10-15-2018

Adjusting for guessing and applying a statistical test to the disaggregation of value-added learning scores

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

Jamie Wagner
University of Nebraska at Omaha

Adjusting for guessing and applying a statistical test to the disaggregation of value-added learning scores*

Ben O. Smith†[1] and Jamie Wagner[1]

[1]College of Business Administration, University of Nebraska at Omaha

Abstract

In 2016, Walstad and Wagner developed a procedure to split pre- and post-test responses into four learning types: positive, negative, retained, and zero learning. This disaggregation is not only useful in academic studies, it also provides valuable insight to the practitioner: an instructor would take different mitigating actions in response to zero versus negative learning. However, the original disaggregation is sensitive to student guessing. This paper extends the original work by accounting for guessing and provides adjusted estimators using the existing disaggregated values. Further, Monte Carlo simulations of the adjusted learning type estimates are provided. Under certain assumptions, an instructor can determine if a difference in positive (or negative) learning is the result of a true change in learning or 'white noise.'

Keywords: Difference score, learning disaggregation, guessing, simulation

JEL classification: A20, A22, A23, C63

*The authors thank Bill Goffe, Matt Rousu, Bill Walstad, Dustin White, Mark Wohar, and the four anonymous referees for helpful suggestions on this project. We further thank Halli Tripe for editorial suggestions.

†Corresponding author: bosmith@unomaha.edu.

Traditionally, there have been two main ways of measuring learning within a course: as a 'stock' of knowledge or as 'value-added' (or flow) of knowledge (Siegfried and Fels 1979, p. 929). The stock of knowledge is how much students know at a given time, while the flow measures the change in the stock over some period of time. Often the stock of knowledge is measured using an exam while the flow of knowledge is measured using either the difference in a pre- and post-test or the pre-test as a control variable in explaining a post-test score. The difference in test scores is more applicable to practitioners and thus will be the definition used in this manuscript.

Walstad and Wagner (2016) suggest that the pre- and post-test score difference is insufficient in determining the amount of learning that occurred in the classroom. Thus, they propose disaggregating the answers to create new learning variables: positive learning ($\hat{\text{pl}}$), retained learning ($\hat{\text{rl}}$), negative learning ($\hat{\text{nl}}$), and zero learning ($\hat{\text{zl}}$). The insight from Walstad and Wagner is that different types of learning have different meanings to a researcher and require different mitigating responses from an instructor.

In the short amount of time the paper has been available, the method has been used in two published studies (Emerson and English 2016; Happ, Zlatkin-Troitschanskaia, and Schmidt 2016) and others are in progress. More importantly, the paper has encouraged instructors to disaggregate their classroom pre- and post-test scores to help diagnose problems and make pedagogical improvements.

This current paper extends the disaggregation by accounting for guessing in the context of multiple choice exams. A student's performance on a given exam is an estimate of their stock of knowledge at the time they took the exam. A student could have answered a question correctly because they knew the answer or because they were lucky enough to guess correctly. Therefore, the disaggregation technique proposed by Walstad and Wagner (2016) is a disaggregation of *performance* which are estimates of underlying learning. We show in this paper that these estimates should be adjusted to account for the expected number of

correct answers due to guessing.[1] Using a simulation, we provide the practitioner a tool to distinguish true changes in learning from statistical noise. The technique presented in this paper is useful to a researcher investigating a pedagogical technique, or an exogenous phenomena's impact on one of the learning types. More importantly, the more accurate measurements of learning presented in this manuscript will help instructors, departments, and assessment committees make better decisions.

The paper will proceed as follows. In *Guessing-Adjustment*, we will determine the expected value of the unadjusted disaggregated learning types and present adjusted estimators. In *Monte Carlo Simulation of the Counterfactual*, we will simulate the distributions of both the adjusted and unadjusted measures under assumptions of no learning. This results in a statistical test given that the underlying data generating assumptions are correct (tables in the appendix). In *Applications*, we apply our modified technique to assessment questions and a commonly used nationally-normed exam. We will then conclude the paper.

Guessing-Adjustment

The learning type disaggregation (Walstad and Wagner 2016) separates pre- and post-test scores into one of four learning types: positive ($\hat{\text{pl}}$), negative ($\hat{\text{nl}}$), zero ($\hat{\text{zl}}$), and retained learning ($\hat{\text{rl}}$)[2]. A student is said to have positive learning if they missed the question on the pre-test then correctly answered the question on the post-test. Negative learning is said to occur when the student answered a question correctly on the pre-test but incorrectly on the post-test. Zero learning is when the student answered a question incorrectly on both the pre- and post-test. Finally, retained learning is said to occur when the student answered a question correctly on both the pre- and post-test.

These item level measurements are aggregated either to the question or the student. Aggregated at the question level, these measurements represent the proportion of students by learning type. For example, an instructor can use the disaggregation technique to find

that a question on his/her exam has 40 percent positive ($\hat{\text{pl}}$), 3 percent negative ($\hat{\text{nl}}$), 20 percent zero ($\hat{\text{zl}}$), and 37 percent retained learning ($\hat{\text{rl}}$). When aggregated at the student level, they represent the proportion of questions by learning type for that specific student.

These aggregated measures have specific properties. First, $\hat{\text{pl}} + \hat{\text{rl}}$ equals the post-test score. Similarly, $\hat{\text{nl}} + \hat{\text{rl}}$ equals the pre-test score. Therefore, the flow of knowledge is $\hat{\text{pl}} - \hat{\text{nl}}$. Finally, the sum of the four learning types equals one. For a more thorough description of the learning types and their properties see Walstad and Wagner (2016, p. 123).

Notably, this disaggregation of performance includes students who guessed. In this section, we find the underlying learning parameters in the expectation when we assume a particular probability of guessing correct. For the purposes of this paper, we will define the probability of guessing correctly as the probability that a student selects the correct answer on a multiple choice exam question given that they do not know the answer. We assume this probability is independent of the student's ability level.

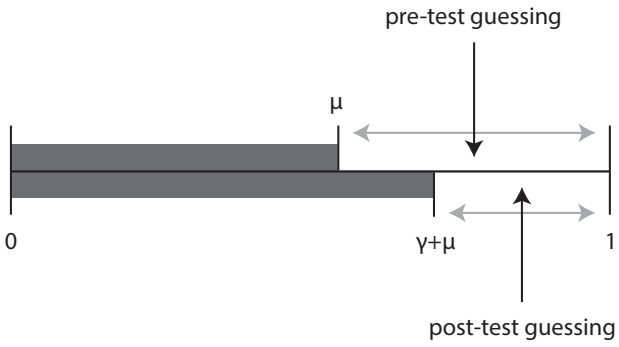


Figure 1: Diagram of the proportion of guessing students on a question on the pre- and post-test. μ is the proportion of students who know the answer at the time of the pre-test. γ is the proportion of students that learned the material by the time of the post-test.

As shown in Figure 1, the proportion of the class guessing on a question, or the proportion of questions guessed by a student, varies between the two exams with a larger proportion of guessing on the pre-test than the post-test. We see that μ proportion of the students knew the material when they entered the class and thus did not guess on either the pre- or

post-test (for the sake of this explanation, assume zero negative learning). An additional γ proportion of students learned the material during the course and thus did not guess on the post-test. But, because the number of students who are guessing is greater on the pre-test than the post-test, the *proportion* of correct answers due to guessing is greater on the pre-test than the post-test even if the probability of successfully guessing the answer is the same.[3] Therefore, subtracting the pre-test from the post-test does not find γ.

In determining the expected value of each learning type, consider a single exam question taken by an infinite number of students. We could consider a single student answering an infinite number of questions on an exam and the math would remain the same. For the sake of brevity, we will describe the parameters in terms of proportions of students knowing or not knowing an answer to a single question for the remainder of the paper. However, these proportions can be restated as the proportions of questions that a student knows or does not know.

Using similar notation to the diagram, we will define μ as the proportion of the students who know the answer to the question when taking the pre-test. γ is the proportion of the students who did not know the material at the time of the pre-test but learned the material by the post-test. Therefore, $1 - \mu - \gamma$ students never learned the material. A proportion of students, α, forgot the material at some point after the pre-test and before the post-test.

As a practical matter, we will make a few assumptions regarding the probability of guessing correct. First, we assume the probability of guessing correct is $1/n$ on both exams, where n is the number of available answers including both the correct answer and the distractors.[4] The reader might disagree that the probability of guessing correctly is $1/n$ for all possible questions (and the authors agree). However, to achieve usable estimates, one must assume or estimate a specific probability of guessing correctly. If a more appropriate probability of guessing exists, one can set n in the adjusted estimators presented in this section such that $n = 1/p$, where p is the probability of guessing correct. One way to obtain this probability is

to use the three parameter logistic (3PL) model developed in the item response theory (IRT) literature (De Ayala 2013), which includes a pseudo-guessing parameter, usually denoted c or c_i. This estimation technique has the advantage of detecting situations where the specific distractors of a question might modify the probability of an unknowing, low ability student selecting the correct response. We will use this approach in the section *Research: Test of Understanding of College Economics* later in the manuscript.

We will further assume the probability of guessing correctly is identical across individual students and occurs independently on the pre- and post-test. If guessing is largely not based on ability, the former assumption seems reasonable. The latter assumption assumes that enough time passes between the pre- and post-test such that students do not remember their responses on the pre-test.

We will describe the process of constructing the expected value of the unadjusted flow of knowledge (post-test minus pre-test – $E[\hat{\text{flow}}]$) and unadjusted positive learning ($E[\hat{\text{pl}}]$); all other expected values are similarly constructed. If $\mu - \alpha + \gamma$ students knew the answer to the question at the time of the post-test, they selected the correct answer without guessing. However, $1-\mu-\gamma+\alpha$ students did not know the answer and guessed with $1/n$ probability of guessing correct. Similarly, at the time of the pre-test μ students selected the correct answer and $1 - \mu$ students guessed. Therefore, the expected value of the flow of knowledge is:

$$E[\hat{\text{flow}}] = \underbrace{(\mu - \alpha + \gamma + \frac{1}{n}(1 - \mu - \gamma + \alpha))}_{\text{Post-test}} - \underbrace{(\mu + \frac{1}{n}(1 - \mu))}_{\text{Pre-test}} = \frac{(n-1)(\gamma - \alpha)}{n} \qquad (1)$$

If $\gamma - \alpha$ is the term of interest, then $E[\hat{\text{flow}}]$ is an underestimate of the flow of knowledge (i.e. $E[\hat{\text{flow}}]$ is less than the true value of $\gamma - \alpha$ in the expectation; all other uses of the terms 'underestimate' and 'overestimate' are similarly defined in relation to their underlying parameter values). Nonetheless, this is less problematic than it may seem at first. Because

$E[\hat{\text{flow}}]$ only depends on γ and α, all values are simply rescaled with a factor dependent on the probability of guessing correct – it is an underestimate but a uniform underestimate.[5] Fortunately, this means that it is possible to compare two $\hat{\text{flow}}$ values ordinally as long as the question and distractors are the same; this holds true even when the probability of guessing $(1/n)$ is unknown. The disaggregated learning types, however, are more complex.

Two portions of the student population contribute to the unadjusted positive learning value ($\hat{\text{pl}}$): those who learned the material (γ) and those who did not know the material at the time of either exam $(1-\gamma-\mu)$. Those that learned the material had $1/n$ chance of guessing correct on the pre-test, and given they will answer correctly on the post-test, they will be misclassified as retained learning. Thus $(n-1)/n$ of students who learned the material will be properly classified as positive learning. Amongst the group of students who never knew that material, $(n-1)/n$ will guess incorrectly on the pre-test; of the group that guessed incorrectly on the pre-test, $1/n$ will guess correctly on the post-test. With this information, we can construct the expected value of the unadjusted positive learning estimate:

$$E[\hat{\text{pl}}] = (1-\gamma-\mu)\frac{n-1}{n}\frac{1}{n} + \gamma\frac{n-1}{n} = \frac{(n-1)(1-\mu+\gamma(n-1))}{n^2} \tag{2}$$

If the variable of interest is γ then the expected value of $\hat{\text{pl}}$ can be an under- or overestimate depending on the values of μ, γ, and n. In Table 1, we provide the expected value of each of the unadjusted learning types. Further, we provide the expected value as the number of distractors approaches infinity.

Each of the four unadjusted leaning types in Table 1 exhibit the expected behavior: the flow equals equals $E[\hat{\text{pl}}] - E[\hat{\text{nl}}]$, the post-test equals $E[\hat{\text{pl}}] + E[\hat{\text{rl}}]$, the pre-test equals $E[\hat{\text{nl}}] + E[\hat{\text{rl}}]$, and $E[\hat{\text{pl}}] + E[\hat{\text{nl}}] + E[\hat{\text{rl}}] + E[\hat{\text{zl}}] = 1$. Additionally, as the number of distractors approaches infinity and the probability of guessing correctly approaches zero, each estimate equals the underlying learning parameter(s) as the estimates are no longer impacted by the

Table 1: Expected Value of Flow and Disaggregated Learning Types

Estimate	Expected Value	Limit ($n \to \infty$)
Post- Pre-test ($\hat{\text{flow}}$)	$(\mu - \alpha + \gamma + \frac{1}{n}(1 - \mu - \gamma + \alpha)) - (\mu + \frac{1}{n}(1 - \mu)) = \frac{(n-1)(\gamma-\alpha)}{n}$	$\gamma - \alpha$
Positive Learning ($\hat{\text{pl}}$)	$(1 - \gamma - \mu)\frac{n-1}{n}\frac{1}{n} + \gamma\frac{n-1}{n} = \frac{(n-1)(1-\mu+\gamma(n-1))}{n^2}$	γ
Negative Learning ($\hat{\text{nl}}$)	$\frac{n-1}{n}\alpha + \frac{1}{n}\frac{n-1}{n}(1 - \gamma - \mu) = \frac{(n-1)(1-\gamma-\mu+\alpha n)}{n^2}$	α
Retained Learning ($\hat{\text{rl}}$)	$\mu - \alpha + (1 - \gamma - \mu)\frac{1}{n}\frac{1}{n} + \gamma\frac{1}{n} + \alpha\frac{1}{n} = \mu - \alpha + \frac{1-\gamma-\mu}{n^2} + \frac{\gamma}{n} + \frac{\alpha}{n}$	$\mu - \alpha$
Zero Learning ($\hat{\text{zl}}$)	$(1 - \mu - \gamma)\frac{n-1}{n}\frac{n-1}{n} = \frac{(n-1)^2(1-\gamma-\mu)}{n^2}$	$1 - \gamma - \mu$

difference in the size of the population guessing.

However, when the number of distractors is finite, each learning type is inaccurate in the expectation when accounting for guessing. Moreover, the disaggregation has 'split' the inaccuracy across types. It is relatively easy to see that $E[\hat{\text{rl}}]$ overestimates $\mu - \alpha$ (true retained learning) as the other terms in the expected value are always positive. However, $E[\hat{\text{zl}}]$ underestimates $1 - \gamma - \mu$ (true zero learning) as $(n-1)^2/n^2$ is always less than one.[6]Positive and negative learning have a more complex relationship to the true value.

Consider the comparative statics of $E[\hat{\text{pl}}]$ with respect to μ. $\Delta E[\hat{\text{pl}}]/\Delta\mu = (1-n)/n^2$, which is always negative. Therefore, an unadjusted positive learning score can not be compared to another unless the other parameter values match. For example, suppose that you are examining two different classes that answered the same question as a pre- and post-test. Both classes exhibit the same amount of unadjusted positive learning. However, μ (pre-test knowledge) is 0.1 greater for one of the classes than the other. Using $\Delta E[\hat{\text{pl}}]/\Delta\mu$ and assuming $n = 4$ then the amount of positive learning attributed to the two classes were differentially impacted by the difference in μ by about -0.019.[7] Therefore, more true positive learning (γ) occurred in the class with the greater μ.

Using the information in Table 1 we can find estimates that adjust for guessing in the expectation. The expected values of $\hat{\text{pl}}$, $\hat{\text{nl}}$, and $\hat{\text{rl}}$ are known (three equations), and there are three unknown variables (μ, γ, and α). Therefore, we can simultaneously solve for each

variable expressed in terms of the unadjusted disaggregation:[8]

$$\begin{aligned}\hat{\mu} &= \frac{\hat{\text{nl}} + \hat{\text{rl}} - 1}{n-1} + \hat{\text{nl}} + \hat{\text{rl}} \\ \hat{\gamma} &= \frac{n(\hat{\text{nl}} + \hat{\text{pl}}n + \hat{\text{rl}} - 1)}{(n-1)^2} \\ \hat{\alpha} &= \frac{n(\hat{\text{nl}}n + \hat{\text{pl}} + \hat{\text{rl}} - 1)}{(n-1)^2}\end{aligned} \tag{3}$$

Each estimate is now expressed with a hat as the observed values of $\hat{\text{pl}}$, $\hat{\text{nl}}$, and $\hat{\text{rl}}$ may not match their expected value. However, as the number of observations increases, the observed values will approach their expected values in Table 1[9] – assuming the probability of guessing correctly is properly specified. Further, each of the adjusted estimates are equivalent to their unadjusted counterparts as n approaches infinity.[10]

Monte Carlo Simulation of the Counterfactual

The distributional properties of the learning type disaggregation are intrinsically complex and may not follow a standard distribution. The unadjusted positive, negative, retained, and zero learning must sum to one while simultaneously each must not be less than zero. This creates bounds in the distribution that depend on the underlying parameter values of μ, α, and γ. Further, in smaller classes, the proportion of the students who know the answer to a given question before entering the class will vary even if the proportion knowing the answer of the overall population who take the class remains the same.

Researchers in the field of regional economics face similar issues and often use simulations to assist in analysis. Some notable examples include Amrhein (1995), Cassey and Smith (2014), Deltas (2003) and Duranton and Overman (2005).

Two very closely related procedures to our own are those of Duranton and Overman (2005) and Cassey and Smith (2014). In both simulations the authors constructed simulated

counterfactuals of no underlying localization[11] to use as a comparison distribution. This allowed the researchers to statistically say that a given industry is localized beyond what one might observe from randomness alone. This is despite the presence of bounds in the underlying parameter values.

The analogous counterfactual to no localization in our study is no learning. In this section we use a Monte Carlo simulation to create an empirical distribution under a known specification of no positive or negative learning. Briefly, our Monte Carlo simulation attempts to simulate the 'real world' data generating process itself (in our case, guessing) then observe the resulting outcomes. If the simulated data generating process matches the real world data generating process then the resulting distribution of outcomes matches the real world distribution of outcomes given sufficient replications and that the underlying assumptions (no learning) are true.

We then extract the 90 and 95 percent quantiles from the resulting simulated empirical distribution; this constructs a table of critical values. If an observed learning value is greater than the critical value this indicates that the observed value occurred in less than ten (or five) percent of the simulations where there was no underlying learning. Put another way, the observed learning value is greater than what one would expect to observe from randomness alone when there is no learning.

These simulations contain important assumptions. First, we assume the students not only do not learn the material but their ability to guess does not change from the pre- to the post-test. Further, we assume the students do not remember the pre-test questions when answering the post-test. Therefore, each student's answers on the pre- and post-test are independent. We think these assumptions are reasonable given we are simulating no learning where we assume a substantial amount of time has passed between the pre- and post-test.

Simulation Setup

In these simulations, the null hypothesis is zero true positive learning ($\gamma = 0$) and zero true negative learning ($\alpha = 0$) on a single question answered on a pre- and post-test. Therefore, the true γ and α value is set to zero for all simulations (but the simulated observed values can be in the negative range). We performed these simulations at increasing values of μ ($\{0.2, 0.3, ..., 0.8\}$) and class sizes of 15 to 300 students (m). The procedure works as follows:

1. For simulated class i, m students are pulled from a population where μ proportion of the students know the answer to the question and thus answer it on the pre- and post-test correctly.

2. Of the students who do not know the answer, each guesses with $1/n$ probability of guessing correctly. A student's guesses on the pre- and post-test are independent of each other.

3. With the pre- and post-test answers determined, each student is assigned to one of the four unadjusted learning types (assigning '1' for the appropriate learning type and '0' for all other learning type columns).

4. Taking class i in aggregate, the four learning types are determined by taking the average of each column. Adjusted estimates are then calculated using equation 3 from the unadjusted measures. This is a simulation of one question taken by one random class.

5. This procedure is repeated 10,000 times to generate 10,000 random classes with m students and stock knowledge μ. From this, one can generate quantile values for $\hat{\gamma}$, $\hat{\alpha}$ and $\hat{f}$ (adjusted flow: $\hat{\gamma} - \hat{\alpha}$).

Each simulation's mean value of the vectors of $\hat{\gamma}$s, $\hat{\alpha}$s, and $\hat{f}$s have been checked and they

all equal zero, as intended. Similarly, the mean value of the simulated $\hat{\mu}$s equals the seeded μ value of that simulation.

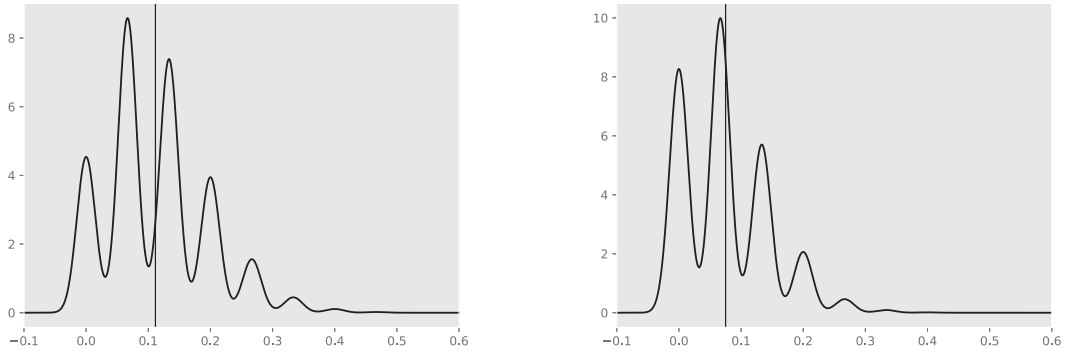
Simulation Results

In simulating class sizes of 15 and 100 students, Kernel Density Estimations plots (KDE – see Scott (2015) for a detailed description) were used to visualize both the adjusted and unadjusted distributions. We have provided KDE plots of adjusted and unadjusted positive learning, but for the sake of brevity, we have not included KDE plots of the other learning types as they tell a similar story to positive learning. However, these plots are available upon request.

For each KDE plot, we used Silverman's (1986) method to calculate the optimal bandwidth. However, for comparison purposes, we then averaged all four of the $m = 15$ optimal bandwidth values and selected the nearest round number; the resulting value is 0.015. Similarly, we averaged the four $m = 100$ optimal bandwidth values and selected the nearest round number: 0.006. Therefore, figures 2 and 3 use a bandwidth value of 0.015 while figures 4 and 5 use a bandwidth value of 0.006.

Examining Figure 2, we see the distributions of unadjusted positive learning estimates ($\hat{\text{pl}}$) when the true amount of positive learning is zero with a class size of 15. As expected, the mean values – plotted as a vertical line – do not match zero. Moreover, due to the differing μ values, they do not match each other. Additionally, the figure is very noisy; the finite number of possible values with a small class is probably the primary contributor to the noisiness of the distributions.

The mean values of $\hat{\text{pl}}$ are in contrast to the adjusted positive learning distributions ($\hat{\gamma}$). Examining Figure 3, we see the mean values – again, plotted as vertical lines – are at zero as expected. However, we do see a relationship between μ and the width of the distributions. Intuitively, as μ moves towards one, a smaller proportion of the class is guessing and thus



(a) $\mu = 0.4$ (b) $\mu = 0.6$

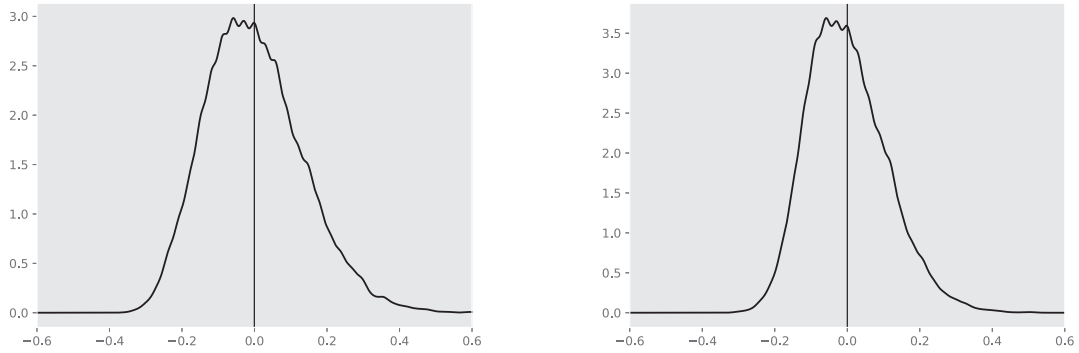
Figure 2: Kernel Density Estimations (KDE) of unadjusted positive learning ($\hat{\text{pl}}$) where μ (stock knowledge) is set to $\{0.4, 0.6\}$ and the true population γ (adjusted positive learning) equals zero. Both simulations are of class sizes of 15 students with a total of four answer options per question. The mean value is plotted as a vertical line. The distribution density is on the y axes and $\hat{\text{pl}}$ is on the x axes.

there is less possible noise from the guessing students.

Simulations of $\hat{\text{pl}}$ and $\hat{\gamma}$ with class sizes of 100 have some familiar themes, but some important differences, to the simulations where the class size was 15. Figure 4 shows the distribution of $\hat{\text{pl}}$ when true learning is zero. Like the previously discussed simulations, the mean values of the distributions are not zero and do not match each other. However, the distributions do appear more narrow and symmetrical. As the size of the class increases, both the uncorrected and corrected distributions narrow due to the law of large numbers. Nonetheless, at any class size, the distributions from the simulations with the lower values of μ are wider as a higher percent of each class is guessing.

In Figure 5, we see distributions of $\hat{\gamma}$ when the class size is 100 students. Like the 15 student simulations of $\hat{\gamma}$, the distribution is centered on zero; like $\hat{\text{pl}}$, the distributions have narrowed with the larger classes.

In aggregate we have produced approximately four hundred simulated distributions of 10,000 classes with differing class sizes and values of μ. We present samples of these results in Tables 4 through 7. Tables 4 and 5 assume $n = 4$ (number of question options), while Tables



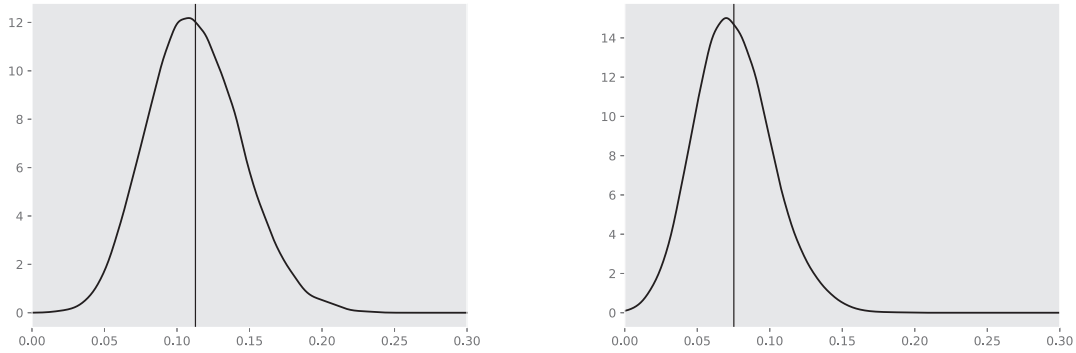
(a) $\mu = 0.4$ (b) $\mu = 0.6$

Figure 3: Kernel Density Estimations (KDE) of adjusted positive learning ($\hat{\gamma}$) where μ (stock knowledge) is set to $\{0.4, 0.6\}$ and the true population γ (adjusted positive learning) equals zero. Both simulations are of class sizes of 15 students with a total of four answer options per question. The mean value is plotted as a vertical line. The distribution density is on the y axes and $\hat{\gamma}$ is on the x axes.

6 and 7 assume $n = 5$. We have not included our simulation results for adjusted negative learning in the paper as the critical values are extremely similar to adjusted positive learning (examining equation 3 one can see why this would be the case). However, these results, along with many more class sizes for the other distributions, can be found in the online version of these tables (https://goo.gl/mjIYvN). Additionally, we include simulations where $n = 3$ and $n = 6$ in the online version of the tables. Finally, our Python simulation code has been posted on GitHub (https://goo.gl/zOqlDx) so that the interested reader can simulate other values of n (including non-whole numbers) pertinent to their specific application.

In each one of these simulations, we extracted the 90% and 95% quantile corresponding to a one-sided test with type I error of 10% and 5%, respectively. Further, we present the 95% confidence interval for $\hat{\mu}$ for that particular distribution. The mean value of $\hat{\mu}$ is the underlying μ value for that simulation. The $\hat{\mu}$ confidence interval will assist the practitioner in selecting the appropriate critical value for their class.

Suppose you are attempting to assess the positive learning in a class with eighty students where you used a four-option multiple choice pre- and post-test to assess a set of learning

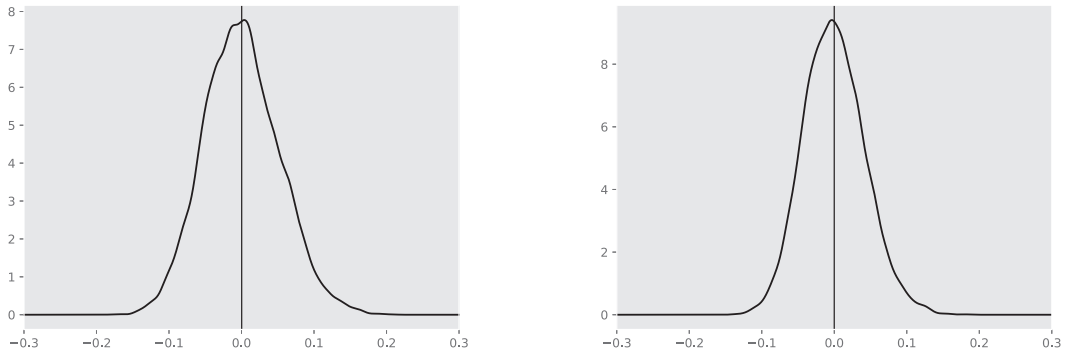


(a) $\mu = 0.4$ (b) $\mu = 0.6$

Figure 4: Kernel Density Estimations (KDE) of unadjusted positive learning ($\hat{\text{pl}}$) where μ (stock knowledge) is set to $\{0.4, 0.6\}$ and the true population γ (adjusted positive learning) equals zero. Both simulations are of class sizes of 100 students with a total of four answer options per question. The mean value is plotted as a vertical line. The distribution density is on the y axes and $\hat{\text{pl}}$ is on the x axes.

objectives for the class. Using equation 3 you have found the adjusted estimates for $\hat{\gamma}$, $\hat{\alpha}$, and $\hat{\mu}$ for a specific question. Suppose that you calculate $\hat{\gamma}$ to be about 0.12 and $\hat{\mu}$ to be about 0.5. Looking at Table 4, there are eight distributions for a class of eighty students. Nonetheless, comparing your $\hat{\mu}$ of 0.5 to the simulated $\hat{\mu}$ 95% confidence interval, only three rows likely correspond to the $\hat{\mu}$ value you found in the data. The greatest critical values of these three can be found with a $\mu = 0.4$ where we see that the 90% critical value is 0.078 and the 95% critical value is 0.100. If the observed adjusted positive learning ($\hat{\gamma}$) is greater than 0.078 than you can say that in less than 10% of the simulated classes where there was no true positive learning did we simulate a $\hat{\gamma}$ value greater than the observed value from the data. Similarly, in less than 5% of the classes with no true positive learning did we simulate a $\hat{\gamma}$ value greater than 0.100. With a $\hat{\gamma}$ value of 0.12 you can be reasonably certain that positive learning occurred in the class. Using Table 5 and the online appendix, you could perform a similar statistical test for the adjusted flow of knowledge ($\hat{f}$) and adjusted negative learning ($\hat{\alpha}$).

However, suppose the class consisted of only 30 students and all other details remained



(a) $\mu = 0.4$ (b) $\mu = 0.6$

Figure 5: Kernel Density Estimations (KDE) of adjusted positive learning ($\hat{\gamma}$) where μ is set to $\{0.4, 0.6\}$ and the true population $\gamma = 0$. Both simulations are for class sizes of 100 students with a total of four answer options per question. The mean value is plotted as a vertical line. The distribution density is on the y axes and $\hat{\gamma}$ is on the x axes.

the same. From Table 4, the appropriate critical values for a class of 30 would be 0.133 for the 90% threshold and 0.163 for the 95% threshold. Under this scenario, you would not be able to reject the null that the positive learning you observed comes from randomness alone. However, as the critical values on Table 6 are slightly smaller than on Table 4, an observed adjusted positive learning value of 0.12 would be significant at the 10% level when $n = 5$.

While our statistical test is useful for instructors with large classes, we believe it is also useful for those who teach small classes. Our simulation tables provide insight to instructors about the randomness in the disaggregated distributions. Some instructors teaching very small classes might never see a disaggregated learning score on a question that is statistically significant. This, of course, does not mean their students have failed to learn, but rather they should expect a large amount of noise from one semester to another due to students guessing. For classes of this size, disaggregating and adjusting pre- and post-test scores might be more confusing than useful.

Applications

We have provided two example uses of our method in this section. In *Assessment: Principles of Microeconomics*, we present ten assessment questions from a principles of microeconomics course taught by one of the authors. In *Research: Test of Understanding of College Economics*, we examine the nationally-normed data from the microeconomics *Test of Understanding of College Economics* (Walstad, Watts, and Rebeck 2007) and adjust the results to account for guessing.

Assessment: Principles of Microeconomics

In the Fall of 2016, one of the authors taught a little under 90 principles of microeconomics students. Each student received a pre-test on the first day of class and the same questions were embedded into the final exam. In total, 83 students (m) took both the pre- and post-test.

Using software by Smith (2018), we disaggregated the test results into both the unadjusted and adjusted learning types. In the absence of other information, the instructor believes the ten assessment questions in Table 2 can be guessed correct with a probability of about 1/4. Thus, we present adjusted learning type estimates using $n = 4$.

Examining the results in Table 2, on average, we see nearly the same amount of positive learning from both the unadjusted and adjusted measures. However, for some of the questions, unadjusted positive learning is greater than the adjusted measure while in others it is less than the adjusted measure. As discussed in *Guessing-Adjustment*, unadjusted positive learning can be an over- or underestimate in the expectation. Adjusted negative learning has trended towards zero indicating most unadjusted negative learning in this dataset was likely due to guessing.

Using Table 4 and the procedure discussed in the previous section, we can easily see

Table 2: Assessment Questions, 83 Students

Question #	Unadjusted Leaning Types					Adjusted Leaning Types				
	$\hat{\text{pl}}$	$\hat{\text{nl}}$	$\hat{\text{rl}}$	$\hat{\text{zl}}$	$\hat{\text{flow}}$	adj. $\hat{\text{pl}}$ ($\hat{\gamma}$)	adj. $\hat{\text{nl}}$ ($\hat{\alpha}$)	adj. $\hat{\text{rl}}$ ($\hat{\mu} - \hat{\alpha}$)	adj. $\hat{\text{zl}}$ ($1 - \hat{\gamma} - \hat{\mu}$)	adj. $\hat{\text{flow}}$ ($\hat{\gamma} - \hat{\alpha}$)
1	0.38	0.09	0.23	0.29	0.29	0.38	-0.01	0.11	0.52	0.39
2	0.26	0.21	0.13	0.41	0.05	0.16	0.09	0.02	0.73	0.07
3	0.41	0.08	0.22	0.29	0.33	0.42	-0.03	0.09	0.52	0.44
4	0.50	0.01	0.41	0.08	0.49	0.63	-0.02	0.25	0.14	0.65
5	0.46	0.10	0.18	0.26	0.36	0.50	0.02	0.02	0.46	0.48
6	0.29	0.15	0.23	0.32	0.14	0.25	0.06	0.12	0.57	0.19
7	0.24	0.15	0.31	0.29	0.09	0.19	0.07	0.21	0.52	0.12
8	0.33	0.09	0.23	0.35	0.24	0.29	-0.03	0.13	0.62	0.32
9	0.53	0.05	0.26	0.17	0.47	0.63	-0.01	0.08	0.30	0.63
10	0.31	0.10	0.29	0.29	0.21	0.28	0.01	0.19	0.52	0.27
Averages	0.37	0.10	0.25	0.28	0.27	0.37	0.02	0.12	0.49	0.36

Notes: Assessment questions from principles of microeconomics with 83 students (m) and the instructor assumed $n = 4$. All adjusted positive learning values are statistically different from randomness. However, in all but two questions (2 & 7), we can not reject the null that the observed negative learning is from randomness alone.

that each of the adjusted positive learning values are greater than we would expect from randomness alone. However, the same can not be said of adjusted negative learning. With eight of the $\hat{\alpha}$ values we can not reject the null that the true α is zero; the remaining two (2 & 7) can be rejected at the 10% level. The instructor has adjusted the corresponding lesson plans to (hopefully) reduce confusion in future versions of the class. However, perhaps the most striking result of Table 2 is the large amount of adjusted zero learning. Clearly, this should be a focus of the instructor in future versions of the course. For the next semester, the instructor has reallocated some of the course homework questions from topics with relatively low adjusted zero learning to topics with relatively high adjusted zero learning.

Research: Test of Understanding of College Economics

The *Test of Understanding of College Economics* (TUCE) is a nationally-normed four-option multiple choice exam at the college principles of economics level. The TUCE includes both a micro- and macroeconomics exam, both of which contain thirty questions. The microeconomics TUCE is nationally-normed using a sample of 3255 matched-pairs: students who took both the pre- and post-test. Given the matched sample, Walstad and Wagner (2016)

disaggregated the results into the four unadjusted learning types; these are the estimates that we use in our analysis.

In Table 3, we present both the original unadjusted and adjusted measures of learning. Unlike our assessment results, we do not assume $n = 4$. Instead, we estimate the pseudo-guessing parameter (c_i) for each question using the standard three parameter logistic model (3PL). This parameter represents the chance that a very low ability student correctly answers a given question – mathematically, it is the lower asymptote. For more on item response theory and this empirical approach see De Ayala (2013). The estimates for the pseudo-guessing parameter are presented in the column $1/n$ with the standard error in parentheses. These estimates were then used to adjust the original disaggregation.

The three parameter logistic model (3PL) approach in Table 3 is of some debate in the psychometric community (Żółtak and Golonka 2015). In fact, Han (2012) has suggested that practitioners should fix the guessing parameter at $1/n$ (or $1/k$ in the notation of the paper) as a means of improving the estimates of other parameters in the 3PL model. For this reason, in addition to the results provided in Table 3, we have provided estimates of the adjusted learning types assuming $n = 4$ and made them available online (https://goo.gl/BO6B4Z).

It is important to emphasize that the values presented in Table 3 are estimates. There are two sources of noise that could impact our results. The first source is the guessing process itself, as simulated in the previous section. Comparatively, this is a small source of noise as simulations with 3000 students and $n = 4$ have resulted in 95% critical values for $\hat{\gamma}$, $\hat{\alpha}$, and $\hat{\text{f}}$ of about 0.02 and 0.01 when $\mu = 0.1$ and $\mu = 0.8$, respectively. Nonetheless, this source compounds the (larger on average) second source of noise which is the estimate of the guessing parameter itself. Unfortunately, there is no way to combine these two sources of noise for a statistical test as the noise estimates are built using differing methodologies. To perform a test, the practitioner would have to assume either the guessing distribution or the estimate of $1/n$ is a point.

Table 3: TUCE: Principles of Microeconomics

		Unadjusted Leaning Types					Adjusted Leaning Types				
Question #	$1/n$ (se)	$\hat{\text{pl}}$	$\hat{\text{nl}}$	$\hat{\text{rl}}$	$\hat{\text{zl}}$	$\hat{\text{flow}}$	adj. $\hat{\text{pl}}$ ($\hat{\gamma}$)	adj. $\hat{\text{nl}}$ ($\hat{\alpha}$)	adj. $\hat{\text{rl}}$ ($\hat{\mu}-\hat{\alpha}$)	adj. $\hat{\text{zl}}$ ($1-\hat{\gamma}-\hat{\mu}$)	adj. $\hat{\text{flow}}$ ($\hat{\gamma}-\hat{\alpha}$)
1	0.18 (0.04)	0.26	0.14	0.24	0.36	0.11	0.22	0.08	0.17	0.53	0.14
2	0.18 (0.05)	0.23	0.16	0.17	0.44	0.07	0.16	0.08	0.11	0.66	0.08
3	0.18 (0.05)	0.29	0.15	0.21	0.35	0.15	0.26	0.08	0.13	0.52	0.18
4	0.29 (0.05)	0.49	0.06	0.08	0.37	0.43	0.47	-0.13	-0.08	0.73	0.60
5	0.19 (0.06)	0.26	0.20	0.20	0.34	0.06	0.22	0.15	0.11	0.52	0.07
6	0.22 (0.02)	0.31	0.08	0.15	0.46	0.22	0.23	-0.06	0.08	0.76	0.29
7	0.14 (0.04)	0.21	0.16	0.28	0.35	0.04	0.18	0.13	0.23	0.47	0.05
8	0.12 (0.02)	0.23	0.07	0.14	0.57	0.16	0.17	-0.01	0.11	0.73	0.18
9	0.17 (0.03)	0.22	0.13	0.09	0.56	0.09	0.13	0.02	0.04	0.81	0.11
10	0.22 (0.09)	0.23	0.16	0.21	0.40	0.07	0.15	0.06	0.13	0.66	0.09
11	0.08 (0.02)	0.26	0.06	0.05	0.63	0.21	0.23	0.00	0.03	0.74	0.22
12	0.27 (0.03)	0.33	0.12	0.12	0.42	0.21	0.24	-0.05	0.01	0.80	0.29
13	0.22 (0.07)	0.29	0.15	0.21	0.34	0.14	0.25	0.07	0.11	0.56	0.18
14	0.24 (0.03)	0.31	0.16	0.14	0.38	0.15	0.25	0.05	0.03	0.67	0.19
15	0.23 (0.02)	0.22	0.10	0.11	0.56	0.12	0.08	-0.08	0.07	0.94	0.16
16	0.22 (0.05)	0.25	0.18	0.25	0.32	0.07	0.21	0.12	0.15	0.53	0.09
17	0.08 (0.03)	0.27	0.15	0.17	0.42	0.12	0.25	0.12	0.13	0.49	0.13
18	0.22 (0.04)	0.26	0.15	0.15	0.44	0.11	0.17	0.03	0.07	0.72	0.14
19	0.35 (0.02)	0.21	0.21	0.23	0.36	0.00	0.02	0.02	0.11	0.85	0.00
20	0.21 (0.01)	0.25	0.11	0.06	0.58	0.14	0.12	-0.06	0.01	0.93	0.17
21	0.09 (0.04)	0.23	0.21	0.22	0.35	0.02	0.21	0.19	0.18	0.42	0.02
22	0.18 (0.08)	0.24	0.20	0.35	0.21	0.04	0.23	0.19	0.27	0.31	0.04
23	0.20 (0.02)	0.21	0.14	0.10	0.55	0.07	0.09	0.00	0.05	0.86	0.09
24	0.17 (0.06)	0.27	0.18	0.22	0.33	0.08	0.24	0.14	0.14	0.47	0.10
25	0.15 (0.02)	0.22	0.12	0.12	0.55	0.10	0.14	0.02	0.07	0.76	0.12
26	0.30 (0.01)	0.23	0.18	0.11	0.48	0.05	0.04	-0.03	0.03	0.97	0.07
27	0.16 (0.06)	0.26	0.14	0.15	0.45	0.12	0.21	0.07	0.09	0.64	0.14
28	0.20 (0.02)	0.24	0.14	0.11	0.51	0.10	0.14	0.01	0.04	0.80	0.13
29	0.16 (0.03)	0.23	0.16	0.15	0.47	0.06	0.16	0.09	0.09	0.66	0.08
30	0.08 (0.04)	0.26	0.17	0.23	0.34	0.09	0.25	0.15	0.20	0.40	0.10
Averages	0.19 (0.04)	0.26	0.15	0.17	0.43	0.11	0.19	0.05	0.10	0.66	0.14

Notes: Unadjusted and adjusted estimates of the nationally-norming sample of 3255 students who took the microeconomics TUCE pre- and post-test. Probability of guessing correct ($1/n$) estimated using a 3PL procedure (De Ayala 2013); standard errors for this estimate presented in parentheses.

We can further emphasize that these values are estimates by considering question four. Given the estimate of $1/n$ of 0.29, the adjusted negative learning is about -0.13 and the adjusted retained learning is about -0.08. However, the estimate of $1/n$ has a 95% confidence interval of $[0.19, 0.39]$ (setting aside the additional noise from the guessing process itself). Using the lower bound, we can reestimate the adjusted learning values for question four; we find adj. $\hat{\text{pl}} = 0.49$, adj. $\hat{\text{nl}} = -0.03$, adj. $\hat{\text{rl}} = -0.02$, and adj. $\hat{\text{zl}} = 0.56$. These results seem more reasonable, however other results may be an under- or overestimate in a less detectable way.

More generally, consider the adjusted negative learning column as a whole. Averaging about 0.05, many items stray into the negative range. While others may have a different interpretation, what the authors see is statistical noise from both the guessing process and estimates of $1/n$ resulting in some $\hat{\alpha}$ values substantially above the average and some (noticeably) below the average in an impossible range.

Overall, our average adjusted estimates of learning suggest that positive learning on the microeconomics TUCE may have been overestimated by the original disaggregation by Walstad and Wagner (2016). This is in contrast to the results in *Assessment: Principles of Microeconomics* where the average unadjusted and adjusted positive learning values are nearly the same. Similar to our assessment results, negative learning has moved towards zero suggesting that a portion of the unadjusted measure is due to guessing alone. Finally, zero learning is substantively underestimated by the unadjusted measure. While the original disaggregation might have encouraged instructors to focus their attention on students that are somehow 'unlearning' material (perhaps through a confusing explanation), our results suggest that true negative learning is rare and instructors should instead focus on the high proportion of zero learners.[12]

Conclusion

Disaggregated learning analysis provides researchers, practitioners, and assessment committees new insights into the learning occurring in classrooms. Different types of learning require different strategies to encourage (or discourage) their growth and should not be intermixed. The work of Walstad and Wagner (2016) represents a major step forward by disaggregating pre- and post-test results into four distinct learning types. This paper builds on that work by creating learning type estimates that adjust for student guessing. This will help researchers more accurately detect phenomena as well as help practitioners make better decisions.

Further, this paper provides Monte Carlo simulation code and results that can be used

as a statistical test given that certain assumptions about the data generating process are true. Using this information, a practitioner can determine whether their students improved better than chance on a particular question given their class size and the observed adjusted estimates of learning. Additionally, these simulations suggest the amount of statistical noise an instructor of a small class can expect could be substantial and in some cases may not be worth the effort to disaggregate and adjust the results. In the *Applications* section we demonstrated the usefulness of this approach.

One opportunity for future research presented by this paper is determining the probability of a student guessing the correct answer on specific exam questions that are used nationwide. Through this manuscript, we have often described the probability of guessing correct as $1/n$ where n is the number of question options. However, as demonstrated with the microeconomics TUCE, this is not always the case. Using an item response theory approach, one can estimate the probability of guessing correctly on a specific question. However, such an approach requires a large amount of data; even some of the largest classes in the country would not be of sufficient size to make such an estimate and would have to resort to a rule of thumb (e.g. $1/n$). If the instructor wishes to use an empirically estimated probability of guessing correct, they should consider using a nationally-normed exam. An instructor could use the empirical estimates of the probability of guessing correct (i.e. $1/n$ or c) using the nationally-norming sample then use those same estimated probabilities to adjust their classroom results.

Notes

[1]While guessing is addressed in Salemi and Tauchen (1980), most studies concerned with inaccuracies associated with estimating learning relate to either sampling error (e.g. Bowles and Jones (2003); Koedel, Mihaly, and Rockoff (2015), pp. 183–185) or data loss (Becker and Powers 2001; Becker and Walstad 1990).

[2]We will use hats (e.g. $\hat{\text{pl}}$) throughout this manuscript to emphasis when we are discussing an estimate. Thus, μ is a true underlying parameter while $\hat{\mu}$ is an estimate of μ using the data.

[3]This holds true as long as true positive learning is greater than true negative learning (forgetting). The proportion of students guessing on the pre-test is $1 - \mu$, while the proportion guessing on the post-test (assuming no true negative learning) is $1 - \mu - \gamma$.

[4]It is possible that the probability of guessing correctly is different on the pre- and post-test assuming success is correlated to student ability. However, an usable modification to our assumed probability $(1/n)$ is not readily available. Further, the primary estimation technique available in the item response theory literature is ill-suited to detect such a difference; the pseudo-guessing parameter in the three parameter logistic model is the probability that a very low ability student responds correct nonetheless. By design, the model is attempting to remove ability as a factor.

[5]However, the bounds of γ depend on μ. γ has an upper constraint at $1 - \mu$.

[6]However, like $E[\hat{\text{flow}}]$, $E[\hat{\text{zl}}]$ can be treated as an ordinal or count variable as it only depends on the probability of guessing correct.

[7]$\Delta E[\hat{\text{pl}}]/\Delta\mu$ shows the resulting change in $E[\hat{\text{pl}}]$ with a one unit change in μ. Therefore, the resulting change in $\Delta E[\hat{\text{pl}}]$ from a 0.1 unit change in μ can be calculated as follows: $0.1 \times ((1 - 4)/4^2)$.

[8]For the reader's convenience we have included equation 3 expressed in terms of the

probability of guessing correct ($\hat{p}$) instead of the number of answer options (n). Equation 3 can be re-expressed as:

$$\begin{aligned}
\hat{\mu} &= \frac{\hat{\text{nl}} + \hat{\text{rl}} - \hat{p}}{1 - \hat{p}} \\
\hat{\gamma} &= \frac{\hat{p}(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}}{(\hat{p} - 1)^2} \\
\hat{\alpha} &= \frac{\hat{p}(\hat{\text{pl}} + \hat{\text{rl}} - 1) + \hat{\text{nl}}}{(\hat{p} - 1)^2}
\end{aligned} \tag{4}$$

[9]These same adjusted estimates can be used when aggregating to the student instead of the question. However, the results should be interpreted carefully. As there are often fewer questions than there are students in a class, the adjusted learning types will be a noisier signal due to small numbers randomness when aggregating to the student.

[10]As n approaches infinity, each of the adjusted estimators becomes equivalent to their unadjusted counterparts. See equation 5.

$$\begin{aligned}
\lim_{n\to\infty} \hat{\mu} &= \lim_{n\to\infty} \frac{\hat{\text{nl}} + \hat{\text{rl}} - 1}{n - 1} + \hat{\text{nl}} + \hat{\text{rl}} = \hat{\text{nl}} + \hat{\text{rl}} \\
\lim_{n\to\infty} \hat{\gamma} &= \lim_{n\to\infty} \frac{n(\hat{\text{nl}} + \hat{\text{pl}}n + \hat{\text{rl}} - 1)}{(n - 1)^2} = \hat{\text{pl}} \\
\lim_{n\to\infty} \hat{\alpha} &= \lim_{n\to\infty} \frac{n(\hat{\text{nl}}n + \hat{\text{pl}} + \hat{\text{rl}} - 1)}{(n - 1)^2} = \hat{\text{nl}}
\end{aligned} \tag{5}$$

[11]Localization is the tendency for an industry's plants to be geographically near each other. Notable examples are the software, car, and oil industry.

[12]We have estimated the adjusted learning types using the macroeconomics TUCE national data with similar results (published online – https://goo.gl/BO6B4Z).

References

Amrhein, Carl G. 1995. "Searching for the elusive aggregation effect: evidence from statistical simulations". *Environment and Planning A* 27(1): 105–119.

Becker, William E and John R Powers. 2001. "Student performance, attrition, and class size given missing student data". *Economics of Education Review* 20(4): 377–388.

Becker, William E and William B Walstad. 1990. "Data loss from pretest to posttest as a sample selection problem". *The Review of Economics and Statistics* 72(1): 184–188.

Bowles, Tyler J and Jason Jones. 2003. "An analysis of the effectiveness of supplemental instruction: The problem of selection bias and limited dependent variables". *Journal of College Student Retention: Research, Theory & Practice* 5(2): 235–243.

Cassey, Andrew J and Ben O Smith. 2014. "Simulating confidence for the Ellison–Glaeser index". *Journal of Urban Economics 81*: 85–103.

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

Deltas, George. 2003. "The small-sample bias of the Gini coefficient: results and implications for empirical research". *The Review of Economics and Statistics* 85(1): 226–234.

Duranton, Gilles and Henry G Overman. 2005. "Testing for localization using micro-geographic data". *The Review of Economic Studies* 72(4): 1077–1106.

Emerson, Tisha L. N. and Linda K. English. 2016. "Classroom experiments: Teaching specific topics or promoting the economic way of thinking?" *The Journal of Economic Education* 47(4): 288–299.

Han, Kyung T. 2012. "Fixing the c parameter in the three-parameter logistic model". *Practical Assessment, Research & Evaluation* 17(1): 1–24.

Happ, Roland, Olga Zlatkin-Troitschanskaia, and Susanne Schmidt. 2016. "An analysis of economic learning among undergraduates in introductory economics courses in Germany". *The Journal of Economic Education* 47(4): 300–310.

Koedel, Cory, Kata Mihaly, and Jonah E Rockoff. 2015. "Value-added modeling: A review". *Economics of Education Review 47*: 180–195.

Salemi, Michael K. and George E. Tauchen. 1980. "Guessing and the error structure of learning models". *The American Economic Review* 70(2): 41–46.

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

Siegfried, John J and Rendigs Fels. 1979. "Research on teaching college economics: A survey". *Journal of Economic Literature* 17(3): 923–969.

Silverman, Bernard W. 1986. *Density estimation for statistics and data analysis*, Volume 26. London: Chapman & Hall.

Smith, Ben. 2018. "Multiplatform software tool to disaggregate and adjust value-added learning scores". *The Journal of Economic Education*: Forthcoming.

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

Walstad, William B, Michael Watts, and Ken Rebeck. 2007. *Test of understanding of college economics: Examiner's manual* (4th ed.). New York: Council for Economic Education.

Żółtak, Tomasz and Grzegorz Golonka. 2015. "Does guessing matter? Differences between ability estimates from 2PL and 3PL IRT models in case of guessing." *EDUKACJA* 134(3): 65–78.

Appendix: Critical Value Tables

Table 4: 90% and 95% critical values from simulated $\hat{\gamma}$ distribution where $n = 4$ and underlying $\gamma = 0$ and $\alpha = 0$

Students	μ	$\hat{\gamma}_{0.90}$ CV	$\hat{\gamma}_{0.95}$ CV	$\hat{\mu}$ 95% CI	
15	0.1	0.207	0.296	-0.156	0.467
15	0.2	0.207	0.267	-0.156	0.556
15	0.3	0.178	0.237	-0.067	0.644
15	0.4	0.178	0.237	0.022	0.733
15	0.5	0.178	0.207	0.200	0.822
15	0.6	0.148	0.207	0.289	0.911
15	0.7	0.119	0.148	0.378	0.911
15	0.8	0.089	0.148	0.556	1.000
30	0.1	0.148	0.193	-0.111	0.333
30	0.2	0.148	0.193	-0.022	0.422
30	0.3	0.133	0.163	0.067	0.556
30	0.4	0.119	0.163	0.156	0.644
30	0.5	0.119	0.148	0.244	0.733
30	0.6	0.104	0.133	0.378	0.822
30	0.7	0.089	0.119	0.511	0.867
30	0.8	0.074	0.104	0.600	0.956
50	0.1	0.116	0.151	-0.067	0.280
50	0.2	0.107	0.142	0.013	0.387
50	0.3	0.107	0.133	0.120	0.493
50	0.4	0.089	0.124	0.227	0.573
50	0.5	0.089	0.116	0.307	0.680
50	0.6	0.071	0.098	0.413	0.760
50	0.7	0.071	0.089	0.547	0.840
50	0.8	0.053	0.071	0.653	0.920
80	0.1	0.089	0.117	-0.033	0.233
80	0.2	0.083	0.111	0.067	0.350
80	0.3	0.083	0.106	0.150	0.450
80	0.4	0.078	0.100	0.250	0.550
80	0.5	0.067	0.083	0.350	0.633
80	0.6	0.061	0.078	0.467	0.733
80	0.7	0.056	0.072	0.567	0.817
80	0.8	0.044	0.056	0.700	0.900
100	0.1	0.080	0.107	-0.013	0.227
100	0.2	0.076	0.102	0.080	0.333
100	0.3	0.071	0.093	0.173	0.427
100	0.4	0.071	0.084	0.267	0.533
100	0.5	0.062	0.080	0.373	0.627
100	0.6	0.053	0.071	0.480	0.720
100	0.7	0.049	0.062	0.587	0.800
100	0.8	0.040	0.053	0.707	0.893
300	0.1	0.047	0.061	0.031	0.173
300	0.2	0.044	0.059	0.124	0.276
300	0.3	0.041	0.053	0.222	0.378
300	0.4	0.039	0.050	0.324	0.476
300	0.5	0.036	0.046	0.427	0.573
300	0.6	0.033	0.041	0.529	0.667
300	0.7	0.028	0.036	0.636	0.760
300	0.8	0.022	0.028	0.742	0.853

Notes: Each row indicates the critical values at the 90% threshold ($\hat{\gamma}_{0.90}$ CV) and 95% threshold ($\hat{\gamma}_{0.95}$ CV). The 95% confidence interval of the estimated $\hat{\mu}$ values ($\hat{\mu}$ 95% CI) are provided to assist in finding the appropriate $\hat{\gamma}$ critical value. A much larger set of critical values are available at https://goo.gl/mjIYvN. All simulations run with 10,000 repetitions.

Table 5: 90% and 95% critical values from simulated $\hat{f}$ distribution where $n = 4$ and underlying $\gamma = 0$ and $\alpha = 0$

Students	μ	$\hat{f}_{0.90}$ CV	$\hat{f}_{0.95}$ CV	$\hat{\mu}$ 95% CI	
15	0.1	0.267	0.356	-0.156	0.467
15	0.2	0.267	0.267	-0.156	0.556
15	0.3	0.267	0.267	-0.067	0.644
15	0.4	0.178	0.267	0.022	0.733
15	0.5	0.178	0.267	0.200	0.822
15	0.6	0.178	0.178	0.289	0.911
15	0.7	0.178	0.178	0.378	0.911
15	0.8	0.089	0.178	0.556	1.000
30	0.1	0.178	0.222	-0.111	0.333
30	0.2	0.178	0.222	-0.022	0.422
30	0.3	0.133	0.222	0.067	0.556
30	0.4	0.133	0.178	0.156	0.644
30	0.5	0.133	0.178	0.244	0.733
30	0.6	0.133	0.133	0.378	0.822
30	0.7	0.089	0.133	0.511	0.867
30	0.8	0.089	0.089	0.600	0.956
50	0.1	0.133	0.187	-0.067	0.280
50	0.2	0.133	0.160	0.013	0.387
50	0.3	0.133	0.160	0.120	0.493
50	0.4	0.107	0.133	0.227	0.573
50	0.5	0.107	0.133	0.307	0.680
50	0.6	0.080	0.107	0.413	0.760
50	0.7	0.080	0.107	0.547	0.840
50	0.8	0.053	0.080	0.653	0.920
80	0.1	0.117	0.133	-0.033	0.233
80	0.2	0.100	0.133	0.067	0.350
80	0.3	0.100	0.133	0.150	0.450
80	0.4	0.083	0.117	0.250	0.550
80	0.5	0.083	0.100	0.350	0.633
80	0.6	0.067	0.100	0.467	0.733
80	0.7	0.067	0.083	0.567	0.817
80	0.8	0.050	0.067	0.700	0.900
100	0.1	0.093	0.120	-0.013	0.227
100	0.2	0.093	0.120	0.080	0.333
100	0.3	0.093	0.107	0.173	0.427
100	0.4	0.080	0.107	0.267	0.533
100	0.5	0.067	0.093	0.373	0.627
100	0.6	0.067	0.080	0.480	0.720
100	0.7	0.053	0.067	0.587	0.800
100	0.8	0.040	0.053	0.707	0.893
300	0.1	0.058	0.076	0.031	0.173
300	0.2	0.053	0.071	0.124	0.276
300	0.3	0.053	0.067	0.222	0.378
300	0.4	0.049	0.062	0.324	0.476
300	0.5	0.044	0.053	0.427	0.573
300	0.6	0.040	0.049	0.529	0.667
300	0.7	0.036	0.044	0.636	0.760
300	0.8	0.027	0.036	0.742	0.853

Notes: Each row indicates the critical values at the 90% threshold ($\hat{f}_{0.90}$ CV) and 95% threshold ($\hat{f}_{0.95}$ CV). The 95% confidence interval of the estimated $\hat{\mu}$ values ($\hat{\mu}$ 95% CI) are provided to assist in finding the appropriate $\hat{f}$ critical value. A much larger set of critical values are available at https://goo.gl/mjIYvN. All simulations run with 10,000 repetitions.

Table 6: 90% and 95% critical values from simulated $\hat{\gamma}$ distribution where $n = 5$ and underlying $\gamma = 0$ and $\alpha = 0$

Students	μ	$\hat{\gamma}_{0.90}$ CV	$\hat{\gamma}_{0.95}$ CV	$\hat{\mu}$ 95% CI	
15	0.1	0.188	0.250	-0.167	0.417
15	0.2	0.167	0.229	-0.083	0.500
15	0.3	0.167	0.229	0.000	0.583
15	0.4	0.146	0.208	0.083	0.669
15	0.5	0.146	0.188	0.167	0.833
15	0.6	0.125	0.167	0.333	0.835
15	0.7	0.104	0.146	0.417	0.917
15	0.8	0.083	0.125	0.583	1.000
30	0.1	0.125	0.167	-0.083	0.292
30	0.2	0.125	0.167	0.000	0.417
30	0.3	0.115	0.156	0.083	0.501
30	0.4	0.104	0.135	0.167	0.625
30	0.5	0.094	0.125	0.292	0.708
30	0.6	0.083	0.115	0.375	0.792
30	0.7	0.073	0.104	0.500	0.875
30	0.8	0.063	0.083	0.625	0.958
50	0.1	0.100	0.131	-0.050	0.250
50	0.2	0.094	0.125	0.050	0.375
50	0.3	0.088	0.119	0.125	0.475
50	0.4	0.081	0.106	0.225	0.575
50	0.5	0.075	0.094	0.325	0.675
50	0.6	0.069	0.088	0.425	0.750
50	0.7	0.056	0.075	0.550	0.850
50	0.8	0.050	0.063	0.675	0.925
80	0.1	0.074	0.098	-0.016	0.234
80	0.2	0.074	0.098	0.078	0.328
80	0.3	0.066	0.090	0.172	0.438
80	0.4	0.063	0.082	0.266	0.531
80	0.5	0.055	0.078	0.359	0.625
80	0.6	0.051	0.070	0.469	0.719
80	0.7	0.043	0.059	0.578	0.813
80	0.8	0.035	0.051	0.688	0.891
100	0.1	0.069	0.091	-0.013	0.213
100	0.2	0.066	0.084	0.088	0.313
100	0.3	0.063	0.078	0.175	0.425
100	0.4	0.056	0.072	0.275	0.525
100	0.5	0.050	0.069	0.375	0.625
100	0.6	0.047	0.059	0.488	0.713
100	0.7	0.041	0.053	0.600	0.800
100	0.8	0.034	0.044	0.713	0.888
300	0.1	0.041	0.053	0.038	0.163
300	0.2	0.038	0.049	0.133	0.267
300	0.3	0.035	0.046	0.229	0.371
300	0.4	0.033	0.043	0.329	0.471
300	0.5	0.029	0.039	0.429	0.567
300	0.6	0.026	0.034	0.533	0.667
300	0.7	0.023	0.030	0.642	0.758
300	0.8	0.019	0.024	0.746	0.850

Notes: Each row indicates the critical values at the 90% threshold ($\hat{\gamma}_{0.90}$ CV) and 95% threshold ($\hat{\gamma}_{0.95}$ CV). The 95% confidence interval of the estimated $\hat{\mu}$ values ($\hat{\mu}$ 95% CI) are provided to assist in finding the appropriate $\hat{\gamma}$ critical value. A much larger set of critical values are available at https://goo.gl/mjIYvN. All simulations run with 10,000 repetitions.

Table 7: 90% and 95% critical values from simulated $\hat{f}$ distribution where $n = 5$ and underlying $\gamma = 0$ and $\alpha = 0$

Students	μ	$\hat{f}_{0.90}$ CV	$\hat{f}_{0.95}$ CV	$\hat{\mu}$ 95% CI	
15	0.1	0.250	0.250	-0.167	0.417
15	0.2	0.167	0.250	-0.083	0.500
15	0.3	0.167	0.250	0.000	0.583
15	0.4	0.167	0.250	0.083	0.669
15	0.5	0.167	0.167	0.167	0.833
15	0.6	0.167	0.167	0.333	0.835
15	0.7	0.083	0.167	0.417	0.917
15	0.8	0.083	0.167	0.583	1.000
30	0.1	0.167	0.208	-0.083	0.292
30	0.2	0.167	0.208	0.000	0.417
30	0.3	0.125	0.167	0.083	0.501
30	0.4	0.125	0.167	0.167	0.625
30	0.5	0.125	0.167	0.292	0.708
30	0.6	0.083	0.125	0.375	0.792
30	0.7	0.083	0.125	0.500	0.875
30	0.8	0.083	0.083	0.625	0.958
50	0.1	0.125	0.150	-0.050	0.250
50	0.2	0.125	0.150	0.050	0.375
50	0.3	0.100	0.150	0.125	0.475
50	0.4	0.100	0.125	0.225	0.575
50	0.5	0.100	0.125	0.325	0.675
50	0.6	0.075	0.100	0.425	0.750
50	0.7	0.075	0.100	0.550	0.850
50	0.8	0.050	0.075	0.675	0.925
80	0.1	0.094	0.125	-0.016	0.234
80	0.2	0.094	0.109	0.078	0.328
80	0.3	0.078	0.109	0.172	0.438
80	0.4	0.078	0.094	0.266	0.531
80	0.5	0.078	0.094	0.359	0.625
80	0.6	0.063	0.078	0.469	0.719
80	0.7	0.063	0.078	0.578	0.813
80	0.8	0.047	0.063	0.688	0.891
100	0.1	0.088	0.113	-0.013	0.213
100	0.2	0.088	0.100	0.088	0.313
100	0.3	0.075	0.100	0.175	0.425
100	0.4	0.063	0.088	0.275	0.525
100	0.5	0.063	0.088	0.375	0.625
100	0.6	0.063	0.075	0.488	0.713
100	0.7	0.050	0.063	0.600	0.800
100	0.8	0.038	0.050	0.713	0.888
300	0.1	0.050	0.067	0.038	0.163
300	0.2	0.046	0.058	0.133	0.267
300	0.3	0.046	0.058	0.229	0.371
300	0.4	0.042	0.054	0.329	0.471
300	0.5	0.038	0.050	0.429	0.567
300	0.6	0.033	0.042	0.533	0.667
300	0.7	0.029	0.038	0.642	0.758
300	0.8	0.025	0.029	0.746	0.850

Notes: Each row indicates the critical values at the 90% threshold ($\hat{f}_{0.90}$ CV) and 95% threshold ($\hat{f}_{0.95}$ CV). The 95% confidence interval of the estimated $\hat{\mu}$ values ($\hat{\mu}$ 95% CI) are provided to assist in finding the appropriate $\hat{f}$ critical value. A much larger set of critical values are available at https://goo.gl/mjIYvN. All simulations run with 10,000 repetitions.