9-28-2023

Assessing proxies of knowledge and difficultywith rubric-based instruments

Ben O. Smith
University of Nebraska at Omaha, bosmith@unomaha.edu

Jadrian J. Wooten
The Pennsylvania State University, jjw27@psu.edu

Received: 17 October 2022 | Accepted: 24 August 2023

DOI: 10.1002/soej.12658

TARGETING TEACHING

Southern Economic Journal WILEY

Assessing proxies of knowledge and difficulty with rubric-based instruments

Ben O. Smith[1] | **Jadrian J. Wooten**[2]

[1]Department of Economic, University of Nebraska at Omaha, Omaha, Nebraska, USA

[2]Department of Economic, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

Correspondence
Ben O. Smith, University of Nebraska at Omaha, Omaha, NE, USA.
Email: bosmith@unomaha.edu

Abstract

The fields of psychometrics, economic education, and education have developed statistically-valid methods of assessing knowledge and learning. These methods include item response theory, value-added learning models, and disaggregated learning. These methods, however, focus on multiple-choice or single response assessments. Faculty and administrators routinely assess knowledge through papers, thesis presentations, or other demonstrations of knowledge assessed with rubric rows. This paper presents a statistical approach to estimating a proxy for student ability and rubric row difficulty. Moreover, we have developed software so that practitioners can more easily apply this method to their instruments. This approach can be used in researching education treatment effects, practitioners measuring learning outcomes in their own classrooms, or estimating knowledge for administrative assessment. As an example, we have applied these new methods to projects in a large Labor Economics course at a public university.

KEYWORDS

assessment, measuring knowledge, rubric assessment

JEL CLASSIFICATION

A20, A22, A23, C63

1 | INTRODUCTION

One of the goals of the university education system is to increase the subject-area knowledge of the students. Within economics there is a debate on the comparative importance of the signaling value of a degree versus knowledge accumulation, but few would argue that knowledge accumulation is unimportant. In many ways what a given university is selling is the knowledge accumulation by a student by attending their institution.

A university education, however, is subject to information asymmetry in which the institution, and particularly its faculty, are more informed of what students will learn than the students at first enrollment. Information asymmetry creates market failure, and thus accrediting organizations serve as intermediaries to certify the quality of knowledge being sold. To the knowledge of the authors, *all* legitimate accrediting organizations put a particular emphasis on assurance of learning (AoL). This is true of the major regional accrediting organizations (e.g. Higher Learning Commission, Northwest Commission on Colleges and Universities) or field/degree specific accrediting organizations (e.g. Association to Advance Collegiate Schools of Business).

Beyond accreditation requirements, faculty are also interested in measuring learning. Classical test theory (Ebel, 1954; Kuder & Richardson, 1937; Traub, 1997) and Item Response Theory (De Ayala, 2013; Lord, 1980) work to separate item difficulty from student characteristics in multiple-choice exams. Flow of student exam performance estimation techniques such as Hake (1998) and Walstad and Wagner (2016) attempt to separate the exam performance of the student before the class from gains in exam performance during the class. Most recently, a family of empirical techniques have been developed to find the underlying learning parameters when employing a pre- and post-test measurement strategy (Smith, 2022; Smith & Wagner, 2018; Smith & White, 2021).

Prior measurement techniques focus primarily on multiple-choice or single-response exams. While immensely useful for measuring knowledge or learning in a program, they are not universally applicable. Some courses are more likely to focus on projects or applications rather than on standard multiple-choice or numerical exams. Some exam questions may move beyond being explicitly right or wrong, and instead focus on varying degrees of accuracy. Knowledge and learning are harder to measure in these contexts, but this is when measurements are the most useful for decision makers: Departments want to know the knowledge of their students at the end of their program, universities want to know the knowledge of their students at the end of their college education, and accreditation bodies want to know the total knowledge produced in higher education.

Measuring knowledge with projects (or papers, theses, etc.) is hard. Students are graded using rubrics where reaching a certain box on the rubric merits a particular score. Notably, students who go beyond what a rubric row requires do not often earn more points than those who simply met the requirement (i.e., the distribution is right censored). Moreover, the way the scores are weighted plays a role in the overall score for a given student compared to a multiple-choice assessment in which questions are likely to be weighted equally. In short, one cannot simply average the scores and call the output knowledge or learning. While some attempts have been made in the psychometric literature to measure knowledge with rubrics (e.g., Uto, 2021), to our knowledge, none have explicitly modeled the censoring problem.[1] In this paper, we show

[1] We have found three articles that involve a censored geometric distribution outside of the education context (Mori, Woolson & Woodworth 1994, Okada, Vandekerckhove & Lee 2018, Patel & Gajjar, 2010). However, all three involve very different approaches and contexts than the approach and context of this paper.

how to model and estimate this problem in a robust, understandable, way.[2] The end result is a measure of the probability of a student failing to achieve the goals in a rubric row while accounting for rubric row difficulty and censoring on the top box. Moreover, rubric row difficulty is estimated and can be used to diagnose issues in the instrument, class, or program in a similar manner as a difficulty index for multiple-choice questions.

2 | PROBABILITY MODEL

Typically a rubric row outlines a set of requirements students must achieve in order to earn marks in a particular box (row score). Implied in this structure, the student met all of the requirements listed in boxes prior to their final marking. For instance, consider the following hypothetical rubric row in which the first column is the lowest scored box and increases as students achieve more criteria:

Did not meet any of the requirements	The chosen topic was on-topic for the course, but the literature was lacking	The literature review was complete but contained errors	Full credit

We assume that a student marked in the third box has also completed all of the requirements for the second box. Similarly, a student marked in the fourth box would have also completed the requirements of the third box, and by extension the second box. In this way, we can think of a student as "surviving" the trials of a given box to move onto the next box. This can be viewed as a geometric distribution where the number of trials (k) simply end at the end of the rubric row. Using the example rubric row above, a student who was marked in the box on the far left would score $k=0$, if they were marked in the next box they would score $k=1$, and earning a mark in the fourth box above would result in $k=3$.

Assuming the probability of failure is $p\left(q_j, s_i\right)$ where q_j characterizes row j's difficulty and s_i characterizes student i's ability, the probability of a single row score then can be expressed as:

$$\Pr\left(k_{ij}\right) = \underbrace{\left(1-p\left(q_j, s_i\right)\right)^{\lfloor k_{ij} \rfloor}}_{1-\text{CDF } at\ k_{ij}} - \overbrace{\left(\left\lceil -\frac{k_{ij}}{b_j} \right\rceil + 1\right)}^{\text{Indicator function}} \underbrace{\left(1-p\left(q_j, s_i\right)\right)^{\lfloor k_{ij}+1 \rfloor}}_{1-\text{CDF } at\ k_{ij}+1}, \tag{1}$$

where k_{ij} is score k by student i on rubric row j and b_j is the highest achievable score on rubric row j. This expression is constructed using the difference in 1-Cumulative Distribution Function (CDF) values with an indicator function. When $k_{ij} < b_j$, the probability function is algebraically equivalent to the Probability Mass Function (PMF). However, when $k_{ij} = b_j$ the probability cap-

[2] The method outlined in this paper necessitates the use of high quality rubrics. If the grading is inconsistent or biased then the resulting estimates will be inaccurate. Like any statistical method, the estimates can only be as accurate as the data.

tures achieving the top box *or higher*.[3,4] Some students would continue to succeed with additional requirements (i.e. "trials"); all we know is they reached *at least* the top box.

With some algebraic simplification, Equation (1) can be used to construct the likelihood function (Equation 2) and log-likelihood function (Equation 3):

$$L=\prod_{i=1}^{n}\prod_{j=1}^{m}\left(1-p\left(q_j,s_i\right)\right)^{\lfloor k_{ij}\rfloor}\left(p\left(q_j,s_i\right)+\left(p\left(q_j,s_i\right)-1\right)\left\lceil-\frac{k_{ij}}{b_j}\right\rceil\right), \tag{2}$$

$$\text{Log}(L)=\sum_{i=1}^{n}\sum_{j=1}^{m}\lfloor k_{\text{ij}}\rfloor\text{Log}\left(1-p\left(q_j,s_i\right)\right)+\text{Log}\left(p\left(q_j,s_i\right)+\left(p(q_j,s_i)-1\right)\left\lceil\frac{-k_{\text{ij}}}{b_j}\right\rceil\right). \tag{3}$$

To estimate all j values of q_j and all i values of s_i, one must assume a specific functional form for $p\left(q_j,s_i\right)$. We suggest two options: $p\left(q_j,s_i\right)=q_j+s_i$ (where q_j and s_i must be between 0 and 1) or $p\left(q_j,s_i\right)=1/\left(1+e^{-\left(q_j+s_i\right)}\right)$. The former has the advantage of interpretability while the latter prevents the possibility of the estimated $p\left(q_j,s_i\right)$ exceeding one for a given student rubric row. Note that $p\left(q_j,s_i\right)$ is the probability of *failure* thus the probability that the estimated $p\left(q_j,s_i\right)$ would exceed one is comparatively low.

A Maximum Likelihood Function (MLE) can be solved by either solving the first order conditions or numerically maximizing the likelihood or log-likelihood function. To explore the first option, the partial derivative of Equation (3) with respect to a specific q_j can be seen below:

$$\frac{\partial\text{Log}(L)}{\partial\hat{q}_j}=\sum_{i=1}^{n}-\lfloor k_{ij}\rfloor\frac{p_1\left(\hat{q}_j,s_i\right)}{1-p\left(\hat{q}_j,s_i\right)}+\frac{p_1\left(\hat{q}_j,s_i\right)+p_1\left(\hat{q}_j,s_i\right)\left\lceil\frac{-k_{ij}}{b_j}\right\rceil}{p\left(\hat{q}_j,s_i\right)+\left(p\left(\hat{q}_j,s_i\right)-1\right)\left\lceil\frac{-k_{ij}}{b_j}\right\rceil}. \tag{4}$$

Note that the use of the hats in Equation (4) is to emphasize that this is the derivative with respect to a specific q_j. It would be repetitive to provide the derivative with respect to a specific s_i as it is the second argument in $p\left(q_j,s_i\right)$ and appears in no other places in the equation; thus the derivative is obvious from Equation (4). For completeness, we have included an exploration of the second and cross-partial derivatives in Appendix C.

It is our belief that the first order conditions cannot be solved algebraically with either of the suggested functional forms of $p\left(q_j,s_i\right)$. However, the log-likelihood function can be maximized numerically.

[3] For instance, suppose $k_{ij}=2$ and $b_j=3$, then Equation (1) would equal $\left(1-p\left(q_j,s_i\right)\right)^2-\left(\left\lceil-\frac{2}{3}\right\rceil+1\right)\left(1-p\left(q_j,s_i\right)\right)^3\Rightarrow\left(1-p\left(q_j,s_i\right)\right)^2-(0+1)\left(1-p\left(q_j,s_i\right)\right)^3\Rightarrow\left(1-p\left(q_j,s_i\right)\right)^2\left(1-\left(1-p\left(q_j,s_i\right)\right)\right)\Rightarrow\left(1-p\left(q_j,s_i\right)\right)^2p\left(q_j,s_i\right)$. If, however, $k_{ij}=b_j=2$ then Equation (1) would equal $\left(1-p\left(q_j,s_i\right)\right)^2-\left(\left\lceil-\frac{2}{2}\right\rceil+1\right)\left(1-p\left(q_j,s_i\right)\right)^3\Rightarrow\left(1-p\left(q_j,s_i\right)\right)^2-(-1+1)\left(1-p\left(q_j,s_i\right)\right)^3\Rightarrow\left(1-p\left(q_j,s_i\right)\right)^2$.

[4] The concept applied here is similar to a Tobit regression (Tobin, 1958) where a Probability Mass Function or Probability Density Function is applied to uncensored observations and some portion of a CDF is applied to censored observations.

3 | ESTIMATION SOFTWARE

As part of this paper, the authors have developed software to estimate the MLE presented in Section 2.[5] The software is a Python package named ProjectAssessment (PA). Python is an open source and free programming language that by some measures is the most popular in data science (Ozgur et al., 2021). In order to be the most useful to the most universities, we have opted for a popular, open source, and free platform. In this software section we will cover the functionality of the software (Section 3.1), using the software in Google Colaboratory (Section 3.2), installing the software on a local computer (Section 3.3), and using the software on a local computer (Section 3.4).

3.1 | Functionality of the software

The PA software numerically solves models presented in Section 2 using the modified Powell method (Powell, 1964; Press et al., 2007); this approach performs well with noisy data. This results in estimates for all j values of q_j and all i values of s_i. The program supports both of the proposed specifications for $p\left(q_j,s_i\right)$ but defaults to $p\left(q_j,s_i\right)=1/\left(1+e^{-\left(q_j+s_i\right)}\right)$. The program then performs one of two block bootstrapping procedures (Künsch, 1989): one that treats the students as independent blocks and one that treats the rubric rows as independent blocks. These procedures produce confidence intervals and p-values that account for the potential lack of independence in the disturbance across student i's rows (or rubric row j's rows). Finally, the program produces fit measures including Akaike information criterion (Akaike, 1974), Bayesian information criterion (Schwarz, 1978), McFadden's pseudo-R^2 (McFadden, 1979)[6], the likelihood-ratio test statistic, and a χ^2 p-value of the model based on Wilks' theorem (Wilks, 1938). The results can be displayed to the screen and saved to a set of comma separated values (CSV) files.

3.2 | Using the software in Google Colaboratory

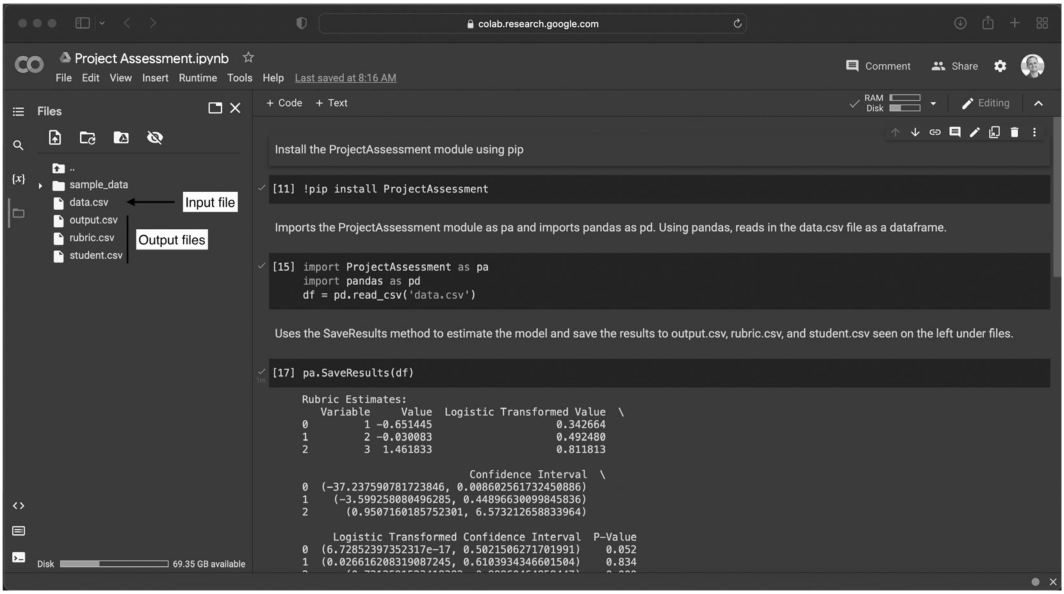
Google Colaboratory (Colab)[7] is a web-based Python environment that allows the user to use Python without installing Python on their own computer. Moreover, "notebooks" can be shared and duplicated for other users. We have created a notebook that walks the practitioner through the process of using the PA software and saving the results. This notebook can be found at https://bit.ly/3y176KQ.

From within Google Colab the practitioner will first need to save a copy of the notebook to their Google Drive account. They will then need to upload a data file to their Colab notebook and follow the instructions in the file. This will result in three saved files that the practitioner can download: output.csv, rubric.csv, and student.csv (Figure 1). We have created a video demonstrating this process at https://vimeo.com/735183858/9b6eb1f32e.

[5] The source code can be found at https://github.com/tazzben/project-based-assessment.

[6] One cannot expect McFadden's pseudo-R^2 to be as high as R^2 values produced in an ordinary regression. For instance McFadden suggested pseudo-R^2 values as low as 0.2 can be considered an excellent fit (McFadden, 1979, p. 306).

[7] https://colab.research.google.com.



Install the ProjectAssessment module using pip

```
[11] !pip install ProjectAssessment
```

Imports the ProjectAssessment module as pa and imports pandas as pd. Using pandas, reads in the data.csv file as a dataframe.

```
[15] import ProjectAssessment as pa
     import pandas as pd
     df = pd.read_csv('data.csv')
```

Uses the SaveResults method to estimate the model and save the results to output.csv, rubric.csv, and student.csv seen on the left under files.

```
[17] pa.SaveResults(df)

     Rubric Estimates:
        Variable     Value  Logistic Transformed Value  \
     0         1 -0.651445                    0.342664
     1         2 -0.030083                    0.492480
     2         3  1.461833                    0.811813

                             Confidence Interval  \
     0  (-37.237590781723846, 0.008602561732450886)
     1    (-3.599258080496285, 0.44896630099845836)
     2      (0.9507160185752301, 6.573212658833964)

          Logistic Transformed Confidence Interval  P-Value
     0  (6.72852397352317e-17, 0.5021506271701991)    0.052
     1  (0.026616208319087245, 0.6103934346601504)    0.834
```

FIGURE 1 An image of the Google Colab notebook. On the left side the input and output files are highlighted.

TABLE 1 The beginning of an example CSV file.

k	Bound	Student	Rubric
3	3	s1	1
3	3	s1	2
2	3	s1	3
2	3	s2	1
2	3	s2	2
3	3	s5	1
2	3	s5	2
1	3	s5	3

Note: "*k*" is the score achieved by the student on a given rubric row, "bound" is the maximum score possible on that rubric row and "student" and "rubric" are identifiers for the student and rubric row. The identifier columns need not be numeric or follow a specific pattern.

In the above notebook, we use an example data file (data.csv) that has the following columns: "*k*", "bound", "student", and "rubric". The top of the example data file can be seen in Table 1. In this file, we see k_{ij} values for each student rubric row (the column "*k*"), the upper bound of the given rubric row (b_j; the column "bound"), identifiers for students (the column "student"), and identifiers for the rubric rows (the column "rubric"). This format can easily be created by any spreadsheet application.

As Google Colab is a free service, it has obvious advantages over installing Python on a local computer. However, the disadvantage is performance. Code in Colab is executing on a virtual machine hosted by Google. The amount of computer resources allocated to the machine is a function of the number of other users on Google Colab. This is particularly pertinent for those

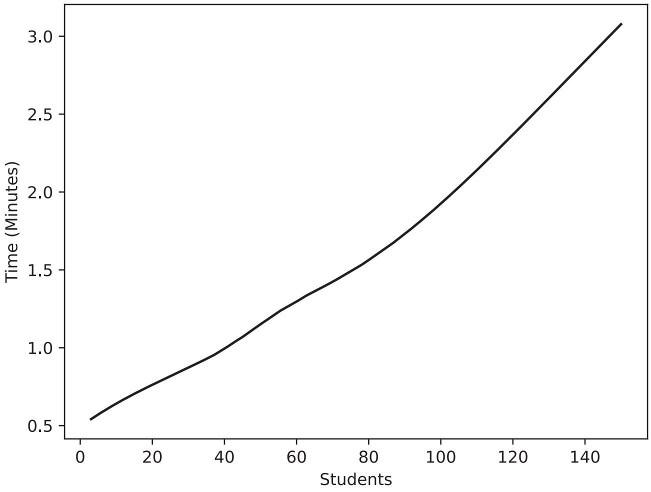


FIGURE 2 The number of minutes to execute the estimation procedure as a function of students in the dataset (eight rubric rows each). Time measurements were performed on an M2 MacBook Air with the internal cooling system saturated before initiating the test.

working with large datasets. In Figure 2 we graph execution times as a function of the number of students in the dataset; each student was represented by eight rubric rows. While the exact times on the left are not comparable across machines, this figure shows that as the number of students increase, the time to execute the procedure grows at a more than linear rate. When working with large datasets this can be mitigated by more processor cores. The PA software is designed to use all processor cores in the bootstrapping procedure. Therefore, execution time is nearly inversely proportional to the number of processor cores in the machine.

3.3 | Installing the software on the practitioner's computer

Before the practitioner can use the software on their own computer, they must install Python itself. The authors recommend users install the Anaconda distribution of Python available at https://www.anaconda.com/products/distribution.[8] The Anaconda distribution includes Python along with commonly used statistical and machine-learning packages. Moreover, it provides easy-to-use graphical interface installers for Windows and macOS.

Once Python is installed, the practitioner will need to install the PA software itself. This can be accomplished using one of two different package managers: pip or conda. pip is a built-in package manager included with most versions of Python while conda is specifically included with Anaconda.

From PowerShell on Windows or Terminal on macOS, the practitioner can type the following to install PA:

```
pip install ProjectAssessment
```

[8] As of the time of this writing, practitioners using Apple Silicon computers should scroll to the bottom of this page to make sure they are downloading the appropriate version of the distribution.

Or, using conda, the practitioner can type the following:

```
conda install -c tazzben projectassessment
```

Note that on some Windows installations, the conda application is not available in PowerShell. In these cases, start the Anaconda PowerShell from the start menu and type the conda command above.

3.4 | Using the software on the practitioner's computer

Once the PA package is installed and the practitioner has created some data, the module can be included in a Python script in the following way:

```
import pandas as pd
import ProjectAssessment as pa

def main():
    df = pd.read_csv('data.csv')
    pa.DisplayResults(df)

if __name__ == '__main__':
    main()
```

Here the pandas library is loaded as pd so that we can read a CSV into a dataframe and PA is loaded as pa. Within the main function, the CSV file data.csv is loaded into the variable df. df is passed to the PA method DisplayResults. This will display the estimates, confidence intervals, and fit estimates to the screen. The user could have called the method SaveResults instead of DisplayResults and saved the estimates, confidence intervals, and fit estimates to a set of CSV files.

In the above example, the method defaults were used. However, the practitioner might want to change these properties. In the example below, the practitioner has specified the "rubric", "linear", and "n". Setting rubric to true tells the module to bootstrap the rubric rows instead of the students; setting linear to true sets $p\left(q_j,s_i\right)=q_j+s_i$ instead of $p\left(q_j,s_i\right)=1/\left(1+e^{-\left(q_j+s_i\right)}\right)$; specifying n indicates the specific number of repetitions in the bootstrap procedure instead of assuming 1000; c indicates the probability in each tail of the generated confidence interval instead of assuming 0.025 for a 95% confidence interval.

```
import pandas as pd
import ProjectAssessment as pa

def main():
    df = pd.read_csv('data.csv')
    pa.DisplayResults(df, rubric=True, linear=True, n=10000, c=0.05)

if __name__ == '__main__':
    main()
```

The SaveResults method has the additional optional arguments rubricFile, studentFile, and outputFile. These arguments specify where to save the results instead of assuming "rubric.csv", "student.csv", and "output.csv", respectfully.

4 | EXAMPLE USAGE

We use data from an undergraduate Labor Economics course at a very large state school. This Labor Economics class regularly has in excess of 70 students and is offered every fall and spring semester. The class is structured around four written projects each covering a key content area. However, the rubric also evaluates writing quality. We will estimate the parameters using the data from Project 1 in Section 4.1. We interpret the rubric results in Section 4.2, and show the student outcomes in Section 4.3. In Section 4.4 we present Average Conditional Probability and Average Marginal Effect estimates. In Section 4.5, we will examine writing-specific identical rubric rows found in projects 1 and 4 to show the growth in ability over the semester. In Appendix A, we provide the Project 1 estimates for the Fall 2021 (Table A1 and Figure A1) and Spring 2022 (Table A2 and Figure A2) semesters individually.[9]

4.1 | Project 1: Estimates

In Table 2, we present the result of the Maximum Likelihood Estimation; we used the logistic form of the probability of failure ($p\left(q_j, s_i\right) = 1/\left(1 + e^{-\left(q_j + s_i\right)}\right)$) in this estimation procedure. In the far left column we present the rubric row indicator followed by the estimated values for q_j. The columns that follow are designed to help the practitioner interpret these estimations. The Average Logistic column evaluates the logistic function ($p\left(q_j, s_i\right)$) with the estimated value of q_j and each of the estimated s_i values. An average is then calculated. In the context of this estimation, for each of the rubric row estimations, the 161 corresponding student observations are extracted resulting in a subset where all values of q_j are the same (the value presented in the table) and each value of s_i corresponds to the 161 estimated values of s_i. The logistic function is calculated 161 times and then averaged. This can be interpreted as the average probability of failure to pass a trial given a specific rubric row.

The Average Marginal Logistic is calculated in a similar way to the Average Logistic but evaluates the derivative of the logistic function instead of the logistic function itself.[10] There are two possible equations for this derivative depending on if we are interpreting a rubric row (Equation 5) or a student row (Equation 6).

$$p_j\left(q_j, s_i\right) = q_j \frac{e^{q_j + s_i}}{\left(e^{q_j + s_i} + 1\right)^2}, \tag{5}$$

$$p_i\left(q_j, s_i\right) = s_i \frac{e^{q_j + s_i}}{\left(e^{q_j + s_i} + 1\right)^2}. \tag{6}$$

[9] Our analysis code and data can be found at https://bit.ly/knowledgediffdatacode.

[10] See Appendix B for an alternative calculation method.

Note that these equations represent the derivative with respect to a specific column of dummy values and not q_j or s_i itself. That is, implied in the formulation of the problem there are a set of columns for each rubric row (or student row) where only one of the columns has the value of one for any given observation (all else zero). This is why the marginal effect in the context of OLS is the value of β and not the value of the data in the specified column. In terms of notation, we will define $p_j\left(q_j, s_i\right)$ and $p_i\left(q_j, s_i\right)$ as these derivatives.

In the context of Table 2, Equation (5) is evaluated for each of the 161 estimated values of s_i and the value of q_j presented in the table. These values are then averaged. This procedure of evaluating a function on the subset of treated rows and taking the average of those evaluations will be used in calculating Average Conditional Probability (ACP) and Average Marginal Effect (AME). Thus, we will be less pedantic in describing those procedures.

4.2 | Project 1: Interpretations

It is the view of the authors that the Average Logistic and Average Marginal Logistic columns are the most useful. The Average Logistic shows the overall probability of trial failure given the estimated value while Average Marginal Logistic shows how that probability of failure changes given the estimated value; they are also the most comparable (in terms of how the practitioner might use them in interpretation) to the estimates of q_j and s_i when using the linear form of $p\left(q_j, s_i\right)$. In the context of rubric rows, this helps diagnose the difficulty of a rubric row. In the context of students, this helps evaluate student ability.

Consider the results for rubric rows one and seven presented in Table 2. Examining the Average Logistic and Average Marginal Logistic values for rubric row one suggests that the probability of trial failure (Average Logistic) is low (about 7%) and this rubric row decreases trial failure (Average Marginal Logistic) an average of 9%. In short, this rubric row is considerably easier for students to attain high marks. This particular rubric row tests basic knowledge on structuring a paper related to length, spacing, and font size (a low Bloom's area). In this context, it might be both intentional and desirable that the rubric row is easy for students.

Rubric row seven equivalently tests basic knowledge related to paper structuring; in this case, creating a properly formatted reference page. However, unlike rubric row one, the Average Logistic is 57% (Average Marginal Logistic is about 36%). This level of difficulty may be expected, and perhaps even desired, with complex subject areas. But, in this case it is in conflict with the relatively low Bloom's level. An instructor who was not previously aware of the difficulty students had with a particular rubric row may pose additional considerations for the next assignment or the next term: How is this content taught (if taught within the class)? Is this content being taught in a prerequisite class? Should the program integrate this content area into the program curriculum? In short, whether a given difficulty level is 'good' or 'bad' is context dependent. The results of the program, however, provide a statistically valid measure of row difficulty rather than simple threshold metrics or average scores.

TABLE 2 Estimates of the rubric parameters along with model fit information for Project 1 in the Labor Economics course.

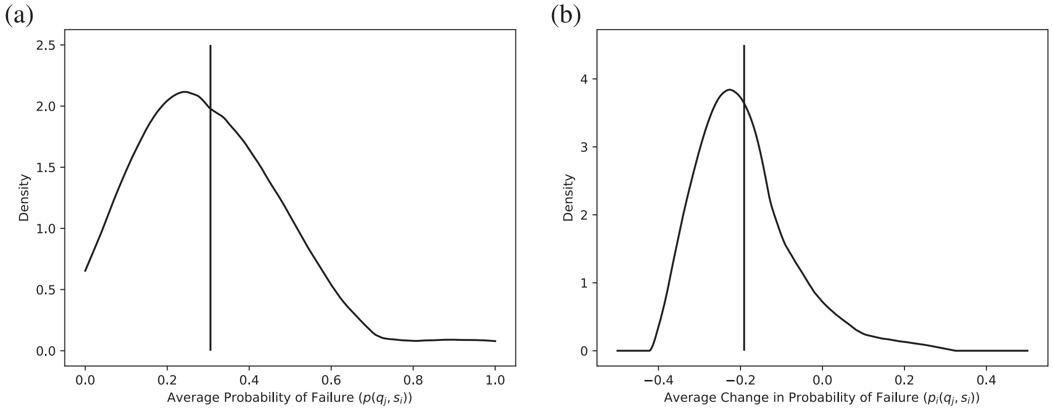
Rubric variable	*E.* value	Average Logistic	Average Marginal Logistic	95% CI	*p*-value
1	−1.724	0.071	−0.088	(−2.009, −1.472)	0.000
2	0.665	0.345	0.123	(0.489, 0.838)	0.000
3	1.050	0.420	0.209	(0.822, 1.268)	0.000
4	−0.141	0.214	−0.019	(−0.381, 0.081)	0.214
5	0.481	0.312	0.085	(0.293, 0.665)	0.000
6	0.741	0.360	0.140	(0.493, 0.976)	0.000
7	1.810	0.574	0.364	(1.576, 2.081)	0.000
8	−0.698	0.147	−0.071	(−0.959, −0.477)	0.000
Number of observations					1296[a]
Number of parameters					169
AIC					3138.259
BIC					4011.488
McFadden pseudo-R^2					0.223
Likelihood ratio test statistic					803.829
χ^2 LR test *p*-value					0.000

Note: One hundred and sixty-one students were in the dataset representing both the Fall 2021 and Spring 2022 semesters. The columns Average Logistic and Average Marginal Logistic show the average value of the logistic function, or derivative of the logistic function, for all treatment observations for the given rubric row. Average Logistic is the average probability of failure given the estimated value while Average Marginal Logistic represents the change in the average probability of failure given the estimated value. The confidence interval and *p*-value are calculated using a block bootstrapping (Künsch, 1989) routine treating each student as an independent block. Student variable estimates are not presented to keep the table a manageable size. These estimates are available upon request.

[a]The number of observations is not exactly equal to 161 × 8 as one student repeated the project as they took the class in both the fall and spring semester. We have not dropped these observations as the student has non-random characteristics. Thus, the best option is to keep the data in the dataset. Nonetheless, this is one student among 161; this does not have a substantive impact on the results.

4.3 | Project 1: Student ability distributions

In Table 2 we presented the rubric estimates but omitted the student estimates from the table. Instead, in Figure 3 we show the estimated student Average Logistic and Average Marginal Logistics as distributions. The distributions were generated using a Kernel Density Estimation (Scott, 2015) procedure. Here we can interpret the two estimates as the average probability of failure and the change in the average probability of failure, respectively. The distributions and averages presented here might not be that useful in isolation but would be of immense value as part of an assessment program designed to assess knowledge or learning. Assuming the same rubric and project was used in multiple cycles of the class, the instructor or program could compare these values across time. While there likely will be a focus on the mean values, it is also possible instructors or committees could focus on the distribution as a whole. For instance, some programs might be concerned by the spread



Probability of Failure

Change in Probability of Failure

FIGURE 3 Kernel Density Estimates (KDE) (Scott, 2015) of student estimates. (a) The Average Logistic given the student estimates. The mean value of 0.305 is plotted as a vertical line in (a). (b) The Average Marginal Logistic given the student estimates. The mean value of −0.191 is plotted as a vertical line in (b).

TABLE 3 The Average Conditional Probability (ACP) for each of the possible k values for a given rubric row.

Rubric variable	ACP $k=0$	ACP $k=1$	ACP $k=2$	ACP $k=3$	ACP $k=4$	ACP $k=5$
1	0.071	0.051	0.044	0.040	0.036	0.759
2	0.345	0.185	0.116	0.354		
3	0.420	0.199	0.113	0.268		
4	0.214	0.138	0.102	0.546		
5	0.312	0.176	0.115	0.397		
6	0.360	0.189	0.451			
7	0.574	0.201	0.224			
8	0.147	0.102	0.082	0.669		

Note: ACP values sum to 1 for any given row. The top box for any given rubric will contain a high percentage of the outcomes. As the outcome is censored, the top box is capturing unobserved successes in the right tail of the distribution.

of student ability as measured by extracting percentiles in the distribution or standard deviation. In Section 4.5 we will show how to use this procedure to estimate changes in student ability.

4.4 | Average Conditional Probability and Average Marginal Effect

While the authors believe the best way to interpret the MLE estimates is using the Average Logistic or Average Marginal Logistic transformations presented above, the PA software also present alternative calculations that might be more familiar to some practitioners. In Table 3 we present the Average Conditional Probabilities (ACP) for each of the rubric rows. This is calculated using a procedure similar to the procedures described to calculate Average Logistic or

TABLE 4 The Average Marginal Effect (AME) for each of the possible k values for a given rubric row.

Rubric variable	AME $k=0$	AME $k=1$	AME $k=2$	AME $k=3$	AME $k=4$	AME $k=5$
1	−0.088	−0.064	−0.053	−0.045	−0.038	0.287
2	0.123	0.030	0.000	−0.154		
3	0.209	0.027	−0.017	−0.219		
4	−0.019	−0.009	−0.004	0.033		
5	0.085	0.026	0.004	−0.114		
6	0.140	0.031	−0.171			
7	0.364	−0.037	−0.327			
8	−0.071	−0.043	−0.028	0.142		

Note: Like all AME values, each row sums to zero.

Average Marginal Logistic. However, the evaluated function is Equation (1). This procedure is repeated for each possible value of k. As this is evaluating all possible values in a PMF, the sum of the columns will sum to one.

Table 4 presents the Average Marginal Effect (AME) of each rubric row. These values are calculated in a similar manner to the ACP values presented above, but the evaluation function is the derivative of Equation (1). Specifically, the following equation is used in this calculation:

$$-p_{j|i}\left(q_j,s_i\right)\left(\left(1-p\left(q_j,s_i\right)\right)^{\lfloor k_{ij}\rfloor-1}\right)\left(\left(-\lfloor k_{ij}\rfloor-1\right)\left\lfloor\frac{k_{ij}}{b_j}\right\rfloor\left(p\left(q_j,\,s_i\right)\text{-}1\right)+\left(\lfloor k_{ij}\rfloor+1\right)p\left(q_j,s_i\right)-1\right), \tag{7}$$

where $p_{j|i}\left(q_j,s_i\right)$ is either Equation (5) or (6) depending on if the AMEs are being determined for a rubric row or student. Like all AMEs the values will sum to zero.

When calculating ACP or AME values for the student estimates it is possible that some of the subset of observations (as described in the calculation procedure) will have different maximum bins. In these cases, the minimum of the maximum bins will be used to calculate the ACP and AME values and all observations in that calculation will be treated *as if* that minimum of the maximum bin is the maximum bin for all observations in the subset. The net result is that the ACP values and AME values will behave as expected (summing to one or zero, respectively) but a high percentage of the outcomes can be represented in the top bin.

While we have presented the ACP and AME values as separate values in this paper, the PA software outputs the information in tables 2-4 as a single data frame; a similar data frame is produced for the student estimates. Finally, the fit estimates presented at the bottom of Table 2 are placed in their own data frame as they do not belong in the same columns.

4.5 | Comparing Project 1 to Project 4

The Labor Economics course included multiple projects throughout the semester using similar rubrics; Project 1 occurred near the beginning of the semester while Project 4 occurred near the

TABLE 5 Estimates of the rubric parameters along with model fit information for projects 1 and 4 in the Labor Economics course.

Rubric variable	*E.* value	Average Logistic	Average Marginal Logistic	95% CI	*p*-value
1	−1.914	0.055	−0.081	(−2.244, −1.653)	0.000
6	0.829	0.330	0.134	(0.642, 1.010)	0.000
7	2.464	0.612	0.399	(2.203, 2.756)	0.000
Number of observations					857[a]
Number of parameters					285
AIC					1913.404
BIC					3268.134
McFadden pseudo-R^2					0.418
Likelihood ratio test statistic					964.388
χ^2 LR test *p*-value					0.000

Note: One hundred and forty-one students were in the dataset representing both the Fall 2021 and Spring 2022 semesters. Here we examine students who completed both Project 1 and Project 4. Rubric rows that were testing the same ability characteristic were extracted from the dataset (rows 1, 6, and 7). Moreover, the proxies for student ability were estimated for each student at the time of Project 1 and at the time of Project 4. These matched pairs were then compared. The columns Average Logistic and Average Marginal Logistic show the average value of the logistic function, or derivative of the logistic function, for all treatment observations for the given rubric row. In simple terms, Average Logistic is the average probability of failure given the estimated value while Average Marginal Logistic represents the change in the average probability of failure given the estimated value. The confidence interval and *p*-value are calculated using a block bootstrapping (Künsch, 1989) routine treating each student as an independent block. Student variable estimates are not presented to keep the table a manageable size. These estimates are available upon request.

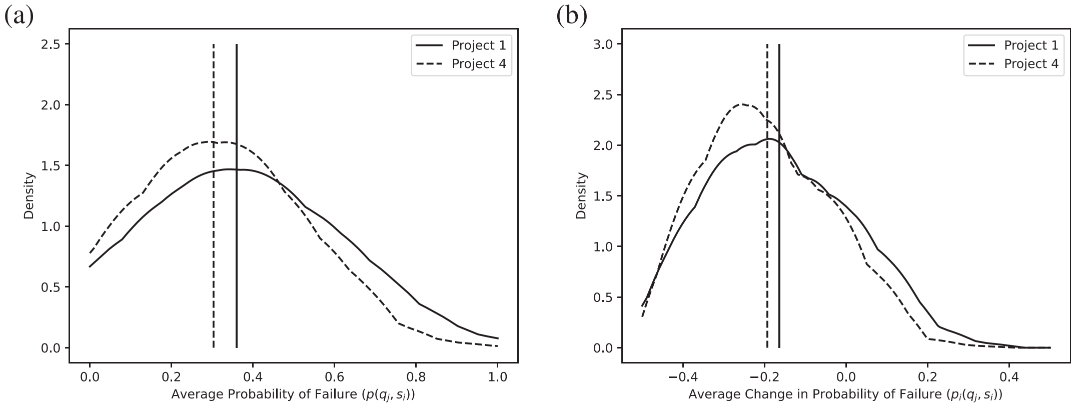
[a] The observation count is not exactly equal to 141 × 2 × 3 because three students repeated parts of Project 4 due to repeating the class. Like the singular student who repeated Project 1, we decided the most unbiased option was to keep these observations. Nonetheless, they do not have a substantive impact on the results.

end. In this section we will explore if the students' writing ability increased during the semester.

Rubric rows 1, 6, and 7 were identical on projects 1 and 4. These rows related to writing ability and not the specifics of the project. Therefore, we extracted these rows from both datasets. For the student variables in Project 4, we labeled each student with a different number from in Project 1. The student estimates assume that the ability of the student is the same. However, these students are being measured at two time periods. Their ability might have increased and we want to measure that increase. In total, we have 141 students who completed both Project 1 and Project 4.

Table 5 shows the rubric estimates and estimated fit parameters. Like the estimates in Section 4.1, we can interpret each estimate using Average Logistic or Average Marginal Logistic values. However, for our purposes in comparing student ability, we are more interested in the proxies for student ability.

Like Figure 3, Figure 4 shows the KDEs of the Average Logistic and Average Marginal Logistic distributions. However, Figure 4 shows two KDEs per panel: one for Project 1, and one for Project 4. The differences in the Average Logistic results for projects 1 and 4 are statistically tested using both a *t*-test and a Mann–Whitney (Mann & Whitney, 1947) test. By this measure, the students writing ability increased throughout the semester by a statistically significant



Probability of Failure

Change in Probability of Failure

FIGURE 4 Kernel Density Estimates (Scott, 2015) of student Average Logistic estimates for projects 1 and 4. Matched rubric rows of 1, 6, and 7 were used. (a) The Average Logistic given the student estimates. Vertical lines are plotted at the mean of each distribution: 0.360 and 0.303 for projects 1 and 4, respectively. This difference is statistically significant using both a *t*-test (*p*-value = 0.008) and a Mann-Whitney (Mann & Whitney (1947)) test (*p*-value =0.026). (b) The Average Marginal Logistic given the student estimates. Vertical lines are plotted at the mean of each distribution: −0.163 and −0.193 for projects 1 and 4, respectively. This difference is not statistically significant using a *t*-test (*p*-value = 0.058) or a Mann–Whitney test (*p*-value = 0.179).

amount; the difference in the Average Marginal Logistic results was not statistically significant. This can be seen in Figure 4 by the leftward movement of the KDEs. Because the measure is of a failed trial, a leftward movement of the KDE for Project 4 implies that fewer students are failing to perform in the selected subject areas.

5 | CONCLUSION

In this paper we provide the researcher, educator, department, or academic administrator a statistical understanding of instruments measured with a rubric. This new understanding can be used by instructors to improve their class, departments to improve their programs, and institutions to improve their core curriculum. To allow the maximum number of people to adopt these new techniques, we provide software that can be used in a standard Python environment on the user's computer or in a web-based Python environment.

For instance, a researcher could test the effectiveness of an educational treatment by randomly assigning the students to treatment and control groups. The students would then be graded using a rubric-based instrument. The distributions of the proxies of student ability could be compared in a similar fashion to what is presented in Section 4.5. A similar approach could be used at the department or university level. For instance, suppose a university emphasizes the importance of writing and thus requires all undergraduate students to complete a specific essay in a testing environment to graduate.[11] These essays are then graded using a rubric and results are stored for analysis. Multiple years of data could be analyzed and it could be determined if the writing ability of graduating students changed over time; a similar approach

[11] This specific example was selected as such a procedure existed at one of the authors' undergraduate institutions.

could be used for end-of-program assessment by a specific department, the scale would simply be smaller.

Individual instructors can adopt the methods presented in this paper to both measure proxies of student ability and diagnose issues with their course or rubric. For instance, an instructor introducing a new rubric-based assignment into their class might be particularly interested to learn if rubric-row difficulty aligns with Bloom's level of the rubric row. The procedure the instructor would follow is highlighted in Section 4.2. While the example presented in this paper uses a very large dataset, the techniques presented here can be used with small classes. Some of the test datasets used when developing the software (Section 3) had as few as five students with three rubric rows. While the authors do not recommend using a dataset that small, this demonstrates that the estimation procedure can work with normal-sized classes.

We suspect many adopters of this procedure will be individual instructors interested in diagnosing issues with their assessments. This can include diagnosing rubric row difficulty or examining the distribution of student ability. However, the impact of these methods might be greatest when used in academic assessment. Academic assessment is often not taken seriously by serious educators. Anecdotally, we believe this is because academic assessment is not typically treated as a scientific endeavor to measure knowledge or learning. Instead, assessment often rests on convenient, but borderline meaningless, measures. This paper demonstrates that assessment does not have to be done this way. As highlighted in the introduction, there are existing statistically-valid methods of measuring knowledge or learning when using multiple-choice or single-response exams. This paper demonstrates how rubric-based assessments can be used in a systematic way. Therefore, scientific approaches to measuring knowledge or learning are available regardless of the assessment method. This allows faculty, units, and institutions to make adjustments to their classes or programs based on better information. It is the opinion of the authors that this will result in more faculty buy-in as the measures will have been generated through a scientific approach.

ACKNOWLEDGMENTS

The authors thank Charles Courtemanche and the two anonymous referees for their helpful feedback. Further, we thank Savannah Adkins and the other participants at the 2023 CTREE conference.

ORCID

Ben O. Smith https://orcid.org/0000-0003-1286-0852
Jadrian J. Wooten https://orcid.org/0000-0002-7838-9349

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

De Ayala, R.J. (2013) *The theory and practice of item response theory*. New York, NY: Guilford Publications.

Ebel, R.L. (1954) Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, 14(2), 352–364.

Hake, R.R. (1998) Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74.

Horn, R.A. & Johnson, C.R. (2013) *Matrix analysis*. New York, NY: Cambridge University Press.

Kuder, G.F. & Richardson, M.W. (1937) The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151–160.

Künsch, H.R. (1989) The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3), 1217–1241.

Lord, F.M. (1980) *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

Mann, H.B. & Whitney, D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50–60.

McFadden, D. (1979) *Quantitative methods for analysing travel behaviour of individuals: some recent developments, in 'Behavioural travel modelling'*. New York, NY: Routledge, pp. 279–318.

Mori, M., Woolson, R. F. & Woodworth, G. G. (1994) *Slope estimation in the presence of informative right censoring: modeling the number of observations as a geometric random variable*. Biometrics, pp. 39–50.

Okada, K., Vandekerckhove, J. & Lee, M. D. (2018) Modeling when people quit: bayesian cen- sored geometric models with hierarchical and latent-mixture extensions. *Behavior Research Methods*, 50(1), 406–415.

Ozgur, C., Colliau, T., Rogers, G. & Hughes, Z. (2021) Matlab vs. Python vs. R. *Journal of Data Science*, 15(3), 355–372.

Patel, M. & Gajjar, K. A. (2010) Inference under progressive interval censoring for geometric distri- bution. *International Journal of Statistics and Economics*, 5, 21–36.

Powell, M.J.D. (1964) An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2), 155–162.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (2007) *Numerical recipe: the art of scientific computing*. New York, NY: Cambridge University Press.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

Scott, D.W. (2015) *Multivariate density estimation: theory, practice, and visualization*. Hoboken, NJ: John Wiley & Sons.

Smith, B.O. (2022) Assessment disaggregation: a new tool to calculate learning types from nearly any exam platform, including online systems. *The Journal of Economic Education*, 53(2), 194–195.

Smith, B.O. & Wagner, J. (2018) Adjusting for guessing and applying a statistical test to the disaggregation of value-added learning scores. *The Journal of Economic Education*, 49(4), 307–323.

Smith, B.O. & White, D.R. (2021) On guessing: an alternative adjusted positive learning estimator and comparing probability misspecification with Monte Carlo simulations. *Applied Psychological Measurement*, 45(6), 441–458.

Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24– 36.

Traub, R.E. (1997) Classical test theory in historical perspective. *Educational Measurement: Issues and Practices*, 16(4), 8–14.

Uto, M. (2021) A multidimensional generalized many-facet rasch model for rubric-based performance assessment. *Behaviormetrika*, 48(2), 425–457.

Walstad, W.B. & Wagner, J. (2016) The disaggregation of value-added test scores to assess learning outcomes in economics courses. *The Journal of Economic Education*, 47(2), 121–131.

Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62.

How to cite this article: Smith, B. O., & Wooten, J. J. (2023). Assessing proxies of knowledge and difficulty with rubric-based instruments. *Southern Economic Journal*, 1–25. https://doi.org/10.1002/soej.12658

APPENDIX A

A.1 | Additional Project 1 estimates

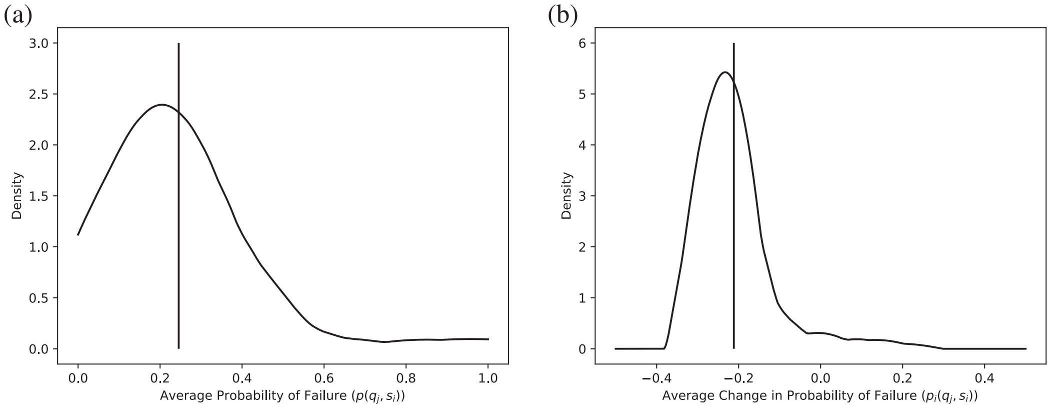
In the main body of the paper, we estimate a pooled model containing students from both the Fall 2021 and Spring 2022 sections of the Labor Economics class. For completeness, we have included the estimates of the individual sections in this appendix. Like the pooled model, we include the rubric row estimates along with KDEs of the student results.

A.1.1 | Fall 2021 Project 1 estimates

TABLE A1 Estimates of the rubric parameters along with model fit information for Project 1 in the Labor Economics course.

Rubric variable	*E.* value	Average Logistic	Average Marginal Logistic	95% CI	*p*-value
1	−1.681	0.067	−0.068	(−2.317, −1.241)	0.000
2	0.587	0.286	0.099	(0.299, 0.841)	0.000
3	0.811	0.325	0.149	(0.474, 1.167)	0.000
4	−0.153	0.180	−0.018	(−0.561, 0.206)	0.404
5	0.612	0.290	0.105	(0.358, 0.873)	0.000
6	0.817	0.326	0.150	(0.4033, 1.170)	0.000
7	0.918	0.345	0.174	(0.512, 1.293)	0.000
8	−0.472	0.145	−0.045	(−0.987, −0.131)	0.004
Number of observations					496
Number of parameters					70
AIC					1213.260
BIC					1507.720
McFadden pseudo-R^2					0.181
Likelihood ratio test statistic					237.049
χ^2 LR Test *p*-value					0.000

Note: Sixty-two students were in the dataset representing the Fall 2021 semester. The columns Average Logistic and Average Marginal Logistic show the average value of the logistic function, or derivative of the logistic function, for all treatment observations for the given rubric row. Average Logistic is the average probability of failure given the estimated value while Average Marginal Logistic represents the change in the average probability of failure given the estimated value. The confidence interval and *p*-value are calculated using a block bootstrapping (Künsch, 1989) routine treating each student as an independent block. Student variable estimates are not presented to keep the table a manageable size. These estimates are available upon request.



Probability of Failure

Change in Probability of Failure

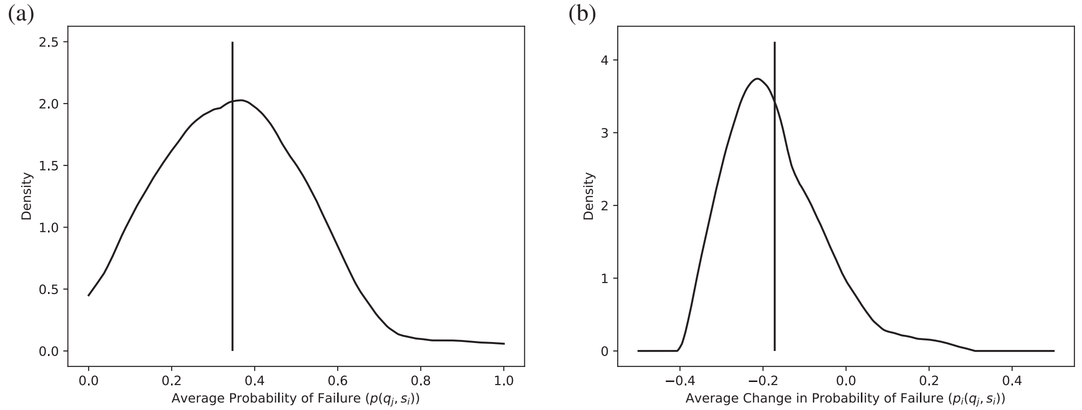
FIGURE A1 Kernel Density Estimates (KDE) (Scott, 2015) of student estimates using the Fall 2021 data. (a) The Average Logistic given the student estimates. The mean value of 0.245 is plotted as a vertical line in (a). (b) The Average Marginal Logistic given the student estimates. The mean value of −0.212 is plotted as a vertical line in (b).

A.1.2 | Spring 2022 Project 1 estimates

TABLE A2 This table presents estimates of the rubric parameters along with model fit information for Project 1 in the Labor Economics course.

Rubric variable	*E.* value	Average Logistic	Average Marginal Logistic	95% CI	*p*-value
1	−1.736	0.073	−0.100	(−2.098, −1.460)	0.000
2	0.713	0.383	0.139	(0.474, 0.942)	0.000
3	1.215	0.483	0.246	(0.958, 1.529)	0.000
4	−0.130	0.235	−0.020	(−0.440, 0.158)	0.384
5	0.396	0.323	0.072	(0.177, 0.613)	0.000
6	0.698	0.380	0.135	(0.367, 0.981)	0.000
7	2.620	0.745	0.409	(2.316, 3.006)	0.000
8	−0.804	0.148	−0.086	(−1.121, −0.533)	0.000
Number of observations					800
Number of parameters					108
AIC					1909.012
BIC					2414.950
McFadden pseudo-R^2					0.256
Likelihood ratio test statistic					582.905
χ^2 LR test *p*-value					0.000

Note: One hundred students were in the dataset representing the Spring 2022 semester. The columns Average Logistic and Average Marginal Logistic show the average value of the logistic function, or derivative of the logistic function, for all treatment observations for the given rubric row. Average Logistic is the average probability of failure given the estimated value while Average Marginal Logistic represents the change in the average probability of failure given the estimated value. The confidence interval and *p*-value are calculated using a block bootstrapping (Künsch, 1989) routine treating each student as an independent block. Student variable estimates are not presented to keep the table a manageable size. These estimates are available upon request.



Probability of Failure

Change in Probability of Failure

FIGURE A2 Kernel Density Estimates (KDE) (Scott, 2015) of student estimates using the Spring 2022 data. (a) The Average Logistic given the student estimates. The mean value of 0.347 is plotted as a vertical line in (a). (b) The Average Marginal Logistic given the student estimates. The mean value of −0.172 is plotted as a vertical line (b).

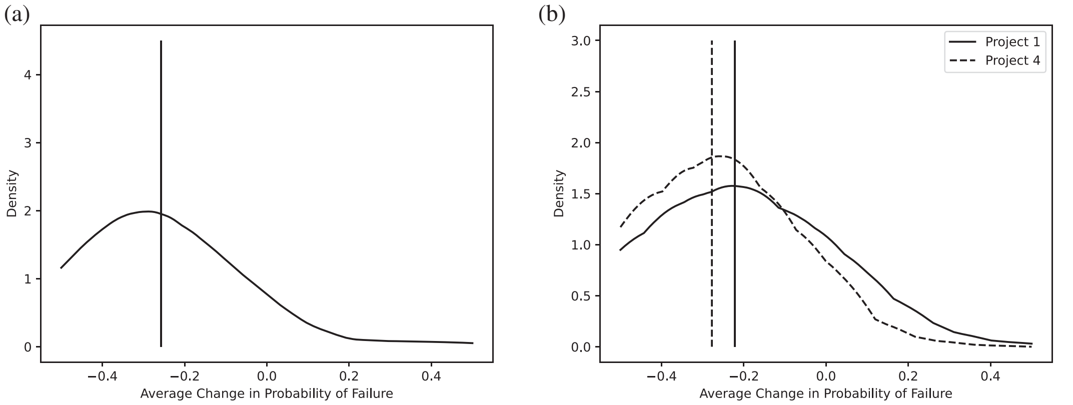
APPENDIX B: DISCRETE AVERAGE MARGINAL LOGISTIC ESTIMATIONS

In the main body of the paper, we describe the procedure to calculate the Average Marginal Logistic. As part of that procedure, the derivative of the logistic function is used to calculate the value for a given row of data in the average. However, in the case of dummy variables, it is common to calculate the marginal effect by evaluating the function with the variable in question set to 1 and 0 then taking the difference. For completeness, we have included this alternative method of calculating the Average Marginal Logistic in our software. In most cases, this Discrete Average Marginal Logistic is very similar to Average Marginal Logistic (Table B1).

Values will vary between the two calculation methods when the estimated value is in the tails. For instance, one of the student estimates in the Project 1 sample is 78.849. For this student estimate, the average logistic value is 1.000 and the Average Marginal Logistic is 0.000. The estimated value is so far into the tail that the instantaneous slope of the logistic curve is effectively zero at that point. The Discrete Average Marginal Logistic, however, is 0.437 as it uses the average slope instead of the instantaneous slope. This difference in calculation method can be seen in the distributions of student ability (Figure B1).

TABLE B1 A comparison of the Average Marginal Logistic and Discrete Average Marginal Logistic from Project 1.

Rubric variable	*E.* value	Average Marginal Logistic	Discrete Average Marginal Logistic
1	−1.724	−0.088	−0.163
2	0.665	0.123	0.111
3	1.050	0.209	0.186
4	−0.141	−0.019	−0.020
5	0.481	0.085	0.078
6	0.741	0.140	0.126
7	1.810	0.364	0.340
8	−0.698	−0.071	−0.087



Change in Probability of Failure (Project 1)

Change in Probability of Failure (Project 1 & 4)

FIGURE B1 Kernel Density Estimates (Scott, 2015) of student Discrete Average Marginal Logistic estimates for projects 1 and 4. (a) The Discrete Average Marginal Logistic distribution from the Project 1 estimate (Section 4.1). (b) The Discrete Average Marginal Logistic distribution from the writing rubric row comparison in Section 4.5. Vertical lines are plotted at the mean of each distribution. In (a) the mean value is −0.258. In (b) the means are −0.222 and −0.278 for Project 1 and Project 4, respectfully. This difference is statistically significant using a *t*-test (p-value = 0.009) and a Mann-Whitney (Mann & Whitney (1947)) test (p-value = 0.031).

APPENDIX C: SECOND AND CROSS-PARTIAL DERIVATIVES OF THE LOG-LIKELIHOOD FUNCTION

The second derivatives with respect to a specific q_j for both of the suggested functional forms of $p\left(q_j, s_i\right)$ are provided below. Equation (C1) is the second derivative when $p\left(q_j, s_i\right) = q_j + s_i$ and Equation (C2) is the second derivative when $p\left(q_j, s_i\right) = 1/\left(1 + e^{-\left(q_j + s_i\right)}\right)$.

$$\frac{\partial^2 \mathrm{Log}(L)}{\partial \hat{q}_j^2} = \sum_{i=1}^{n} -\lfloor k_{ij} \rfloor \frac{1}{\left(1-\hat{q}_j - s_i\right)^2} - \frac{\left(1+\left\lceil \frac{-k_{ij}}{b_j} \right\rceil\right)^2}{\left(\hat{q}_j + s_i + \left(\hat{q}_j + s_i - 1\right)\left\lceil \frac{-k_{ij}}{b_j} \right\rceil\right)^2}, \tag{C1}$$

$$\frac{\partial^2 \mathrm{Log}(L)}{\partial \hat{q}_j^2} = \sum_{i=1}^{n} \frac{e^{\hat{q}_j + s_i}\left(\frac{\overbrace{\left(\left\lfloor \frac{k_{ij}}{b_j} \right\rfloor - 1\right)}^{\text{Term 1}}\left(e^{2\left(\hat{q}_j + s_i\right)} - \left\lfloor \frac{k_{ij}}{b_j} \right\rfloor\right)}{\left(\left\lfloor \frac{k_{ij}}{b_j} \right\rfloor + e^{\hat{q}_j + s_i}\right)^2} - \lfloor k_{ij} \rfloor\right)}{\left(e^{\hat{q}_j + s_i} + 1\right)^2}. \tag{C2}$$

Both are negative. This is easy to see with Equation (C1). It is less obvious with Equation (C2). However, remember that $\left\lfloor \frac{k_{ij}}{b_j} \right\rfloor$ is either zero or one. When $\left\lfloor \frac{k_{ij}}{b_j} \right\rfloor = 0$, Term 1 above becomes negative one; this is multiplied by the positive $e^{2\left(\hat{q}_j + s_i\right)}$. If $\left\lfloor \frac{k_{ij}}{b_j} \right\rfloor = 1$ then Term 1 equals zero and the numerator simplifies to $-e^{\hat{q}_j + s_i}\lfloor k_{ij} \rfloor$.

The cross-partial derivatives with respect to a specific s_i and q_j are provided below. Equation (C3) is the cross-partial derivative when $p\left(q_j, s_i\right) = q_j + s_i$ and Equation (C4) is the cross-partial derivative when $p\left(q_j, s_i\right) = 1/\left(1+e^{-\left(q_j + s_i\right)}\right)$.

$$\frac{\partial^2 \mathrm{Log}(L)}{\partial \hat{q}_j \hat{s}_i} = -\lfloor k_{ij} \rfloor \frac{1}{\left(1-\hat{q}_j - \hat{s}_i\right)^2} - \frac{\left(1+\left\lceil \frac{-k_{ij}}{b_j} \right\rceil\right)^2}{\left(\hat{q}_j + \hat{s}_i + \left(\hat{q}_j + \hat{s}_i - 1\right)\left\lceil \frac{-k_{ij}}{b_j} \right\rceil\right)^2}, \tag{C3}$$

$$\frac{\partial^2 \mathrm{Log}(L)}{\partial \hat{q}_j \hat{s}_i} = \frac{e^{\hat{q}_j + \hat{s}_i}\left(\frac{\left(\left\lfloor \frac{k_{ij}}{b_j} \right\rfloor - 1\right)\left(e^{2\left(\hat{q}_j + \hat{s}_i\right)} - \left\lfloor \frac{k_{ij}}{b_j} \right\rfloor\right)}{\left(\left\lfloor \frac{k_{ij}}{b_j} \right\rfloor + e^{\hat{q}_j + \hat{s}_i}\right)^2} - \lfloor k_{ij} \rfloor\right)}{\left(e^{\hat{q}_j + \hat{s}_i} + 1\right)^2}. \tag{C4}$$

Notably, equations (C3) and (C4) are the inner portion of the summations in equations (C1) and (C2). Thus, $\frac{\partial^2 \mathrm{Log}(L)}{\partial \hat{q}_j^2}$ is equal to the sum of the cross-partial derivatives. For instance, suppose there was two students and two rubric rows. The second derivative matrix could be expressed as follows:

$$M = \begin{array}{ccccc} & s_1 & s_2 & q_1 & q_2 \\ s_1 & \frac{\partial^2 Log(L)}{\partial q_1 s_1} + \frac{\partial^2 Log(L)}{\partial q_2 s_1} & 0 & \frac{\partial^2 Log(L)}{\partial q_1 s_1} & \frac{\partial^2 Log(L)}{\partial q_2 s_1} \\ s_2 & 0 & \frac{\partial^2 Log(L)}{\partial q_1 s_2} + \frac{\partial^2 Log(L)}{\partial q_2 s_2} & \frac{\partial^2 Log(L)}{\partial q_1 s_2} & \frac{\partial^2 Log(L)}{\partial q_2 s_2} \\ q_1 & \frac{\partial^2 Log(L)}{\partial q_1 s_1} & \frac{\partial^2 Log(L)}{\partial q_1 s_2} & \frac{\partial^2 Log(L)}{\partial q_1 s_1} + \frac{\partial^2 Log(L)}{\partial q_1 s_2} & 0 \\ q_2 & \frac{\partial^2 Log(L)}{\partial q_2 s_1} & \frac{\partial^2 Log(L)}{\partial q_2 s_2} & 0 & \frac{\partial^2 Log(L)}{\partial q_2 s_1} + \frac{\partial^2 Log(L)}{\partial q_2 s_2} \end{array}$$

It can be shown a matrix following this structure (n students and m rubric rows) is negative semidefinite (or the test is inconclusive). One can see that $-1M$ is a diagonally dominant matrix. Thus, as a symmetric, diagonally dominant, matrix $-1M$ is positive semidefinite (Horn & Johnson, 2013, see Theorem 6.1.10 on page 392). Therefore, M is negative semidefinite (or the test is inconclusive). The inconclusive case is not problematic for this application, as the numerical optimization algorithm examines the function curvature as part of the routine. Note that the modified Powell (Powell, 1964; Press et al., 2007) algorithm uses the objective function directly and does not calculate or use derivatives to find an optimum.

APPENDIX D: EXAMPLE USAGE: PROJECT 1 RUBRIC

The following rubric was used to grade Project 1. The formatting of the rubric was altered to fit into the structure of the paper. However, the rubric boxes are identical to the version used in Canvas used for grading.

D.1. | Structure of paper

Description: No more than 3 pages double-spaced. Reference page and Dashboard image are included in appendix.

No Grade: No paper was submitted.	*Unacceptable*: Most items are poorly done.	*Poor*: 3 or more items are missing or not well done.	*Acceptable*: 2 items are missing or not well done.	*Good*: Something is missing or not well done.	*Perfect*: All items are included and done well.

D.2. | Measure descriptions

Description: Measures associated with the paper are well-defined and accurate.

No Grade: None of the measures for the project were defined in the analysis paper.	*Poor*: Multiple measures were not defined well or were missing.	*Good*: At least one measure could have been better defined/ described in the paper.	*Perfect*: There are no areas for improvement in the describing/ defining variables.

D.3. | Class connection

Description: There are clear connections to class. Any relevant question in the paper is answered and uses class material to form basis of answer.

No Grade: There is no connection to the material covered in class.	*Poor*: There seems to be very little class connection or questions aren't answered from the project.	*Good*: The connections to class are weak or there are questions in the project that went unanswered. Small improvements could be made.	*Perfect*: The writer has clearly connected the material in the paper to discussions in class or references examples from class. There are no areas for improvement.

D.4. | Unemployment

Description: Analyzes issues with unemployment rate calculation in terms of classification.

No Grade: No mention of issues or calculation at all.	*Poor*: Poor discussion of unemployment calculation and classification issues.	*Good*: Described, but classification issues could be better outlined.	*Full Marks*: Well describe and classification issues are clearly outlined

D.5. | Article(s) and data summaries

Description: Summarizes article(s) listed in project instructions. Summarizes data files used for visualizations.

No Grade: No summary of article(s) or datasets.	*Poor*: Multiple items missing from summaries.	*Above average*: All of the article(s) and data are mentioned, but the summaries could be improved with little additional effort.	*Perfect*: All article(s) and data files are summarized. It is clear the writer read the material and accurately described each item.

D.6. | In-text citations

Description: Parenthetical citations are accurate and well done.

No Grade: There are no in-text citations in the analysis paper.	*Poor*: A lot of missing in-text citations.	*Perfect*: All are accurate and well-done.

D.7. | Reference page

Description: Citations are in APA format, correctly listed, and professional.

No Grade: References exist, but there are multiple problems with their references. The student should consider seeking help from the Writing Center.	*Above Average*: One or two references are entered incorrectly or are missing from the reference sheet.	*Perfect*: Every reference is completed correctly with full information and in the correct order. There are no clear issues with the references.

D.8. | Overall assessment

Description: The paper is well-written, free of grammatical and spelling errors. The material is presented professionally and is easy to read.

Unacceptable: The paper was difficult to read and did not entirely flow well. There were numerous issues throughout the paper that made it difficult to complete.	*Adequate*: The paper is understandable, but at times hard to read. There are multiple spelling and/or grammar mistakes throughout the paper. The paper was not easy to read.	*Above average*: There were portions of the paper that were difficult to read or there were consistent spelling/grammar issues in the paper. The paper is well written, but not entirely smooth.	*Perfect*: There are no noticeable problems while reading the paper. The paper was easy to read and flowed well.