Quibble 2004, Yerushalmi and Polingher 2006, Koehler 2020). Große and Renkl (2007) note that, whereas students are able to explain in their own words why erroneous statements are incorrect, they are less likely to describe the underlying principles that would correct the error; requiring students to also fix the mistake closes the loop in this learning cycle. It is true that identifying and correcting error-filled examples can lead to more "confrustion" (confusion and frustration) compared with simply solving incomplete examples (Richey et al. 2019), but student learning in the long run does not suffer.

Evaluating these HOT skills can be challenging as acceptable answers may not necessarily be a single numerical value. Xiong and Suen (2018) find that grading ill-structured problems in massive open online courses was difficult because student answers vary substantially, and providing individualized feedback could not be easily scaled. Lack of grading resources to efficiently provide meaningful feedback is also a concern (Buswell et al. 2019). To alleviate this shortcoming, some instructors have turned to peer evaluation (e.g., Harris 2011), which outsources a portion of the grading or developed rubrics and processes to enable timely grading practices (e.g., Essig et al. 2014). Still, the logistics of grading can be difficult to coordinate among graders, particularly in MICs with several hundred students.

In our introductory business analytics course, we aimed to evaluate students' HOT abilities without increasing grading effort. We achieved this by designing an assignment, dubbed memo evaluation toward mistake elimination (MEME), that leverages features of Microsoft Word and Visual Basic for Applications (VBA). Our assignment puts the student in the role of a "critic," who is given a memo that outlines, at times erroneously, technical analyses used to arrive at a business recommendation. The student must identify and rectify several errors using the comment feature in a Word document. With a custom VBA script developed by the first author of this paper, these comments are extracted to an Excel database in which they can be filtered and sorted by "error." This enables a straightforward way to divide grading tasks among a team of graders, resulting in efficient grading of HOT skills.

The rest of this paper is structured as follows. We first describe the design of this assessment method and provide more details on the instructions given to students. Next, we address the assessment's scalability and introduce an Excel-based software program to facilitate grading. We then present survey-based evidence demonstrating MEME's effectiveness and conclude with a discussion of best practices for faculty who wish to design their own assignments based on this method.

2. The Assessment

Students are given a memo in a Microsoft Word document that describes technical analyses used to arrive at some final recommendation for a business problem. Each step of the analysis is detailed but at times contains errors in assumptions, reasoning, or explanations. Students must read the memo and identify the errors within the document. For each error identified, students select the relevant text and add a comment to the document (accessed through the review tab from the ribbon or by right-clicking the selected text). The comment should explain why the selected text is an error and, if applicable, how to correct it.

2.1. Preparing the Memo

The memo itself is prepared by the instructor and is written clearly and concisely with figures and equations provided as needed. Topics and concepts we have used in previous course offerings include recommending a real estate listing price using regression analysis and developing a production plan for a small food company using linear programming. Errors included in the memo are intended to test students' understanding of those concepts, so they may be in the form of incorrect application of technique (e.g., using too many dummy variables in a regression) or misinterpretation of output values (such as shadow prices). Several errors of varying severity (as determined by the instructors) are included in the document. Low severity could mean going against a known convention but otherwise correct analysis (e.g., reversing the x and y axis in a chart), and high severity would be something that is blatantly incorrect (e.g., assuming that any data set with more than 30 data points is normally distributed). The memo is uploaded to the course's learning management system (LMS) as an editable Word document so that students can use the comment feature. This enables the custom VBA script to extract the comments automatically.

2.2. Instructions to Students

The full set of instructions we provided to students is in the appendix, and we describe the most important aspects here.

Students are to identify a subset, S, of the errors in the memo focusing on those with higher severity. The size of the subset, $|S|$, is determined by the instructor and communicated to students. Because only the first $|S|$ comments will be graded, there is no incentive to comment

on more errors than required. Students are instructed to consider errors to be independent of one another, commenting on each error only once without considering its effects on later parts of the memo. The students do not need to edit the text of the memo itself, nor do they need to perform any additional analyses.

Each comment should indicate why the selected text is an error and, if possible, describe how to correct it. To encourage concise, technical writing and discourage verbosity, students are instructed to keep comments under 300 characters (not including spaces). Similar to the number of comments, only the first 300 characters of any comment will be graded.

To give context to the comment when grading, students are instructed to select full sentences or equations rather than single words or entire paragraphs. The technical reasons for this approach are detailed in the next section.

As this assignment format may be new to students, we provide an example memo with suggested comments to serve as a guideline for student submissions (Figure 1). The example memo describes the process of building a decision tree in order to provide a recommendation though many sentences contain subtle errors in process or interpretation. We elaborate with a few error–comment pairs with the first demonstrating an interpretation error:

Error 1: The expected value is an important concept because it tells you the most likely outcome.

Comment 1: The EV does not tell you the most likely outcome but rather the long-run average if this were repeated many, many times. This is not necessarily the same as the most likely outcome.

The second is a process error, for which the suggested comment explains why the sentence is an error and *how* it would be corrected.

Error 2: We calculate the expected value by "rolling back" the decision tree given in exhibit 1 from left to right.

Comment 2: Generally speaking, by convention, a decision tree is rolled back from right to left. That should also be done in this case because we see in exhibit 1 that time moves from left to right. Therefore, the first calculations should be done at the furthest right node.

A memo can also contain distractor sentences. Similar to distractor options in multiple choice exams (Kilgour and Tayyaba 2016, Gierl et al. 2017), these sentences act as red herrings, requiring students to evaluate them carefully because they are not actually errors. For instance, the sentence, "The risk profile can be calculated by finding the probability of arriving at each outcome and creating a table of the possible outcomes and their total probability" is a reasonable description of a risk profile but does not conform to its traditional textbook definition (i.e., a histogram or frequency table of plausible outcomes and their associated probabilities).

We note that this example is for illustrative purposes, demonstrating to students the mechanics of selecting text and adding comments. It does not represent the same level of difficulty as the actual graded assessment. The supplemental materials contain a graded memo we used in 2022.

3. Achieving Efficiency Through Parallel Grading

Given the large student enrolment in MICs, a team of graders is necessary to deliver timely feedback on student assessments. Operationally, it is sensible for student submissions to be graded in sequence or in parallel.

Under sequential grading, each student submission is assigned to a specific grader, who evaluates all rubric items for that submission before moving to the next one. This method is useful for rubric items that depend on one another (e.g., when an answer to a part (a) is used to answer something in part (b)) or when a submission requires integrated evaluation. For instance, a full essay or group project report would likely have to be graded sequentially. In a sense, each grader is a jack of all trades. As an example, in the top portion of Figure 2, we see how both graders 1 and 2 are responsible for all four rubric items, and they each grade a mutually exclusive subset of all student assignments.

Alternatively, a grader could specialize, and the entire grading process might resemble an assembly line. Rather than grade an entire assignment from start to finish, each grader becomes proficient at grading a subset of rubric items. We call this grading in parallel because a single student submission is graded by multiple graders. The bottom of Figure 2 shows how grader 1 is responsible for rubric items 1 and 2 for all students, whereas grader 2 handles rubric items 3 and 4. Note that the total grading workload for the MIC is identical under each method. It is merely how the work is divided that differs.

The decision of which operational process to use should be made carefully. Grading in sequence is simple to explain to both graders and students as each submission is assigned to a single grader. Grading in parallel, in turn, has the benefit of grading consistency because a rubric item is always graded by the same grader. That is, it likely results in higher interrater reliability because of specialization. It is also potentially efficient because a grader specializes in a certain portion of the rubric.

However, if each rubric item necessitates evaluating the entire submission, the potential efficiencies are lost because multiple graders are spending time reading the entire submission. Because of this, sequential grading is often used in which rubric items are interdependent or they are comprehensive in nature.

In our MEME assignment, students must decide whether each sentence contains an error and, if so, provide a comment. Considering that their comments are independent of one another, each erroneous sentence can be graded in parallel using a separate grading rubric. This

Figure 1. Sample Student Memo

In this analytical summary, we explain our rationale and methods behind recommending going forward with **Starting the Project.**

We first note that this is a one-time decision that is highly strategic. Therefore, we will use both the expected value (EV) as well as the risk profile as our *decision criteria* to examine. The EV is an important concept because it tells you the most likely outcome. The risk profile, on the other hand is important because it quantifies the probability of each unique outcome occurring.

We calculate the expected value by "rolling back" the decision tree given in Exhibit 1, from left to right. At each decision node, the highest expected value option is taken, and at each chance node the expected value is calculated. This process is done recursively until the entire tree is solved. In this case, using this method, we find that the EV of starting the project is $70, and the EV of not doing so is $0.

The risk profile can be calculated by finding the probability of arriving at each outcome, and creating a table of the possible outcomes and their total probability. The probability of arriving at each outcome is found by multiplying the probabilities of each branch that arrives to that outcome. Since we do not currently know what decision will be made at the Domestic vs International decision, we will assign a 50% probability of each decision occurring when calculating the risk profile. In Exhibit 2 we show the final risk profile.

We recommend that you **Start the project.** It has a higher EV and at the worst case, will only lead to a loss of $50 20% of the time.

Exhibit 1 – Decision Tree

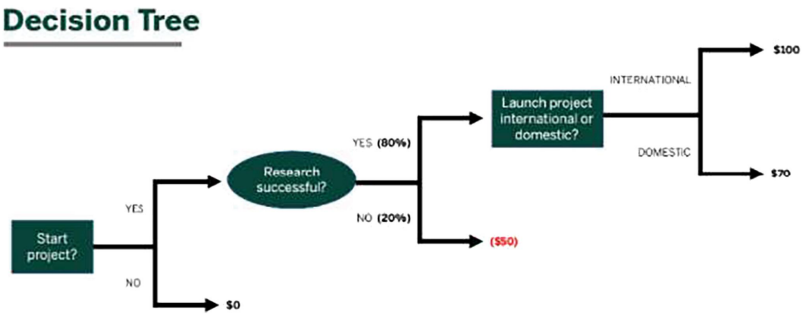


Exhibit 2 – Risk profile of starting the project

Probability

Outcome	Probability
$100	40%
$70	40%
$-50	20%

Commented [A1]: The EV does not tell you the most likely outcome, but rather the *long run average* if this were repeated many many times. This is not necessarily the same as the most likely outcome.

Commented [A2]: Generally speaking, by convention, a decision tree is rolled back from right to left. That should also be done in this case since we see in Exhibit 1 that time moves from left to right. Therefore, the first calculations should be done at the furthest right node.

Commented [A3]: This is not the correct procedure. Instead, the decision that is taken must be given a probability of 100%, and all other decisions can be given a probability of 0%.

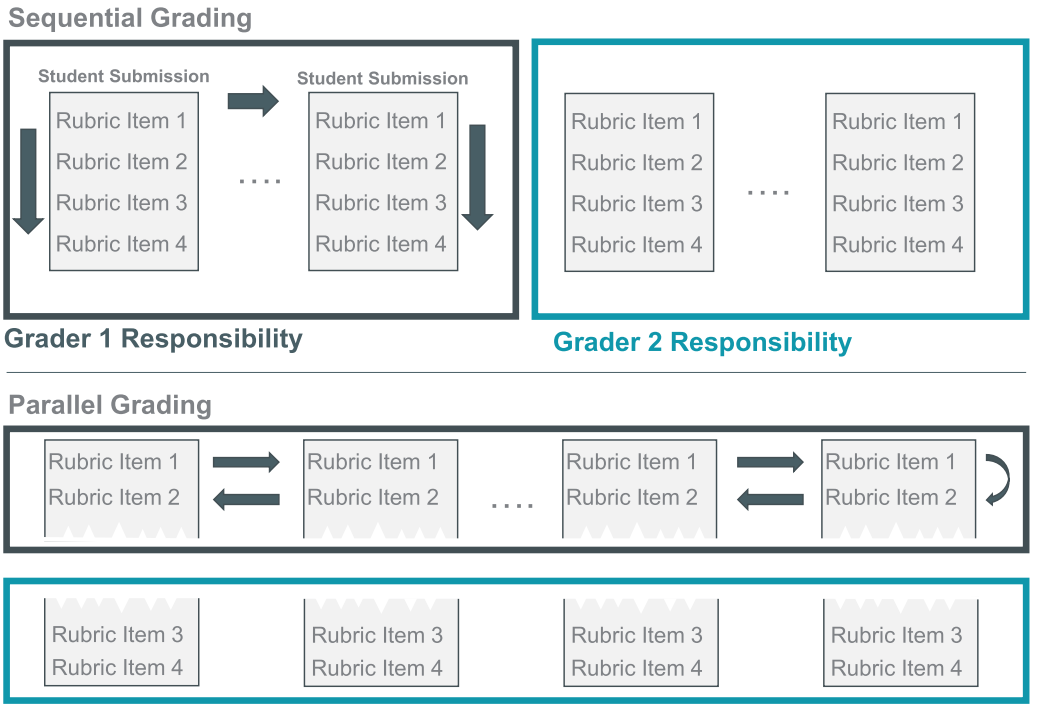
enables MEME grading to be both efficient and consistent, enabling its scalability for MICs.

4. Supplementary Grading Tool

Grading hundreds of assignments in parallel can be challenging to implement, especially for a MEME assessment for which students' error selection and comment content will vary. A grader may need to open many or all student submission files individually, find the specific sentence(s) for which the grader is responsible, and determine whether a student commented on that sentence. This would be time-consuming and inefficient. In this section, we describe a custom-built software tool to operationalize parallel grading, which is available as an online supplement.

Our Excel-based VBA programmed tool allows faculty to aggregate all student comments into one worksheet.

Figure 2. Sequential vs. Parallel Grading

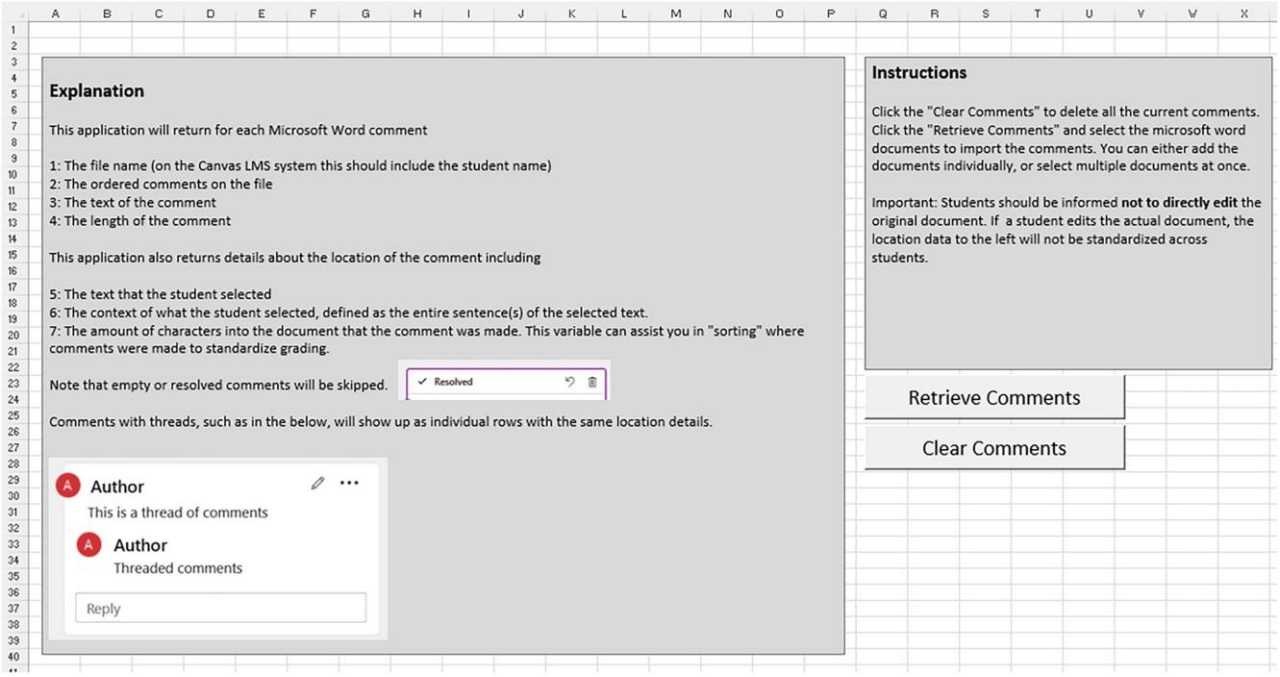


Notes. Conceptual diagram of the differences between sequential grading and parallel grading. The superimposed boxes describe the responsibilities of two hypothetical graders.

This combination was chosen as Microsoft Excel is a widely used spreadsheet application, readily available to faculty at postsecondary institutions. Furthermore, VBA is a well-known programming language for Microsoft Office products, offering easy editing for end-user programmers.

Before using the tool, we assume that all student files have been downloaded and are available on a local or network drive accessible by the user. The program does not automatically access an LMS via application programming interface. However, this is not a practical

Figure 3. User Interface of the VBA Tool



issue. In our experience, commercial LMS systems allow faculty to download student submissions as a zip file. In that case, the user would simply need to unzip these files prior to using our program.

To begin extracting student comments in the Excel/VBA tool, the user clicks the "retrieve comments" button (Figure 3). A file dialog box opens, and the user should select all Microsoft Word documents that students have submitted with comments. If it is not possible to select all documents, for example, because they are located in separate folders, the user can perform this operation multiple times to access all of the documents. If the program has been previously run and the spreadsheet is populated with data, the user can first reset the entire workbook by clicking the "clear comments" button before proceeding with this step.

The program loops through all the Word files selected by the user. For each file, the program loops through all comments in the document. Blank or "resolved" comments are automatically ignored so that content is not extracted to the Excel sheet. In our experience, some students will "resolve" a comment as an alternative to deleting it; the comment is not displayed in the document margin but still exists in the background.

For each comment retrieved, the program adds a row to the Excel sheet with the following information:

1. File name: The file name that the comment is under. For some LMSs (e.g., Canvas), the file name automatically contains information on the student name and student ID. Thus, we use the file name as a unique ID for each student. Under systems in which the file name does not automatically include information, an instructor could ask students to use a particular naming convention.

2. Comment ID: An ordered number of each retrieved comment per student. For instance, one indicates the first comment for a specific file, two indicates the second comment, etc. By concatenating this integer with the file name, you can create a unique ID for each comment.

3. Comment: The string of the comment. This is retrieved using the VBA `.Range.Text` property.

4. Comment length: The character count (not including spaces) of the comment using the VBA method `.Range.ComputeStatistics(3)`. The argument "3" indicates the specific statistic that we want to return.[1] For instance, you could use `.Range.ComputeStatistics(5)` to return the number of characters including spaces. This is useful when an assignment specifies maximum or minimum length requirements, which we discuss further in Section 6.

5. Text selected: Using the `.Scope` property of the comment returns the text that the student selected when creating the comment.

6. Full context: Returns the entire sentence surrounding *text selected*. This is helpful because sometimes a student may select just a fragment of a sentence (e.g., a word or phrase), whereas another student may select the entire sentence. In either situation, the full context of *text selected* for both students is the same.

7. Character start: Returns the number of characters from the beginning of the memo up to the start *full context*. For instance, a value of 1,350 would indicate that *full context* begins 1,350 characters into the memo. This enables sorting student comments by where they appear in the memo. Note that *full context* is used, not *text selected*; if one student selects the first three words of a sentence and another selects the last three words of that sentence, the value of *character start* for each student would be the same because the underlying context is identical.

Figure 4 shows how retrieved comments appear in the Excel spreadsheet once extracted. Using the "full context" column, student comments can be sorted by the sentences with which they are associated and assigned to graders as necessary.

5. Evidence of Effectiveness

We used MEME assessments in Fall 2020, 2021, and 2022 offerings of a mandatory third-year undergraduate business analytics course at a Canadian university. The course enrolled approximately 700 students per year, across 8–10 sections taught by 4 or 5 faculty members. Each section had between 75 and 79 students, and we hired 8–10 graders. Our graders consisted primarily of upper year undergraduate students who had taken the previous course and performed well.

In 2020 and 2022, the memo developed pertained to linear regression. In 2021, the memo was about an optimization problem and designed to test linear programming knowledge. In all years, the assessment was one of three individual assignments given in the course with each assignment worth 10% of the overall course grade. In 2020 and 2021, students were given approximately 30 hours to complete the assessment, whereas in 2022,

Figure 4. Result of Running the VBA Tool

	A	B	C	D	E	F	G
1	File Name	Comment ID	Comment	Length	Text Selected	Full Context	Character Start
2	bsmith_Memo.docx	1	Month_Sold is a poor choice for a variable and cannot be included in the regression analysis since it includes information that is not available at the time the prediction is being made.	487	Month_Sold is the month (1=January, 2=February…) that the sale took place in, to	Month_Sold is the month (1=January, 2=February…) that the sale took place in,	1045
3	bsmith_Memo.docx	2	Both categorical variables "Month_Sold" and "Location" have been incorrectly coded. In N categories, you need N-1 variables; this means that when coding for 4 categories, you include 3	473	Location is a categorical variable coded as 1=Terrible, 2=Good, 3=Very good,	Location is a categorical variable coded as 1=Terrible, 2=Good, 3=Very good,	1272
4	bsmith_Memo.docx	3	There is inconsistency between the visualization of the "Pool?" variable (Exhibit 2), and the final regression model (Exhibit 3). The negative coefficient observed in the final model implies that a	499	Pool? is equal to 0 if the house has a pool, and 1 if the house does not have a pool.	Pool? is equal to 0 if the house has a pool, and 1 if the house does not have a	1365
5	bsmith_Memo.docx	4	While you do want to remove predictor variables listed as non-significant, you must do so one at a time rather than all at once. In order to correct this mistake, you should re-run the regression	520	We then removed all variables that were listed as non-significant to generate a	We then removed all variables that were listed as non-significant to generate a	2126

this window was extended to a five-day period to provide students with additional flexibility and autonomy of their prioritization skills. In all years, the authors anticipated that each assignment would require students, on average, four hours to complete.

For each error, we assigned a maximum possible score based on the severity of the error. This usually ranged in magnitude from one to five points. Partial credit was awarded if comments were incomplete, that is, the error was identified only with no or incomplete explanation as to why it is an error or error rectification was not provided. To compute the final grade, the individual comment scores were added and divided by the maximum score theoretically possible. The maximum score theoretically possible consisted of the $|S|$ highest valued errors (in our assessments, we used $|S| = 7$).

5.1. Operational Effectiveness

We found that MEME worked well to scale grading for hundreds of students, especially with the use of our supplementary grading tool. For the 2022 offering, the resulting spreadsheet contained on the order of 5,000 individual comments from the more than 600-student cohort and an allocation of roughly 500 comments for each of our 10 graders.

Recall that not all sentences in the memo are equal; some have higher or lower severity errors on which nearly all students comment, whereas others are distractors with few student comments. To assign an equitable workload to our graders, we first sorted comments by *character start* to group them by their position in the memo. Then, we assigned groups of comments associated with a sentence to a specific grader.

One unusual issue we encountered when sorting by *character start* was that it occasionally differed among students who selected the same sentence. This discrepancy was sometimes because students added paragraph breaks, whereas other times, the reason was not obvious. We suspect that version differences in .doc/.docx file types and Mac/PC differences were responsible for these inconsistencies. However, it did require some manual sorting to resolve the issue.

We hypothesized and our graders later confirmed that grading 700 comments on a single sentence was easier than evaluating 100 comments on each of 7 different sentences. Memo sentences that generated more student comments tended to be "obvious" errors and were straightforward to grade. In contrast, "long-tail" sentences with fewer and highly varied comments took longer to grade because they were not intended to be errors and, thus, required the creation of new rubric items with associated grading keys.

Even with these "long-tail" sentences, we found that the MEME format worked extremely well in scaling grading for hundreds of students. We were especially encouraged in our first year of piloting MEME when a grader, who was initially skeptical about grading in parallel, described to us how much easier MEME was to grade and maintain consistency compared with other assignments.

5.2. Teaching Effectiveness

We also gathered student feedback regarding their impressions of this assignment format. In 2022, our 10-section MIC had three individual assignments. The first and last assignments acted as a control to MEME and consisted of traditional business analytics problems; students were presented with an extensive word problem and required to perform a numerical analysis along with a concise written recommendation supported by their findings. The second assignment employed the MEME format. After completing each assignment, students were required to complete an anonymous, ungraded survey conducted through the LMS with the following questions:

1. Overall, I found this assignment to be a good learning experience.
2. I found this assignment to be clear and understandable.
3. In total, approximately how many hours did you spend on this assignment?

Students rated their level of agreement with statements in questions 1 and 2 on a five-point scale from *strongly agree* to *strongly disagree*. we also provided a text entry box for additional comments.

To quantitatively analyze survey results, we removed "I don't know/Prefer not to answer" responses and converted Likert-scale responses accordingly (strongly agree = 5 and strongly disagree = 1). Table 1 presents descriptive statistics for each question, and Figure 5 provides the mean value for each assignment and survey question combination with 95% confidence intervals. To determine whether these differences were statistically significant, we performed a pairwise t-test between all assignments with a Bonferroni correction for multiple comparisons.

Our primary interest was in how students perceived MEME as a learning experience (question 1). MEME received a higher average agreement of 3.84 (out of 5) compared with 3.62 and 3.75 for assignments 1 and 3, respectively. MEME was statistically different from assignment 1 ($p < 0.001$) but not statistically different from assignment 3 ($p = 0.233$).

We were also curious about whether students found the MEME format confusing or unclear in any way (question 2). This is a natural concern because of its novelty. The average agreement under MEME was 4.04 as compared with 3.84 and 3.99 for assignments 1 and 3, respectively. MEME was statistically different from assignment 1 ($p < 0.001$) but not statistically different from assignment 3 ($p = 0.633$). Perhaps counterintuitively, students generally found MEME to be as clear or clearer than other, more traditional, assignments. In addition,

Table 1. Descriptive Statistics of Survey Results

Question	Assignment	*N*	Mean	Median	Standard deviation	Minimum	Maximum
Overall, I found this assignment to be a good learning experience.	Control (assignment 1)	678	3.618	4	0.957	1	5
	MEME	724	3.840	4	0.828	1	5
	Control (assignment 3)	667	3.750	4	1.052	1	5
I found this assignment to be clear and understandable.	Control (assignment 1)	701	3.839	4	0.825	1	5
	MEME	740	4.043	4	0.767	1	5
	Control (assignment 3)	673	3.987	4	0.917	1	5
In total, approximately how many hours did you spend on this assignment?	Control (assignment 1)	700	7.000	6	3.112	1.0	24
	MEME	742	3.967	3	4.371	0.5	85
	Control (assignment 3)	671	9.341	8	9.818	2.0	218

Note. Observations are not equal because of the removal of "I don't know" responses.

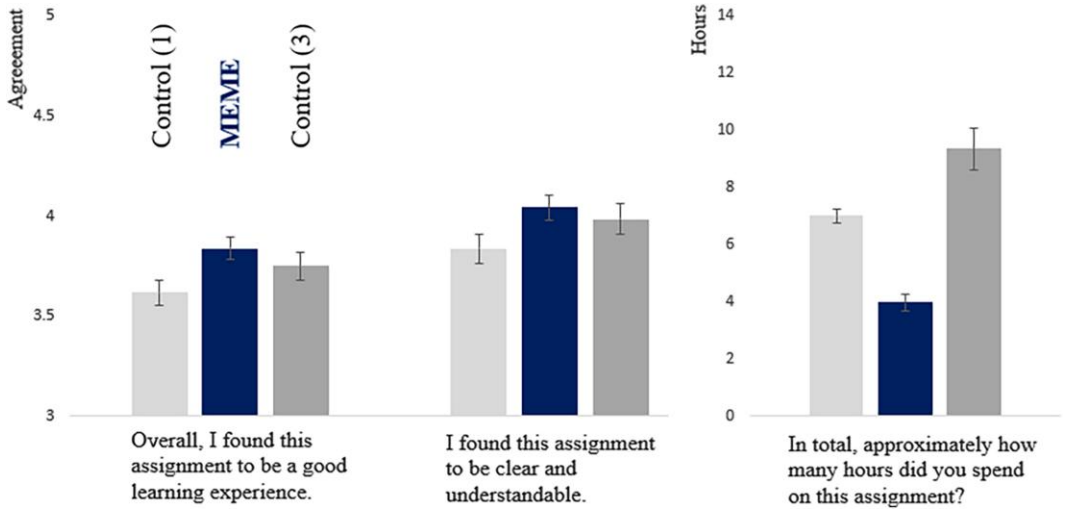
whereas students were working on MEME, we noticed a decrease in the number of clarification questions we typically receive, confirming our previous observations.

Finally, because MEME is a novel format, we were uncertain about the time it would take for students to complete. From question 3, we learned that MEME took significantly less time, with an average of 3.97 hours, compared with assignments 1 and 3, requiring 7.00 and 9.34 hours, respectively. This was statistically different from both assignment 1 ($p < 0.001$) and assignment 3 ($p < 0.001$).

Taken together, the quantitative results from our survey are reassuring. On all dimensions, MEME performed as well as or better than our traditional assignments. The format was just as clear, offered students a comparable learning experience, and took less time for them to complete. The following are selected survey responses from students that describe their experience more qualitatively.

- I found it exciting to get to review others [*sic*] work and spot corrections. I believe it forced me to look at the analytical process from a different lens, and it challenged me to truly understand why something was wrong. Further, I enjoyed practicing explaining mistakes in a concise manner, vs. giving long explanations.
- Found it very interesting editing from this perspective. It makes you realize how having correct wording is so important and in statistics, one word can change the meaning of a sentence.
- I like the format of the assignment because it actually tests our knowledge and understanding—you could've given us a set of data and asked us to do a regression and analysis and we could do it and get the right answers without understanding it at all, just memorizing the process. This forces us to actually think critically about each step of the process and what it all means and why, which I think is a much better test of our understanding. As hard as it was, I think it's a really great format.
- I really liked this one! It had clear real-life applications, completing the assignment reinforced my in-class

Figure 5. Survey Results



Notes. Mean response by survey question and assignment. Error bars indicate 95% confidence intervals.

learning, it wasn't too time-consuming, the instructions were accessible, and it was kind of fun looking for the different errors!

- I have never completed an assignment in this manner before (i.e., analyzing and selecting errors), and as such, I thoroughly enjoyed the learning experience. It was helpful to apply concepts we have learned in class to a real-world problem, and definitely forced me to stretch my critical thinking skills.

Students also provided negative feedback, from which we elicit that the main criticism was that the MEME format was "ambiguous." For the most part, these types of survey comments did not expand upon what was ambiguous, so ironically, there was ambiguity in what those students found ambiguous. One student alluded to difficulty in deciding on which errors to focus because the student found more than the maximum limit specified in our instructions. But, of course, this is by design. Because of the vague nature of these negative comments and the presence of similar comments found in control assignments 1 and 3 and because the average clarity rating was actually highest on MEME, we decided that no changes to assignment structure or instructions were necessary in future iterations.

To supplement this student survey with feedback from other stakeholders, we also solicited thoughts from our graders and faculty. The faculty members from whom we requested feedback were other instructors of our course but not the authors of this paper.

Our graders found this assignment to be straightforward to mark and that grading in parallel enabled them to apply the grading key consistently. One grader remarked that the grader "enjoyed grading the assignment because there was only a set number of errors so it was easy to mark." The grader also referenced a previous experience with MEME as a student in the course: "As a student, I really enjoyed the commenting assignment because correcting something was a challenging way to test my knowledge. The assignment was difficult but the commenting aspect made it very manageable and less time-consuming."

Another grader insightfully pointed out a limitation: there is a distinction between identifying modeling errors and the ability to create a model from scratch: "I believe that learning how to analyze models and conclusions comes with the skill of being able to create models. Whereas I don't believe the opposite is necessarily true, i.e., if you can build a model, you can likely analyze one, but analyzing one does not imply you can build one. I felt that just because someone received a high grade, that did not mean I felt confident they could prepare an equally accurate model themselves."

From the faculty perspective, one instructor noted, "One thing that I particularly like about this type of assignment is that it tests students' understanding of basic fundamental concepts. It is also a good practice for students to identify problems on real-world applications of theoretical concepts that they learn in classroom. It raises an awareness that reports need to be evaluated critically in the real world."

One difficulty this faculty member identified was in deciding how to weight different errors: "In this type of assignment, there are usually more issues than what students are asked to identify. Different issues might have different level of importance. Some might not be as serious as other ones. In this case, do we assign equal weights to each issue? That does not sound fair. If we assign different weights, how do we assign those weights?"

As we discuss in Section 6, having different error severity is an important feature of this assignment. Although assigning weights can be subjective (similar to any question in any assignment), it is not arbitrary.

6. Best Practices and Our Advice for Designing Your Own Analytical Memo

Given the success we had using MEME in our own MIC, we hope other instructors will adopt this format and write memos for their own courses. We provide some considerations and best practices for adopting this assessment type and for developing your own memo. These best practices are based on our own experience, but we also speculate about when it may be appropriate to deviate from them.

1. Be clear about the "ground truth." When writing your memo, pay critical attention to what information should be taken as a given versus what information can be criticized. For instance, in our linear programming memo, some students criticized the company's assumption that they could sell whatever amount of product that could be produced. In a managerial setting, this may be a fair criticism, but it was not the focus of this assignment. We have found it useful to structure the memo as two parts. The first outlines the business problem and any assumptions that are not to be criticized. Narratively, we present this as an email from a client to an analytical firm or consultant. The second part of the memo outlines the analytical methods and the "solution." This acts as a response from the analytical firm or consultant to the client. We then instruct students to only add comments to the second part. By splitting the memo into two segments, it becomes immediately apparent which portions are "fair game" for criticism.

2. Be clear that each error should be *independent* of one another. In the assignment instructions, you should make it clear that every error is independent of other errors. That is, whereas one error may impact multiple conclusions in the memo, the student should select a singular instance of the error and then assume it has been corrected throughout the rest of the report. If this is not clear to students, their comments may be

duplicative and make grading in parallel challenging if not impossible. Of course, for the instructor, this should be considered when developing the memo itself.

3. Be clear about whether you are assessing presentation errors. In our first year of piloting MEME, we produced a memo focused on predictive analytics that contained several scatterplots and a histogram of residuals from a multiple regression model. Some students criticized the presentation of these charts, for example, not including grid lines. However, the intent of the charts was to provide material for students to criticize regarding whether assumptions underpinning an ordinary least squares regression may have been violated. Thus, it is important to be explicit in the assignment instructions whether presentation or communication errors are open for criticism or not. In our assignment, we include the following instruction: "Do not criticize any typos, grammatical errors, or general communication practices."

We recognize, however, that criticizing presentation may be an intentional learning outcome in some courses that can be assessed with this type of assignment. This would likely present a challenge when using our software tool as it is difficult in Microsoft Word to select a specific area of a chart for commenting. As a result, our software is not able to provide the *full context* of the selection, which makes grouping these errors difficult to carry out, ultimately hindering the parallelization of the grading process. This is an area for further exploration for analytics courses with a focus on visualization.

4. Provide errors of different magnitude. Part of the higher order learning that MEME facilitates is enabling students to notice various errors. They must then decide on their own volition which ones are substantial errors. This process of critically thinking through the implications of an assumption is part of higher order learning. Therefore, when designing your own memo, ensure that you are deliberate about introducing errors of different magnitude.

5. Provide an upper character limit for each comment. To prevent students from guessing at multiple concepts and overcomplicating their responses, we recommend setting an upper character limit for each comment. This ensures that comments can be graded in parallel and provides some expectations to students about the level of detail to provide. It also helps develop students' ability to communicate concisely.

It is important to specify this length limit in terms of characters rather than the number of sentences. In our first pilot of MEME, we used a five-sentence limit. In turn, students were creative in their interpretation of what could be considered a "sentence," resulting in some comments exceeding 200 words and with extremely poor grammar.

Though we have not tested this idea, we speculate that setting a word limit may be effective for memos that do not require students to correct equations. However, if equations are involved in the task, it may be cumbersome for students to determine how to count equation terms as words, potentially causing confusion.

7. Discussion and Conclusion

In this paper, we discuss our development of a novel assignment format that challenges students to comment on errors in a technical memo. We provide a custom VBA Excel tool for instructors to aggregate Microsoft Word comments from student submissions so that parallel grading can be executed. This MEME format allows for the assessment of HOT skills, allowing scalable benefits from parallel grading. After using MEME over the last three years in a large MIC on business analytics, our survey-based evidence indicates that students perceive this type of assignment as equally or more comprehensible than our traditional assignments. It offers an equivalent or superior learning experience, requiring less time to complete.

We did not objectively measure the impact of MEME on student learning, but rather relied on students' self-reported perception of learning to gauge its effectiveness. Just as the quality of a multiple-choice question can induce lower or higher order thinking, the quality of a memo will also impact the opportunities students have to demonstrate their learning. We believe that focusing on the outcomes of our individual memos would be misleading. Instead, the contribution of this paper is to introduce MEME as a worthwhile format that is suitable for other instructors to tailor to their own teaching goals in any subject area.

There is also an open question of how recently developed large language models (LLMs), such as ChatGPT, may be used by students in answering MEME assessments. For instance, can MEME be answered by LLMs and, thus, be rendered obsolete? Because all three of our assessments were conducted prior to the announcement of ChatGPT, we do not have personal experience with how they might impact MEME assessments. We are also careful to note that advancements in this area are moving quickly, and making predictions is bound to be error-prone in hindsight.

However, with these cautionary notes out of the way, we surmise that LLMs will not render MEME obsolete. First, if memos are mixed media, including charts, tables, and analysis, LLMs will be more likely to struggle with comprehending them in totality. Currently, LLMs are not able to easily interpret images and certainly not able to easily interpret them in context of an analysis. Second, unless the prompts to the LLM are carefully constructed, LLMs are perhaps more likely to critique typos/grammatical errors or errors in communication, elements that the assessment instructions explicitly say to ignore. Third, if developed well, the memos should relate to what the

course instructor has emphasized in class. The LLM would not be aware of what has been taught in class and what has not and may end up being unnecessarily "picky."

We expect that LLMs will make MEME assessments more valuable, not less. Because LLMs are well-known to "hallucinate" or confidently state incorrect information (Ji et al. 2023, Rudolph et al. 2023, Zhang et al. 2023), it has arguably become more valuable for students to be able to critically examine pieces of writing for errors. This is a skill that MEME emphasizes and can help students develop.

Building on this work, further research can focus on two areas. First, whereas we have written our own memos for this assessment type, we do not have evidence beyond our own experiences of what characteristics make a memo a good or bad learning experience. Further research can explore which types of memo errors drive HOT skills and how this may depend on subject area. Second, it would be worthwhile to explore the extent to which further automation may be helpful in ensuring high-quality grading. For instance, if a memo's rubric is designed such that merely selecting the relevant sentence is rewarded points, it would be feasible to automate the entire process. And, as discussed previously, the advancement of LLMs could also be used to assist in grading. Future research can focus on how to leverage LLMs to pregrade or sort comments, thus reducing the human interaction required and shortening the feedback cycle.

Appendix. Assessment Instructions to Students

Your job is to review the analytical memo for any analytical errors/mistakes/oddities. Do not criticize any typos, grammatical errors, or general communication practices. Rather, select errors/mistakes/oddities within the analytical process completed for the client. Think critically at each step.

In the Microsoft Word file provided, select what you consider to be the most important errors, inconsistencies, and questionable assumptions and add a comment to the associated sentence/equation/etc. To do so, in Microsoft Word go to the Review tab, select something, and click "New Comment" to explain your issue. Be succinct. There is a maximum character count of 300 characters per comment (spaces not included). Once you have made your comment on an issue, move on—do not keep bringing up the same issue across the entire report. See below for other instructions on comments.

You will be submitting

- ONE Microsoft Word file of your selections and comments.
- Please select and comment on full sentences or equations. Do not select entire paragraphs. Be specific about where you find errors/or take issue.
- Your comments should explain/discuss the issue briefly—for instance, when applicable: why you suspect it is an error/mistake/problem, what would be the solution (if there is one), what might be impacted by this error/mistake/problem.
- If you identify a type of error that applies to multiple formulas/equations, select only one, describe the error there, and quickly describe which formulas it applies to, then skip the rest of the formulas that it also applies to.
- Each comment you make should be independent. Once you comment on an issue in one place, do not continue to bring up the effects of the error you've already identified. Once commented, assume that that error has been fixed.
- It is not necessary to open Excel and it is not necessary to use a calculator for this assignment. While there may be logical errors in the formulation or calculations, there are no errors in the actual mathematical calculations themselves.
- Each comment should be a maximum of 300 characters, not including spaces. These should be relatively brief and to the point, not an essay for each.
- You can add a maximum of 7 comments to this memo. If you find more than that number of errors, select only the most critical errors.
- Do not edit the text itself. You are merely adding comments. Any edits to the text itself will be ignored and not graded.

Your grade will be based on

- The number of, and severity of errors that you identify and your succinct discussion of those errors. Not all errors are equally severe, and not everything in the report is an error necessarily.

Endnote

[1] The full list of statistic options can be found at https://docs.microsoft.com/en-us/office/vba/api/word.wdstatistic.

References

Alvidrez M, Louie N, Tchoshanov M (2022) From mistakes, we learn? Mathematics teachers' epistemological and positional framing of mistakes. *J. Math. Teacher Ed.* 1–26.

Anderson LW, Krathwohl DR (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Addison Wesley Longman, Inc., New York).

Bray WS (2013) How to leverage the potential of mathematical errors. *Teaching Children Math.* 19(7):424–431.

Buswell NT, Jesiek BK, Troy CD, Essig RR, Boyd J (2019) Engineering instructors on writing: Perceptions, practices, and needs. *IEEE Trans. Professional Comm.* 62(1):55–74.

Carrithers D, Bean JC (2008) Using a client memo to assess critical thinking of finance majors. *Bus. Comm. Quart.* 71(1):10–26.

Essig RR, Troy CD, Jesiek BK, Boyd J, Buswell NT (2014) Adventures in paragraph writing: The development and refinement of scalable and effective writing exercises for large-enrollment engineering courses. *Proc. 2014 ASEE Annual Conf. & Exposition* (American Society for Engineering Education, Washington, DC), 24.141.1–24.141.24.

Gierl MJ, Bulut O, Guo Q, Zhang X (2017) Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Rev. Ed. Res.* 87(6):1082–1116.

Große CS, Renkl A (2007) Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning Instruction* 17(6):612–634.

Harris JR (2011) Peer assessment in large undergraduate classes: An evaluation of a procedure for marking laboratory reports and a review of related practices. *Adv. Physiology Ed.* 35(2):178–187.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. *ACM Comput. Surveys* 55(12):1–38.

Jonassen DH (2000) Toward a design theory of problem solving. *Ed. Tech. Res. Development* 48(4):63–85.

Kilgour JM, Tayyaba S (2016) An investigation into the optimal number of distractors in single-best answer exams. *Adv. Health Sci. Ed. Theory Practice* 21(3):571–585.

King A, Holder MG, Ahmed RA (2013) Errors as allies: Error management training in health professions education. *BMJ Quality Safety* 22(6):516–519.

Koehler AT (2020) *A Methodology for Teaching from Student Errors in Computer Science Education* (University of California, Riverside, CA).

Mason A, Yerushalmi E, Cohen E, Singh C (2016) Learning from mistakes: The effect of students' written self-diagnoses on subsequent problem solving. *Phys. Teacher* 54(2):87–90.

Miltenburg J (2019) Online teaching in a large, required, undergraduate management science course. *INFORMS Trans. Ed.* 19(2):89–104.

Quibble ZK (2004) Error identification, labeling, and correction in written business communication. *Delta Pi Epsilon J.* 46(3): 155–168.

Richey JE, Andres-Bray JML, Mogessie M, Scruggs R, Andres JM, Star JR, Baker RS, McLaren BM (2019) More confusion and frustration, better learning: The impact of erroneous examples. *Comput. Ed.* 139:173–190.

Rudolph J, Tan S, Tan S (2023) Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *J. Appl. Learning Teaching* 6(1).

Saladin BA, Shafer SM (2006) Turning the tables on student case analysis assignments. *Decision Sci. J. Innovative Ed.* 4(1):169–173.

Williams JAS, Rankin M, Gallamore K, Reid R (2016) Beyond model formulation: Assessment of novices graphing, interpreting, and writing about their model and solution. *INFORMS Trans. Ed.* 17(1):13–19.

Xiong Y, Suen HK (2018) Assessment approaches in massive open online courses: Possibilities, challenges and future directions. *Internat. Rev. Ed.* 64(2):241–263.

Yerushalmi E, Polingher C (2006) Guiding students to learn from mistakes. *Phys. Ed.* 41(6):532.

Zhang M, Press O, Merrill W, Liu A, Smith NA (2023). How language model hallucinations can snowball. Preprint, submitted May 22, https://arxiv.org/abs/2305.13534.