7-27-2021

On Guessing: An Alternative Adjusted Positive learning Estimator and Comparing Probability Misspecification with Monte Carlo Simulations

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

Dustin R. White
University of Nebraska at Omaha

Recommended Citation

Smith BO, White DR. On Guessing: An Alternative Adjusted Positive Learning Estimator and Comparing Probability Misspecification With Monte Carlo Simulations. Applied Psychological Measurement. 2021;45(6):441-458. doi:10.1177/01466216211013905

On Guessing: An Alternative Adjusted Positive Learning Estimator and Comparing Probability Misspecification with Monte Carlo Simulations

Ben O. Smith and Dustin R. White

University of Nebraska at Omaha

Author Note

Author contact info: 6708 Pine St, Omaha, NE 68106; P: 402.554.4816; bosmith@unomaha.edu. The authors thank Michael O'Hara, Ignacio Sarmiento-Barbieri, and Brandon Sheridan for their helpful comments on earlier drafts of this paper. The authors thank John Donoghue, Brian Habing, and the anonymous referees for their helpful comments during the peer review process.

Abstract

Practitioners in the sciences have used the 'flow' of knowledge (post-test score minus pre-test score) to measure learning in the classroom for the past fifty years. Walstad and Wagner (2016) and Smith and Wagner (2018) moved this practice forward by disaggregating the flow of knowledge and accounting for student guessing. These estimates are sensitive to misspecification of the probability of guessing correct. This work provides guidance to practitioners and researchers facing this problem. We introduce a transformed measure of true positive learning that under some knowable conditions performs better when students' ability to guess correctly is misspecified and converges to Hake's (1998) normalized learning gain estimator under certain conditions. We then use simulations to compare the accuracy of two estimation techniques under various violations of the assumptions of those techniques. Using recursive partitioning trees fitted to our simulation results, we provide the practitioner concrete guidance based on a set of yes/no questions.

Keywords: Disaggregated learning, Gain measurement, Value-added learning, Monte Carlo Simulation

On Guessing: An Alternative Adjusted Positive Learning Estimator and Comparing Probability Misspecification with Monte Carlo Simulations

Introduction

For the past half-century, educators in the sciences (e.g. economics, physics, etc.) have used pre- and post-tests to measure the 'flow' of knowledge (Siegfried & Fels, 1979), or aggregate gains in learning over time. Flow-of-knowledge measures play an important role in program/course assessment, instructor self improvement, and in estimating the impact of educational treatments as they have a closer connection to the total learning that occurred during a treatment than stock measures (such as a post-treatment exam). In determining the effectiveness of an intervention or program, the practitioner[1] is usually interested in the increase in student knowledge that occurred because of the intervention. As the effect of the treatment is not directly observable, the increase in student knowledge (the flow) from the beginning of the treatment period (e.g. class or program) to the end of the treatment period is often used instead.

Attempting to measure knowledge increases without accounting for prior knowledge can lead to unreliable measures of student learning. Prior student knowledge can vary widely from one section of a course to another (as the sample is the class or program size). Despite this hurdle, it is necessary to accurately measure student knowledge gains in such settings, as these measurements drive education intervention research, changes that individual faculty might make to a course, and the perceived strengths and weaknesses of an academic program. In its most basic formulation, the practitioner or researcher can find the flow of knowledge by subtracting the matched pre-test scores from the post-test scores (students who did not take both the pre- and post-test are removed from the analysis). While differencing the pre- and post-tests is a common approach, other methods include regression analysis with the pre-test as an independent variable and differencing with a normalization factor (for instance, the Hake (1998) estimator).

Walstad and Wagner (2016) improved this practice by suggesting that the practitioner or

[1] Throughout this paper, we will often use the term "practitioner" to describe an instructor or department interested in measuring learning in one or more classes.

researcher should disaggregate learning into four types: positive, negative, zero, and retained. Positive learning is said to occur when the student answered incorrectly on the pre-test then answered correctly on the post-test. Negative learning is said to occur when the student answered correctly on the pre-test but then answered incorrectly on the post-test. Zero learning is said to occur when the student answered incorrectly both times and retained learning is said to occur when the student answered correctly both times.

While many fields have generated 2x2 tables in order to study specific subsets of populations based on outcomes,[2] the insight of Walstad and Wagner (2016) was that these disaggregated learning types represent different outcomes and should not be intermixed; a practitioner should react very differently to an increase in positive versus a decrease in negative learning (an increase in positive learning is likely due to an improved pedagogical technique while a decrease in negative learning is likely due to the removal of a confusing explanation). Despite this fundamental difference, the flow of knowledge is measured as positive learning minus negative learning. This insight has lead to rapid adoption of the new procedure in research and in assessment. For instance, Emerson and English (2016) used positive learning to estimate the impact of educational treatments. Happ, Zlatkin-Troitschanskaia, and Schmidt (2016) measured the relationship between positive learning and characteristics such as native language and GPA. To enable the adoption of this technique when assessing many courses, Smith (2018) developed software that produced the disaggregated values directly from Scantron formatted exam output files.

The work by Walstad and Wagner (2016) had one notable shortcoming: it did not account for student guessing. Guessing on a multiple choice exam decreases the accuracy of the instrument (Zimmerman & Williams, 2003), and multiple choice exams are in widespread use, especially in lower-level courses. Smith and Wagner (2018) showed that guessing masks the true learning values and creates bias in the expected values of the unadjusted learning types suggested by Walstad and Wagner (2016). However, the paper also showed that the unadjusted learning

[2] See, for example disaggregations like those provided in Figure 1 of Swets (1969).

values could be adjusted to account for guessing when the probability of guessing correct is known. The critical adjusted measures are $\hat{\gamma}$ (true positive learning) and $\hat{\alpha}$ (true negative learning). When determining if a pedagogical technique is effective or when measuring the learning in a class, $\hat{\gamma}$ is the value of interest. (Throughout this work we will use hats to emphasize when we are discussing an estimate of a learning parameter and not the parameter itself.) This modified approach can be applied in educational research. Even more valuable is the improved accuracy it provides to assessment procedures (the software in Smith (2018) also produces these adjusted measures).

Smith and Wagner (2018) adjusted the results from the national-norming sample for the micro- and macroeconomics *Test of Understanding of College Economics* (Walstad, Watts, & Rebeck, 2007) or TUCE. The paper estimated the probability of guessing correct using a workhorse in the psychometrics literature: the Three Parameter Logistic or 3PL (De Ayala, 2013). The 3PL estimates the probability of a very low ability student to answer a given question correct despite not knowing the correct answer to a question. In the literature, this value is referred to as the pseudoguessing parameter. This method can detect when students are able to remove distractors from consideration; many well written questions have point estimates for the pseudoguessing parameter statistically indistinguishable from $1/n$, where n is the number of question options. While this is a reasonable method for the TUCE data, the 3PL cannot be applied to most classes.

The 3PL procedure requires a tremendous amount of data to converge; researchers disagree about the precise minimum number but 1000 students with twenty questions seems to be the median recommendation (De Ayala, 2013, pp. 130-131). Even when this observational threshold is satisfied, the 3PL estimator will only converge on truth under select conditions.[3] This convergence problem is part of the reason Han (2012) suggested fixing the pseudoguessing parameter at $1/n$ to estimate the other parameters in the model.

---

[3] Notably, Orlando and Thissen (2003) show questions of non-monotonic form can be problematic when estimating the monotonic 3PL.

Relationship to General Educational Measurement Literature

While the consideration of guessing in evaluating learning is a relatively new feature in the Economic Education literature, it has been a core element of Educational Measurement for at least 50 years. Birnbaum (1968) extended existing models by statistically accounting for guessing.[4] Later, Maris (1999) elaborated that students could have a partial mastery of question content that changes the probability of answering a question correct. These mental resources may include skills such as educated guessing (partial knowledge), elimination techniques, or other methods of increasing the likelihood of success.

Educated guessing and other techniques to utilize partial knowledge become particularly important where the stakes of the test are not sufficiently high. Wise and DeMars (2006) presents a model that is able to accommodate and adjust for low levels of effort. Under low effort, many students will seek to simply answer questions quickly, either due to indifference or fatigue, and will thereby generate inaccurate estimates of learning or understanding. Accounting for low effort is critical when educators seek to use scores from low-stakes problems (pilot questions given without stakes, small homework assignments, etc.) to project scores or results in high-stakes contexts like certification exams or final exams. The model proposed by Wise and DeMars (2006) can outperform the 3PL model in some cases. Another way for educators to overcome the low-stakes problem is simply to raise the stakes. Wise and DeMars (2005) proposes several methods to make a low-stakes test high-stakes for both the educator as well as the test-takers, including providing incentives for good performance and prompt feedback.

Finally, Gönülateş and Kortemeyer (2017) show that Item Response Theory can be adapted to account for a student's propensity to guess using multidimensional Item Response Theory (MIRT). This approach, like the approach of Wise and DeMars (2006), can be utilized to generate more accurate forecasts from low-stakes settings like homework assignments to higher stakes examinations. Gönülateş and Kortemeyer (2017) emphasize that controlling for more learner

[4] In Item Response Theory, questions are labeled items. We will proceed to call them questions in order to maintain the clarity of the manuscript.

traits will improve models, but mostly in low-stakes settings.

In our work below, we assume that questions are presented in a high-stakes setting, and that educators can accurately state how many effective discriminators exist for a given question. In the context of evaluating homework questions or other relatively low-stakes questions, these assumptions may not be true. In those cases, we urge the reader to refer to the works cited in this section for a more complete understanding of how guessing and partial knowledge affect our understanding of learning measurements.

Contribution to the Literature

This paper extends the work of Walstad and Wagner (2016) and Smith and Wagner (2018) in a number of ways. First, we suggest a transformation of the $\hat{\gamma}$ learning estimator: $\hat{\gamma}/(1-\hat{\mu})$. We show that this transformation (what we call the gain estimator) performs better than the original estimator under conditions knowable to the practitioner or researcher; in some cases, it has a superior interpretation. Moreover, we extend the Monte Carlo simulations of no learning by (Smith & Wagner, 2018, pp. 5-9) to include the the gain estimator. Therefore, the practitioner or researcher can perform the same statistical test regardless of if they use the $\hat{\gamma}$ or gain estimator.

Second, this paper suggests that when the probability of guessing correct on a set of exam questions does not converge to $1/n$, assuming true negative learning is zero could be a superior option. Throughout the paper the modified estimators given this assumption are provided. The gain estimator under the assumption that true negative learning is zero results in Hake's (1998) normalized learning gain – a common learning value used in STEM fields. Therefore, this paper provides a clarified interpretation of previous research that have used the Hake learning gain.

Third, under strong assumptions, we use a large number of Monte Carlo simulations to compare deviations from the true value using the estimated $\hat{\gamma}$ and $\hat{\gamma}/(1-\hat{\mu})$ when assuming the probability equals $1/n$ or setting estimated true negative learning to zero. Using these simulation results we make three general observations and fit the results to decision trees (in Online Appendix A). These trees represent the collective guidance of all of the Monte Carlo simulations

	Correct (Post)	Incorrect (Post)
Correct (Pre)	$\hat{\text{rl}}$	$\hat{\text{nl}}$
Incorrect (Pre)	$\hat{\text{pl}}$	$1 - \hat{\text{pl}} - \hat{\text{nl}} - \hat{\text{rl}}$

Figure 1. Mapping of responses to raw learning types as described by Walstad and Wagner (2016). This original disaggregation is used in all calculations in Smith and Wagner (2018) and this paper. Note that Post-Test $= \hat{\text{pl}} + \hat{\text{rl}}$ and Pre-Test $= \hat{\text{nl}} + \hat{\text{rl}}$.

in a visual form and are critical to providing practitioners guidance.

Finally, we provide a 'practitioner's guide' to the results in this paper. This section walks the practitioner through the process of choosing a learning estimate ($\hat{\gamma}$ or $\hat{\gamma}/(1-\hat{\mu})$), selecting the best method of specifying the probability of guessing correct ($1/n$ or $\hat{\alpha} = 0$), and performing a statistical test against the null hypothesis of no learning. The practitioner who is only interested in applying the resulting techniques in this paper can read the practitioner's guide without first reading the preceding sections.

This paper will proceed as follows: (1) In *A Brief Description of the Estimators in Smith and Wagner (2018)* we review the estimators in Smith and Wagner (2018) and provide those estimators when estimated true negative learning is zero; (2) In *The Gain Measurement* we will propose a transformed measure of positive learning; (3) In *Comparing the Adjusted Positive Learning and the Gain Estimators' Sensitivity to Probability Misspecification* we will show under what conditions the gain estimator is less sensitive than the original estimator to probability misspecification; (4) In *A Statistical Test of the Gain Measurement using a Counterfactual Monte Carlo Simulation* we will provide a statistical test for the gain estimator equivalent to what was provided by Smith and Wagner (2018) for the original estimator; (5) In *Comparative Monte Carlo Simulations* we compare deviations from the true learning values with Monte Carlo simulations; and (6) In *Practitioner's Guide* we provide instructions on how to use the results in this paper. We then conclude the paper.

A Brief Description of the Estimators in Smith and Wagner (2018)

In Smith and Wagner (2018) the authors develop estimators of the true underlying learning parameters using the raw estimates described in Walstad and Wagner (2016). The estimators are as follows:

$$\hat{\gamma} = \frac{\hat{p}(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}}{(\hat{p} - 1)^2} \tag{1}$$

$$\hat{\alpha} = \frac{\hat{p}(\hat{\text{pl}} + \hat{\text{rl}} - 1) + \hat{\text{nl}}}{(\hat{p} - 1)^2} \tag{2}$$

$$\hat{\mu} = \frac{\hat{\text{nl}} + \hat{\text{rl}} - \hat{p}}{1 - \hat{p}} \tag{3}$$

where $\hat{\text{pl}}$ is raw or unadjusted positive learning, $\hat{\text{nl}}$ is unadjusted negative learning, and $\hat{\text{rl}}$ is unadjusted retained learning (Figure 1). $\hat{p}$ is the estimated probability of a student who does not know the correct answer answering correct nonetheless (for the estimates to be valid, the probability of guessing correctly cannot equal one).[5] $\hat{\gamma}$ is the estimate of true positive learning, $\hat{\alpha}$ is the estimate of true negative learning,[6] and $\hat{\mu}$ is the estimate of incoming stock knowledge (what proportion of students knew the answer when they took the pre-test). Therefore, the estimate of true retained learning would be $\hat{\mu} - \hat{\alpha}$ and the estimate of true zero learning would be $1 - \hat{\mu} - \hat{\gamma}$ (for the reader's convenience, we provide a reference of our notation in Figure 2). As the underlying learning values must sum to one, the bounds of learning values depend on each other. Notably, γ can not exceed $1 - \mu$ as μ proportion of students know the question at the time

[5] In Smith and Wagner (2018), the authors describe the probability of guessing correct as $1/n$. As a result, their estimators are filled with n's instead of $\hat{p}$'s. For notational convenience in the next section, we describe the estimators in terms of $\hat{p}$. The estimators are mathematically identical.

[6] Unadjusted negative learning is disaggregation of *performance* which can be generated through multiple mechanisms including guessing, forgetting, and actively learning incorrect information to replace correct information. True negative learning removes guessing in the expectation but still measures negative learning without regard to how it was generated. In practice, people often refer to negative learning as "forgetting" even though that is not completely accurate.

Learning Type	Unadjusted Notation	Adjusted for Guessing Notation
Positive Learning	$\hat{\text{pl}}$	$\hat{\gamma}$
Negative Learning	$\hat{\text{nl}}$	$\hat{\alpha}$
Retained Learning	$\hat{\text{rl}}$	$\hat{\mu} - \hat{\alpha}$
Zero Learning	$1 - \hat{\text{pl}} - \hat{\text{nl}} - \hat{\text{rl}}$	$1 - \hat{\mu} - \hat{\gamma}$
Stock Knowledge	$\hat{\text{rl}} + \hat{\text{nl}}$	$\hat{\mu}$

(a) Notation for learning types in the paper

Meaning	Notation
Probability of a student guessing correct	$\hat{p}$
Number of options on a multiple choice exam	n

(b) Other notation used in the paper

Figure 2. This figure provides notation used throughout the paper for the convenience of the reader. Estimates of an underlying parameter are always presented with a hat (e.g. $\hat{\gamma}$). If we are referring to the underling parameter itself, it will be presented without a hat (e.g. γ).

of the pre-test (this bound holds true of the estimated values as well).

Assuming $\hat{p}$ is properly specified, with an infinite number of observations $\hat{\gamma}$, $\hat{\alpha}$, and $\hat{\mu}$ converge to the true underlying parameters of γ, α, and μ. However, there is no guarantee $\hat{p}$ is properly specified. In the authors' experience, it seems that p is often near $1/n$ on nationally-normed exam questions or other questions where the students are unable to remove distractors or otherwise modify the probability of guessing correct beyond pure chance. However, not all exam questions meet this criteria; for brevity we will refer to questions where the probability of guessing correct is substantively different than $1/n$ as "guessing-probability-deviating questions."

A reasonable alternative to $1/n$ is to set the estimated true negative learning ($\hat{\alpha}$) to zero and solve for the implied probability. Intuitively, true negative learning would occur when students forgot material they learned prior to the class/program and the class/program instruction did not resurrect that knowledge. Alternatively, it can occur when the instructor's explanation of a concept is inaccurate or otherwise confuses a student that previously had a solid understanding of a concept. In practice, true negative learning appears to be rare.

In Smith and Wagner (2018), the estimated $\hat{\alpha}$ values were very low when using the TUCE national-norming sample (Walstad et al., 2007). This suggests true negative learning is quite rare in the sample. Additionally, as the authors have applied the method in Smith and Wagner (2018) to numerous classes for assessment purposes, $\hat{\alpha}$ values are nearly always below 0.05 and trend

towards zero. Therefore, while assuming zero true negative learning is an incorrect assumption, it might be a useful assumption when $1/n$ could be substantially incorrect due to guessing-probability-deviating questions. Further, as we will show on page 12, this assumption has been implicitly made by many authors working in STEM education over the last twenty years.

Setting equation 2 equal to zero and solving for $\hat{p}$ reveals the implied probability in equation 4 (derivations of all new equations presented in this paper are in Online Appendix C).

$$\hat{p} = \frac{\hat{\text{nl}}}{1 - \hat{\text{pl}} - \hat{\text{rl}}} \tag{4}$$

Notably, when unadjusted zero learning equals zero ($1 - \hat{\text{pl}} - \hat{\text{rl}} - \hat{\text{nl}} = 0$) the implied probability equals one (which cannot be true). Conceptually, the way unadjusted zero learning could equal zero is if students can guess correct with 100% probability. Naturally, the accuracy of this estimate will vary greatly and will at times suggest unrealistic probabilities through randomness alone; for the estimates presented here to be correct, $\hat{p}$ cannot equal one. This point will be made clear in the simulation section of this paper.

Assuming adjusted negative learning equals zero, we can substitute equation 4 into equation 1 to find our $\hat{\gamma}$ estimator given $\hat{\alpha} = 0$.

$$(\hat{\gamma}|\hat{\alpha} = 0) = \frac{(\hat{\text{pl}} - \hat{\text{nl}})(1 - \hat{\text{pl}} - \hat{\text{rl}})}{\underbrace{1 - \hat{\text{nl}} - \hat{\text{pl}} - \hat{\text{rl}}}_{\text{Zero Learning}}} \tag{5}$$

One can see the impact of $\hat{p}$ on equation 5: the estimator is undefined as the denominator is zero when unadjusted zero learning equals zero (when the implied probability of guessing correct, $\hat{p}$, equals one). Naturally, the accuracy of this measure will depend on the accuracy of $\hat{p}$. Nonetheless, as we will show in the upcoming simulation, this measure can be preferable when the $1/n$ estimate is particularly unreasonable. This would occur when myopic students have a different than chance probability of guessing correctly (e.g. the ability to eliminate a distractor or confusion caused by the wording of a question or distractor).

The Gain Measurement

As noted in the previous section, γ has an upper bound at $1-\mu$. Therefore, the maximum observable $\hat{\gamma}$ value is question and population dependent. This has an interpretation disadvantage where learning measurements, even for the same question, cannot be compared across student populations if the incoming stock knowledge (μ) differs. In this section, we present a simple transformation that resolves this interpretation disadvantage. Further, under some knowable conditions, this measure is less sensitive to probability misspecification. To be clear, this transformation is not always preferable to the $\hat{\gamma}$ estimator. However, given data, it is knowable if the transformation is preferable.

While γ is the proportion of students that learned ($\gamma \in [0, 1-\mu]$), $\gamma/(1-\mu)$ is the proportion of students who learned the material who could have possibly learned the material (i.e. they didn't already know it). This measure of positive learning has consistent bounds of $[0, 1]$ regardless of stock knowledge. Using equations 1 and 3, the gain measurement is revealed to be the equation below:

$$\frac{\hat{\gamma}}{1-\hat{\mu}} = \frac{\hat{p}(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}}{(\hat{p}-1)(\hat{\text{nl}} + \hat{\text{rl}} - 1)} \tag{6}$$

When $1/n$ might be a particularly poor estimate of p, one may set adjusted negative learning to zero and solve for $\hat{p}$. The probability implied by equation 4 suggests the gain estimator simplifies to equation 7.

$$(\frac{\hat{\gamma}}{1-\hat{\mu}} | \hat{\alpha} = 0) = \frac{\hat{\text{pl}} - \hat{\text{nl}}}{1 - \hat{\text{nl}} - \hat{\text{rl}}} \tag{7}$$

Equation 7 is mathematically equivalent to $(\text{Post-Test} - \text{Pre-Test})/(1 - \text{Pre-Test})$. (Note that Post-Test $= \hat{\text{pl}} + \hat{\text{rl}}$ and Pre-Test $= \hat{\text{nl}} + \hat{\text{rl}}$). Hake (1998) refers to this measure as a 'normalized learning gain;' this measure has been used in many STEM studies including Colt, Davoudi, Murgu, and Rohani (2011), Hamne and Bernhard (2000), and Supasorn (2015). With this work, we show that this measure can be interpreted as the adjusted-for-guessing gain measure

when assuming zero true negative learning.

Comparing the Adjusted Positive Learning and the Gain Estimators' Sensitivity to Probability Misspecification

To determine the sensitivity of the estimators, we will use a common tool in the field of economics: elasticity. Likely the most familiar elasticity is price elasticity of demand, which measures how sensitive a good or service is to changes in price. Formally it is the percent change in quantity divided by the percent change in price. This basic equation can be rearranged as follows: $\frac{\%\Delta Q}{\%\Delta P} = \frac{\Delta Q/Q}{\Delta P/P} = \frac{\Delta Q}{\Delta P}\frac{P}{Q}$. $\frac{\Delta Q}{\Delta P}$ is the change in quantity divided by the change in price; simply put, it is the derivative of the quantity function with respect to price. Therefore, we can re-express the price elasticity of demand as $Q'(P)\frac{P}{Q}$.

In this section, we will develop probability misspecification elasticities. That is, in percent terms, how responsive the estimate is to the probability of guessing correct being misspecified; mathematically, it is the percent change in the estimate divided by the percent change in the probability misspecification. All else equal, a 'less elastic' estimator is desirable as it will be closer to the true learning value when the probability of guessing correct is misspecified.

To compare the sensitivity of the two estimators of interest ($\hat{\gamma}$ and $\hat{\gamma}/(1-\hat{\mu})$ – equations 1 and 6), we will make two notional changes. First, we will describe the specified probability ($\hat{p}$) as $p+\Delta$, where p is the true probability of guessing correct and Δ is the specification error of the success rate of guessing (which can be positive or negative). Second, to simplify the notation later in this section, we will describe the gain estimator as the function $g(\Delta)$ (where the partial derivative with respect to Δ would be described as $g'(\Delta)$) and the adjusted positive learning estimator as the function $f(\Delta)$ (where the partial derivative with respect to Δ would be described as $f'(\Delta)$). With these two notation changes, our estimators can be specified as follows:

$$\begin{aligned} \frac{\hat{\gamma}}{1-\hat{\mu}} &= \frac{(p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(p+\Delta-1)(\hat{\text{nl}}+\hat{\text{rl}}-1)} = g(\Delta) \\ \hat{\gamma} &= \frac{(p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(p+\Delta-1)^2} = f(\Delta) \end{aligned} \tag{8}$$

As the bounds of the estimators differ, we cannot compare the impact of Δ on the estimators directly; we will calculate elasticities. Here we will describe $\varepsilon_{\hat{\gamma}/(1-\hat{\mu})}$ as the probability misspecification elasticity of the gain estimator. Similarly, $\varepsilon_{\hat{\gamma}}$ is the probability misspecification elasticity of the $\hat{\gamma}$ estimator. These elasticities are presented in equation 9:

$$\begin{aligned} \varepsilon_{\hat{\gamma}/(1-\hat{\mu})} &= g'(\Delta)\frac{\Delta}{g(\Delta)} = -\frac{\Delta(\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1)}{(\Delta + p - 1)((\hat{\text{nl}} + \hat{\text{rl}} - 1)(\Delta + p) + \hat{\text{pl}})} \\ \varepsilon_{\hat{\gamma}} &= f'(\Delta)\frac{\Delta}{f(\Delta)} = -\frac{\Delta(\hat{\text{nl}}(\Delta + p + 1) + 2\hat{\text{pl}} + (\hat{\text{rl}} - 1)(\Delta + p + 1))}{(\Delta + p - 1)((\hat{\text{nl}} + \hat{\text{rl}} - 1)(\Delta + p) + \hat{\text{pl}})} \end{aligned} \tag{9}$$

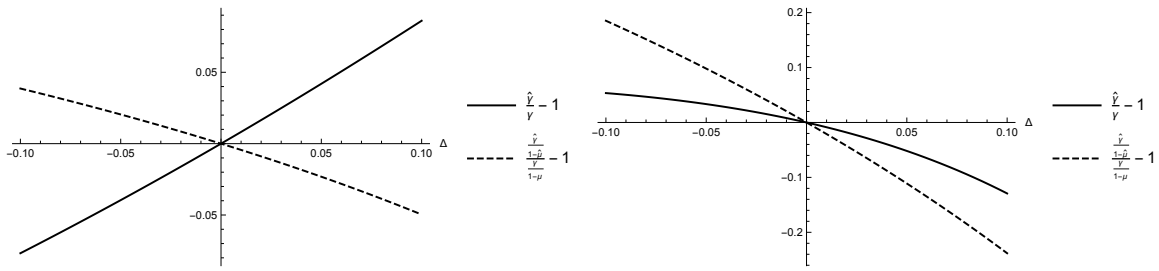
Dividing the equations presented above results in the ratio of elasticities (equation 10):

$$R = \frac{\varepsilon_{\hat{\gamma}/(1-\hat{\mu})}}{\varepsilon_{\hat{\gamma}}} = \frac{\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1}{2\hat{\text{pl}} + (\hat{\text{nl}} + \hat{\text{rl}} - 1)(\Delta + p + 1)} \tag{10}$$

When this ratio (R) is between -1 and 1, the gain estimator is less sensitive to a probability misspecification. When the value is outside of that range, the original $\hat{\gamma}$ estimator is less sensitive to probability misspecification. Concrete examples of this behavior are demonstrated in Figure 3. Notably, while p and Δ are not known, $p + \Delta$ is the value specified by the practitioner or researcher ($\hat{p}$). R is therefore a knowable value for any given dataset, since p and Δ never occur except as the summed value specified by the practitioner or researcher. If the practitioner or researcher is assuming $\hat{\alpha} = 0$, then this ratio simplifies to:

$$(R|\hat{\alpha} = 0) = \frac{\hat{\text{pl}} + \hat{\text{rl}} - 1}{2\hat{\text{pl}} + \hat{\text{rl}} - \hat{\text{nl}} - 1} \tag{11}$$

Whether $(R|\hat{\alpha} = 0)$ is between -1 and 1 depends on the relative size of $\hat{\text{pl}}$ and $\hat{\text{nl}}$. On the positive side of the range, if $\hat{\text{pl}}$ is at least twice as large as $\hat{\text{nl}}$ then the gain measurement will always be less sensitive to probability misspecification. Regardless, this ratio, like R in general, can be calculated by the practitioner or researcher given data.



(a) $\hat{\text{pl}} = 0.50, \hat{\text{rl}} = 0.30, \hat{\text{nl}} = 0.07, R = -0.53$ (b) $\hat{\text{pl}} = 0.35, \hat{\text{rl}} = 0.10, \hat{\text{nl}} = 0.10, R = 2.00$

Figure 3. By comparing $\hat{\gamma}/\gamma - 1$ and $\frac{\hat{\gamma}/(1-\hat{\mu})}{\gamma/(1-\mu)} - 1$ we demonstrate the sensitivity of the estimates to probability misspecification at example learning values; along the x-axis of both graphs is the amount that the estimated probability deviates from the true probability (Δ) and the true probability of guessing correct is $1/5$ ($p = 1/5$) in these graphs. In panel 3(a), $R = -0.53$ (when $\Delta = 0$) with the specified learning values below the figure. This falls into the range of $[-1, 1]$ where the gain estimator is less sensitive to probability misspecification. This can be seen by the steeper slope of $\hat{\gamma}/\gamma - 1$ (solid) in comparison to $\frac{\hat{\gamma}/(1-\hat{\mu})}{\gamma/(1-\mu)} - 1$ (dashed) indicating the misspecified estimator $\hat{\gamma}$ deviates away from the true value at a higher rate than the misspecified $\hat{\gamma}/(1-\hat{\mu})$ estimator. In panel 3(b), $R = 2.00$ (when $\Delta = 0$) with the specified learning values below the figure. In this range the $\hat{\gamma}$ estimator is less sensitive to probability misspecification. This can be seen by the steeper slope of $\frac{\hat{\gamma}/(1-\hat{\mu})}{\gamma/(1-\mu)} - 1$ in comparison to $\hat{\gamma}/\gamma - 1$.

A Statistical Test of the Gain Measurement using a Counterfactual Monte Carlo Simulation

Following an identical methodology to Smith and Wagner (2018), we provide a statistical test for the gain measurement by simulating a counterfactual distribution when there is no true positive or negative learning occurring. Smith and Wagner (2018) created a counterfactual distribution of no learning for the positive learning estimator $\hat{\gamma}$. However, as the gain estimator suggested in this paper ($\hat{\gamma}/(1-\hat{\mu})$) did not exist, it was not included in their critical value tables.

This Monte Carlo simulation replicates student responses to a single question asked on both a pre- and post-test. First, m students are randomly pulled from a population with μ stock knowledge, where m is the number of students in the class. After this random pull, each student either knows or doesn't know the answer to the question (with the proportion of the population knowing the answer equal to μ). The students who know the answer simply answer the question correct on both the pre- and post-test. The students who don't know the answer guess on both the pre- and post-test with $1/n$ probability of guessing correct. Each student's guess on the post-test is independent of their guess on the pre-test, given that the student does not know the correct

answer to the question.

The unadjusted learning types are then calculated and adjusted to produce estimates for positive learning, negative learning, and stock knowledge ($\hat{\gamma}$, $\hat{\alpha}$, and $\hat{\mu}$, respectively). We extend the original simulation by calculating $\hat{\gamma}/(1-\hat{\mu})$ for each simulated class. This process is repeated 10,000 times. The resulting estimated gain measurements are then ordered from lowest to highest and the 90% and 95% critical values are extracted for the simulated empirical distribution. Similarly, the 95% confidence interval for stock knowledge ($\hat{\mu}$) is extracted.

To use these Monte Carlo simulations, the practitioner or researcher should find the set of rows associated with the number of students in their class (m). Then, using their calculated $\hat{\mu}$ value, they should find the set of simulated distributions where their $\hat{\mu}$ value falls in the confidence interval (Online Appendix Table B1 if $\hat{p} = 1/4$ and Online Appendix Table B2 if $\hat{p} = 1/5$). If their calculated gain measurement ($\hat{\gamma}/(1-\hat{\mu})$) exceeds the greatest critical value of the relevant set of rows then the practitioner or researcher can say that their value likely did not occur from randomness alone when there is in fact no learning. Put more simply, the value is statistically different from randomness. We present these tables in Online Appendix B (Tables B1 if $\hat{p} = 1/4$ and Table B2 if $\hat{p} = 1/5$; a larger set of tables can be found at `https://goo.gl/kKAySX` and the simulation code is available at `https://goo.gl/zOqlDx`). For each row in the tables, μ (stock knowledge), m (number of students), and n (question options) are specified; all other values are derived as part of the simulation ($\mu \in \{0.1, 0.2, ..., 0.8\}$, $m \in \{15, 20, 25, ..., 50, 60, 70, ..., 100, 150, 200, 250, 300\}$, and $n \in \{3, 4, 5, 6\}$). For a more detailed description of these Monte Carlo simulations, see the *Monte Carlo Simulation of the Counterfactual* section of Smith and Wagner (2018).

Comparative Monte Carlo Simulations

In this section, we perform Monte Carlo simulations to compare the accuracy of assuming the probability of guessing correct equals $1/n$ against the accuracy of assuming that true negative learning equals zero ($\hat{\alpha} = 0$). It is not surprising that setting the probability correctly (by either

method) would result in the superior estimate. Therefore, we present results based on when these assumptions are violated in Online Appendix E. In the case of the $1/n$ estimator, we assume the practitioner or researcher is setting $\hat{p} = 1/4$ or $\hat{p} = 1/5$ but the true probability is simulated from 0.15 to 0.35 (with steps of 0.01). Similarly, with the true negative learning equaling zero estimator, we assume the practitioner or researcher is setting $\hat{\alpha} = 0$, but the true negative learning ranges from 0.01 to 0.05 (with steps of 0.01). These results have been simulated with all combinations of $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\mu \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ conditioned on $\gamma + \mu + \alpha \leq 1$. We have generated results using class sizes of 30, 50, and 100. In total, we have generated 15,120 distributions: 7560 assuming $\hat{p} = 1/4$, 7560 assuming $\hat{p} = 1/5$. The complete set of results are available here: `https://goo.gl/2f8NSa`. The chosen parameters are motivated by the maximum and minimum values observed by the authors in the TUCE dataset as well as data collected through departmental assessment procedures. If the reader wishes to simulate a combination of parameters not provided in the paper, our simulation code is available here: `https://goo.gl/qjCCUj`. The Monte Carlo simulation works as follows:

1. For a given underlying μ (true stock knowledge), each simulated student's stock ability on an exam with 30 equally difficult questions is drawn from a Binomial distribution with probability such that the population $\hat{\mu}$ equals the seeded μ value. From this procedure, each student knows (randomly) some set of questions on the pre-test.

2. For a given underlying γ (true positive learning), each student randomly 'learns' some set of the questions they did not know during the pre-test. This value is determined by a draw from a Binomial distribution such that the population $\hat{\gamma}$ equals the seeded γ value.

3. For a given underlying α (true negative learning), each student randomly 'forgets' some set of the questions they knew at the time of the pre-test. This value is determined by a draw from a Binomial distribution such that the population $\hat{\alpha}$ equals the seeded α value.

4. Based on the random draws above, each student knows or does not know the answer to each question on each of the exams. For the questions they know, they answer correctly.

On the questions they don't know they guess with the seeded value probability (p) of guessing correct. From the resulting simulated exam responses, the raw learning types can be calculated.

5. From the steps above, the pre- and post-test responses are simulated such that the unadjusted learning types can be calculated. From this, true positive learning ($\hat{\gamma}_{1/n}$) and the gain measurement ($\frac{\hat{\gamma}_{1/n}}{1-\hat{\mu}_{1/n}}$) are estimated assuming $\hat{p} = 1/4$ (regardless of the true probability of guessing correct; simulations have also been conducted assuming $\hat{p} = 1/5$). Similarly, true positive learning ($\hat{\gamma}_Z$) and the gain measurement ($\frac{\hat{\gamma}_Z}{1-\hat{\mu}_Z}$) are calculated under the assumption that true negative learning equals zero (regardless of the actual seeded α value).

6. From the steps above we have simulated estimates of the true positive learning and the gain measurement as calculated by a practitioner or researcher. However, we also know the true learning values for the simulated class (γ and $\gamma/(1-\mu)$). The simulation calculates the absolute deviation from the true value of γ for the class when assuming $\hat{p} = 1/4$ ($\left|\hat{\gamma}_{1/n} - \gamma\right|$) and when assuming true negative learning equals zero ($|\hat{\gamma}_Z - \gamma|$). Similarly, the absolute deviation from the true gain measurement ($\gamma/(1-\mu)$) is calculated when assuming $\hat{p} = 1/4$ ($\left|\frac{\hat{\gamma}_{1/n}}{1-\hat{\mu}_{1/n}} - \frac{\gamma}{1-\mu}\right|$) and when assuming true negative learning equals zero ($\left|\frac{\hat{\gamma}_Z}{1-\hat{\mu}_Z} - \frac{\gamma}{1-\mu}\right|$).

7. This process is repeated for 10,000 classes and the average of each of the four measures of absolute deviation are reported for a given set of seeded values. If unadjusted zero learning equals zero in one of the 10,000 simulated classes (which can occur by randomness alone) then the absolute deviation from truth for the $\hat{\gamma}_Z$ estimator is infinite. Therefore, assuming true negative learning is equal to zero is not feasible for that specification; this is indicated with a blank cell in the tables.

The probabilities used to characterize the Binomial distributions are not specifically defined above as they vary with the specified values of μ, γ, and α. That is, for a given population γ value

to be achieved, the specified probability that characterizes the Binomial distribution must increase with an increase in μ as a smaller proportion of the population have the possibility to learn $(1 - \mu)$. Similarly, for a given value of α, the specified probability that characterizes the Binomial distribution decreases with μ as there is a larger group of students who could forget.

Our rationale for the use of the Binomial distribution in the procedure above is that each question on the simulated exam is mapped to a unique bit of knowledge that could be known, learned, or forgotten. If the knowledge is truly independent (and no two questions are mapped to the same bit of knowledge), then the questions can be thought of as a Binomial distribution. By necessity, the distributional assumption is strong. Without assuming a specific distribution (with relatively few parameters), the Monte Carlo simulations are not feasible given that we are simulating all combinations of the input parameters; a targeted sensitivity check of this assumption was performed in Online Appendix D.

In the educational context, the Binomial assumption indicates that we are simulating an exam with equally difficult questions where each question is entirely independent. In many cases, these are unreasonable assumptions. However, many instructors attempt to write exams that contain independent questions; often in the context of program/course assessment for accreditation, the questions are required be independent. The sensitivity check in Online Appendix D is an attempt to partially address this weakness. In that set of simulations we structured the process of learning and forgetting to be highly correlated with student ability (more details in the appendix). The simulations in Online Appendix D show that the simulations presented here are at least somewhat robust to changes in the correlation structure. If the reader wishes to perform a different targeted simulation using a different distribution or correlation structure, they can do so as the pulls from the Binomial distribution occur in three places in the code (which we've made available).

This distribution assumption emphasizes the intended use of these Monte Carlo simulations. Unlike the simulations provided in the previous section, which can be used more generally, our goal here is to provide *guidance* to the practitioner or researcher. For our simulations to be

feasible, our assumptions are necessarily strong; without strong assumptions we would not be able to provide guidance to the practitioner at all. Therefore, the values in the table should only be compared to each other and one should focus on trends instead of any single simulated distribution. We will describe these trends as *observations*.

In addition to the observations included in this paper, we use the full set of simulation results (`https://goo.gl/2f8NSa`) to generate figures A1 and A2 in Online Appendix A. These recursive decision trees (Breiman, Friedman, Stone, & Olshen, 1984) find the splits in the data that minimize error. In this context, the data (all of the simulations) is recursively split by the parameter (and value on a continuous variable) that best predicts whether the practitioner should use the $1/n$ or $\hat{\alpha} = 0$ probability strategy. The sub nodes are then split using the same logic (and so on). Therefore, these recursive decision trees represent the most important breaks in the simulated data and can assist the practitioner in determining the best probability estimation strategy based on the assumptions they are willing to make about their dataset. These generated decision trees are in concordance with the general observations made in the following section but provide a concrete procedure for the practitioner to follow.

Results of the Monte Carlo Simulations

In Online Appendix E we provide Monte Carlo simulation results for the procedure described above. For each specification, the mean absolute deviation from the true value is provided for the two proposed methods of setting the probability of guessing correct. Further, to visually show any trend, a column ('$1/n$ Pref.') has the value 'T' whenever the $1/n$ method of specifying the probability of guessing correct results in a lower mean absolute deviation from the true value. In total, we have four tables of results (`https://goo.gl/2f8NSa`): The first table compares the methods of estimating $\hat{\gamma}$ when the practitioner believes the probability of guessing correct equals $1/4$; the second table compares the methods of estimating $\hat{\gamma}/(1-\hat{\mu})$ when the practitioner believes the probability of guessing correct equals $1/4$; the third table compares the methods of estimating $\hat{\gamma}$ when the practitioner believes the probability of guessing correct equals

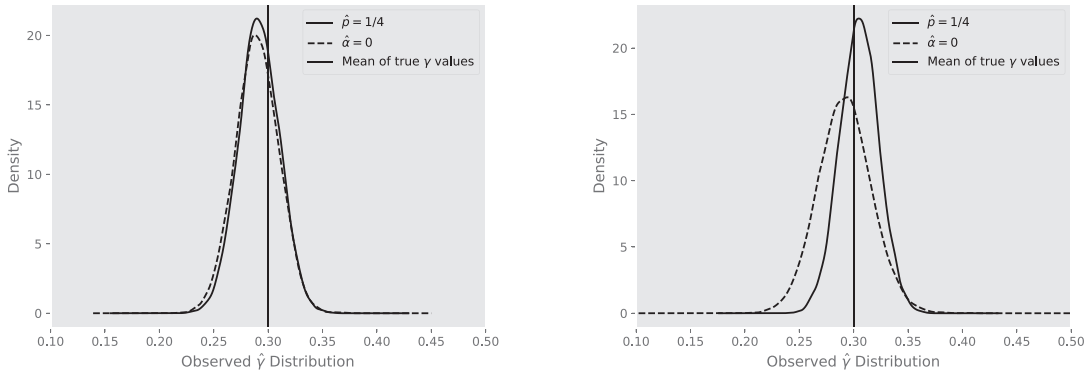
$1/5$; and the final table compares the methods of estimating $\hat{\gamma}/(1-\hat{\mu})$ when the practitioner believes the probability of guessing correct equals $1/5$. The true probability of guessing correct is provided in the column p in each table.

Observation 1 *Under the specified data generating process in our Monte Carlo simulations, when the specified probability of guessing correct using* $1/n$ *deviates from the true probability by* 0.05 *or less, it is more often preferable to use the* $1/n$ *probability estimate.*

The mean absolute deviation is smaller using the $1/n$ estimator in 78% of the 720 simulated distributions estimating $\hat{\gamma}$ when we specify that the true probability deviates by 0.05 or less from $1/n$ $(1/4)$. This is particularly true as α increases. In 82% of cases when $\alpha = 0.03$ and 99% of cases when $\alpha = 0.05$, the $1/n$ estimator produces a smaller mean absolute deviation. The smaller mean absolute deviation can be due to either a less biased estimator or a narrower distribution; when p is overestimated, the distribution produced by assuming $\hat{\alpha} = 0$ is substantially wider than when assuming the probability equals $1/n$ (see Figure 4 for an example). Put simply, due to the increased width of the $\hat{\gamma}_Z$ distribution, the $\hat{\gamma}_{1/n}$ estimator might be a superior option even when the center of the $\hat{\gamma}_Z$ distribution is closer to the true value.

A similar story can be told with the gain estimator: in 65% of the simulated distributions when the $1/n$ probability estimator violates truth by 0.05, the mean absolute deviation is smaller using the $1/n$ probability estimate. Like the $\hat{\gamma}$ estimator, these percentages increase with α: in 75% of the $\alpha = 0.03$ simulated distributions and 99% of the $\alpha = 0.05$ distributions, the $1/n$ estimator is the preferred strategy.

Observation 2 *Under the specified data generating process in our Monte Carlo simulations, when the true* α *value is less than or equal to* 0.03 *and the true probability deviates by* 0.10 *from the estimated* $1/n$ *probability, it is preferable to assume zero true negative learning when estimating the probability of guessing correct. This is particularly true if the* μ *and* γ *values are low.*



(a) $p = 0.20$ (b) $p = 0.30$

Figure 4. Kernel Density Estimates (KDE – for details see Scott (2015)) of $\hat{\gamma}$ distributions when underlying $\mu = 0.3$, $\gamma = 0.3$ and $\alpha = 0.03$. Both simulations are of class sizes of 50 students where the practitioner or researcher assumes the probability of guessing correct is equal to $1/4$. The true probability of guessing correct (p) is below each figure; this is the only difference in the simulation parameters that generated the figures. Optimal bandwidth is calculated using Silverman's (1986) method. The mean value of γ is plotted as a vertical line.

When the $1/n$ probability estimator is incorrect by 0.10, assuming true negative learning (α) is equal to zero often results in a smaller mean deviation from truth. In the case of the $\hat{\gamma}$ estimator, in 50% of the simulated distributions, the $1/n$ estimator results in a lower mean deviation from truth. However, when only looking at the simulated distributions where $\alpha \in \{0.01, 0.02, 0.03\}$, setting $\hat{\alpha} = 0$ is a preferable strategy 60% of the time.

This conclusion is stronger with the gain measurement: in only 26% of cases when the probability deviation of $1/n$ is equal to 0.10 is using the $1/n$ estimator a preferable strategy. Moreover, when $\alpha = 0.03$, setting $\hat{\alpha} = 0$ results in a lower mean absolute deviation 77% of the time; this is increased to 99% when $\alpha = 0.01$.

Further, if γ and μ are low, then setting $\hat{\alpha} = 0$ might result in lower mean absolute deviation. As an example, examining only the rows of the online table where $\gamma = 0.5$ and the true probability deviates by 0.10 from the estimated $1/n$ probability, in 78% of cases using the $1/n$ estimator would be preferable to setting $\hat{\alpha} = 0$. However, when only looking at the rows where $\gamma = 0.1$ (all else equal), in 15% of cases is $1/n$ the preferable strategy. A similar story can be told with μ.

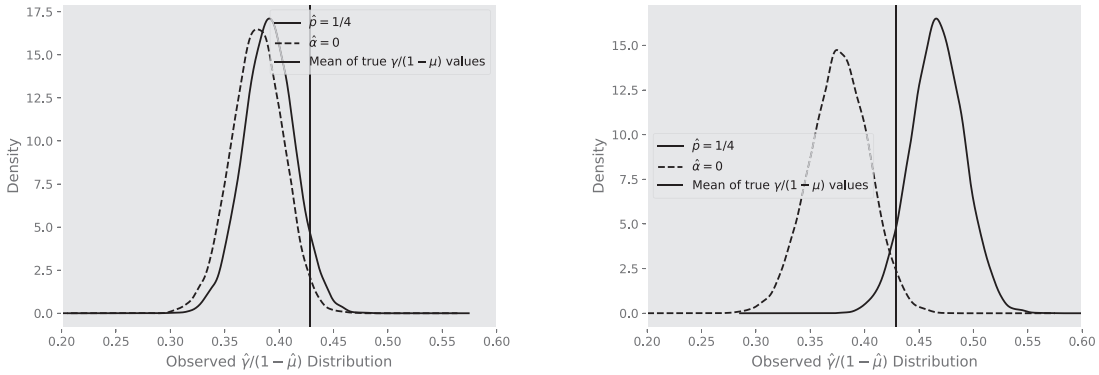
Observation 3 *Under the specified data generating process in our Monte Carlo simulations, calculating the probability of guessing correct under the assumption of zero true negative learning is more often the best strategy when using the gain measurement than when using $\hat{\gamma}$ as the measurement of learning.*

As noted earlier, the mean absolute deviation is smaller using the $1/n$ estimator in 78% of the 720 simulated distributions estimating $\hat{\gamma}$ when we specify that the true probability deviates by 0.05 from $1/n$ $(1/4)$. However, under the same specifications with the gain metric, the $1/n$ strategy is only preferable in 65% of cases. Moreover, when the probability specified by $1/n$ is off by 0.10, then $\hat{\gamma}$ can be estimated using either method with roughly the same chance of the strategy having the lower mean absolute deviation. However, in the case of the gain metric of the same specification, the $1/n$ probability strategy is only preferable 26% of the time. Using the same underlying parameters as Figure 4, Figure 5 compares the width of the $\hat{\gamma}/(1-\hat{\mu})$ distributions when assuming $\hat{\alpha} = 0$ and when assuming the probability equals $1/n$. This shows the difference in distribution width of the two estimation strategies decreases in comparison to the $\hat{\gamma}$ distributions shown in Figure 4. In Figure 5(a), the $1/n$ estimator distribution has a standard deviation of 0.018 while $\hat{\alpha} = 0$ estimator distribution has a standard deviation of 0.020. In Figure 5(b), the $1/n$ estimator distribution has a standard deviation of 0.018 while $\hat{\alpha} = 0$ estimator distribution has a standard deviation of 0.025.

Practitioner's Guide

In this section, we describe how to make actionable decisions based on the information in the study. This section is simplified to show only the relevant conclusions and tools; how these conclusions/tools were derived are left to other sections of the paper.

Assuming the practitioner wishes to use disaggregated value-added learning to estimate the amount of learning in their classroom, study, or assessment procedure, they are faced with the two fundamental questions: (1) should they measure their results in terms of positive learning adjusted for guess ($\hat{\gamma}$) or the gain estimator introduced in this paper ($\hat{\gamma}/(1-\hat{\mu})$), and (2) should they



(a) $p = 0.20$ (b) $p = 0.30$

Figure 5. Kernel Density Estimates (KDE – for details see Scott (2015)) of $\hat{\gamma}/(1-\hat{\mu})$ distributions when underlying $\mu = 0.3$, $\gamma = 0.3$ and $\alpha = 0.03$. Both simulations are of class sizes of 50 students where the practitioner or researcher assumes the probability of guessing correct is equal to $1/4$. The true probability of guessing correct (p) is below each figure; this is the only difference in the simulation parameters that generated the figures. Optimal bandwidth is calculated using Silverman's (1986) method. The mean value of $\gamma/(1-\mu)$ is plotted as a vertical line.

estimate the probability of guessing correct as $1/n$ or assume $\alpha = 0$?

These choices could be influenced many factors. For instance, the practitioner could have a long standing assessment process calculating $\hat{\gamma}$ values for a given course. In that situation, the ability to compare across years might override any statistical factors. Similarly, a practitioner facing populations with very different stock levels of knowledge might prefer to use the gain estimator for increased interpretability. However, for purposes of this section, we will assume that the practitioner wishes to maximize the accuracy of their estimate. This set of choices can be visualized as Table 1.

To determine which cell is the most appropriate, the practitioner must calculate R, defined in *Comparing the Adjusted Positive Learning and the Gain Estimators' Sensitivity to Probability Misspecification*, and follow the decision trees presented in Online Appendix A. We will walk through each of these steps.

The first step is to calculate the unadjusted $\hat{\text{pl}}$, $\hat{\text{nl}}$, $\hat{\text{rl}}$, and $\hat{\text{zl}}$. This disaggregation can be performed by hand or by using software that performs this disaggregation automatically (Smith, 2018). Using these values the practitioner can easily calculate both R equations with a formula:

Table 1
Decision for the practitioner using the tools presented in this paper. R is the ratio of elasticities (or sensitivity) to probability misspecification and E is the average absolute error of given estimation approach.

	$1/n$	$\hat{\alpha}=0$
$\hat{\gamma}$	$\lvert R\rvert>1$ & $\lvert E_{1/n}\rvert<\lvert E_{\hat{\alpha}=0}\rvert$	$\lvert R\rvert>1$ & $\lvert E_{1/n}\rvert>\lvert E_{\hat{\alpha}=0}\rvert$
$\hat{\gamma}/(1-\hat{\mu})$	$\lvert R\rvert<1$ & $\lvert E_{1/n}\rvert<\lvert E_{\hat{\alpha}=0}\rvert$	$\lvert R\rvert<1$ & $\lvert E_{1/n}\rvert>\lvert E_{\hat{\alpha}=0}\rvert$

$R=\frac{\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1}{(\hat{\text{nl}}+\hat{\text{rl}}-1)(\hat{p}+1)+2\hat{\text{pl}}}$ when the practitioner assumes $\hat{p}=1/n$, $R=\frac{\hat{\text{pl}}+\hat{\text{rl}}-1}{2\hat{\text{pl}}+\hat{\text{rl}}-\hat{\text{nl}}-1}$ when the practitioner assumes $\hat{\alpha}=0$.

R represents the ratio of two probability misspecification elasticities ($\frac{\varepsilon_{\hat{\gamma}/(1-\hat{\mu})}}{\varepsilon_{\hat{\gamma}}}$). As the gain estimator is in the numerator of this fraction, when $-1<R<1$ then the gain estimator is less sensitive (in percent terms) to misspecification. The two calculated R values are likely close to each other (both in the range $-1<R<1$ or not); if they are not, the practitioner might need to perform the steps that follow twice. With the estimator selected, the practitioner can move to selecting either $1/n$ or $\hat{\alpha}=0$ to estimate $\hat{p}$.

The goal of the practitioner is to choose a strategy that maximizes the accuracy of the chosen estimator: in essence, which strategy would result in less absolute error. However, this absolute error is not known without assumptions. To provide the practitioner at least some guidance on this choice, a large set of Monte Carlo simulations have been conducted under a given set of assumptions, not least of which is the independence of the questions (details of these assumptions are provided in *Comparative Monte Carlo Simulations*). These results were then fit to decision trees that represent the collective wisdom of all of these simulations in one diagram.

The practitioner should start with what they know about the exam questions. If the question has been carefully designed to give a myopic student chance probability of guessing correct, then probability of guessing correct is likely to converge on $1/n$. If that wasn't a central tenant of question construction, however, this might not be the case. Knowledge of question guessing probability is the first step to putting the trees in Online Appendix A in action. As an example, suppose that the practitioner has chosen to use the gain estimator and believes that their

four-option multiple choice exam was designed to give students chance probability of guessing correct. Following Online Appendix Figure A2(b), there is a yes/no choice indicting that $|\hat{p} - p| > 0.055$. If the practitioner believes that the questions are designed to give the student only chance probability of guessing correct, they likely do not believe that there is such a large deviation from the true probability. Noted in the round node, without further questions, the diagram suggests the practitioner should use the $1/n$ probability estimator. However, if they believe they know how much negative learning (α) occurs in their sample, they could further refine their choice. In this case, the tree asks if there is less than 0.015 true negative learning; a very small amount. The practitioner likely does not believe that is reasonable, answers 'no' and finds the final node (continuing to suggest the $1/n$ estimator).

The final step is to determine if the observed learning value is statistically different from no learning. To give a concrete example, if a practitioner or researcher observed a $\hat{\mu}$ of 0.6, had a class size of 30 (m), estimated the gain measurement ($\hat{\gamma}/(1 - \hat{\mu})$) to be 0.4, and was exploring a question with $n = 4$, then the final five rows of Online Appendix Table B1 corresponding to class sizes of 30 might apply. Because the true μ may lie between 0.4 and 0.8, the practitioner or researcher would use the critical values (at either the 90% or 95% level) to test whether or not the gain was statistically different from 0. With a gain of 0.4, the practitioner or researcher would fail to reject the null hypothesis of no learning at the 95% level but reject the null at the 90% level.

Conclusion

This paper extends the work of Walstad and Wagner (2016) and Smith and Wagner (2018) by: (1) providing a transformed measure of learning that under knowable conditions is less sensitive to probability misspecification; (2) providing a statistical test for the new transformed measure; (3) suggesting an alternative method of determining the probability of guessing correct ($\hat{\alpha} = 0$); (4) simulating the distribution of outcomes when the probability of guessing correct is assumed to be $1/n$ (where n is the number of answer options) and assuming true negative learning is zero and solving for the probability that this implies. These Monte Carlo simulations

show when one probability specification results in a smaller mean absolute deviation from the true value (which are known in the simulations). As our assumptions are very strong, we only make general observations about our simulation results.

One reasonable conclusion one could draw from the Monte Carlo simulations is a reiteration of the importance of exam questions with a known probability of guessing correct. When the specified probability of guessing correct using $1/n$ was deviated by 0.05 or less in either direction from the true value, the simulations indicated it was almost always better to use the $1/n$ probability estimator. Arguably, exam question options that result in more than a 0.05 probability deviation from random guessing is in an indication of exam questions with distractors that are not serving their intended purpose. This suggests that carefully written question distractors aren't just important to create a stronger discriminator between high and low quality students, but are in fact critical for this type of analysis of the data. Widely used exam questions that have been tested, even if a norming sample isn't available, might result in a higher quality analysis.

Further, our simulations suggested that the estimators assuming zero true negative learning performed best when true positive and retained learning were comparatively low. This result was hinted at in *A Brief Description of the Estimators in Smith and Wagner (2018)* where if raw zero learning equaled zero, the implied probability of guessing the correct answer equaled one when assuming $\hat{\alpha} = 0$. This suggests that datasets with very low zero learning would result in impractical implied probabilities of guessing the correct answer on a question.

While the transformed measure, supplemented by the Monte Carlo simulations, provides the practitioner or researcher guidance, it does not replace judgment. Only the practitioner/researcher can decide what assumptions they are willing to make for their particular dataset. However, this paper provides them with tools to make more informed decisions when it comes time to assess learning in a collection of classes, in their own class, or to test the effectiveness of a pedagogical technique; the section *Practitioner's Guide* can help practitioners make this decision.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & E. Novick M.R. (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading: Addison-Wesley.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and regression trees*. Taylor & Francis.

Colt, H. G., Davoudi, M., Murgu, S., & Rohani, N. Z. (2011). Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical endoscopy*, *25*(1), 207–216.

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

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

Gönülateş, E., & Kortemeyer, G. (2017). Modeling unproductive behavior in online homework in terms of latent student traits: An approach based on item response theory. *Journal of Science Education and Technology*, *26*(2), 139–150.

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.

Hamne, P., & Bernhard, J. (2000). Educating pre-service teachers using hands-on and microcomputer based labs as tools for concept substitution. *Physics Teacher Education Beyond*, 663–666.

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

Happ, R., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2016). An analysis of economic learning among undergraduates in introductory economics courses in germany. *The Journal of Economic Education*, *47*(4), 300-310.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of s-x2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289–298.

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

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

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). London: Chapman & Hall.

Smith, B. O. (2018). Multiplatform software tool to disaggregate and adjust value-added learning scores. *The Journal of Economic Education*, *49*(2), 220–221.

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.

Supasorn, S. (2015). Grade 12 students' conceptual understanding and mental models of galvanic cells before and after learning by using small-scale experiments in conjunction with a model kit. *Chemistry Education Research and Practice*, *16*(2), 393–407.

Swets, J. A. (1969). Effectiveness of information retrieval methods. *American Documentation*, *20*(1), 72–89.

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.

Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of understanding of college economics: Examiner's manual* (4th ed.). New York: Council for Economic Education.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems

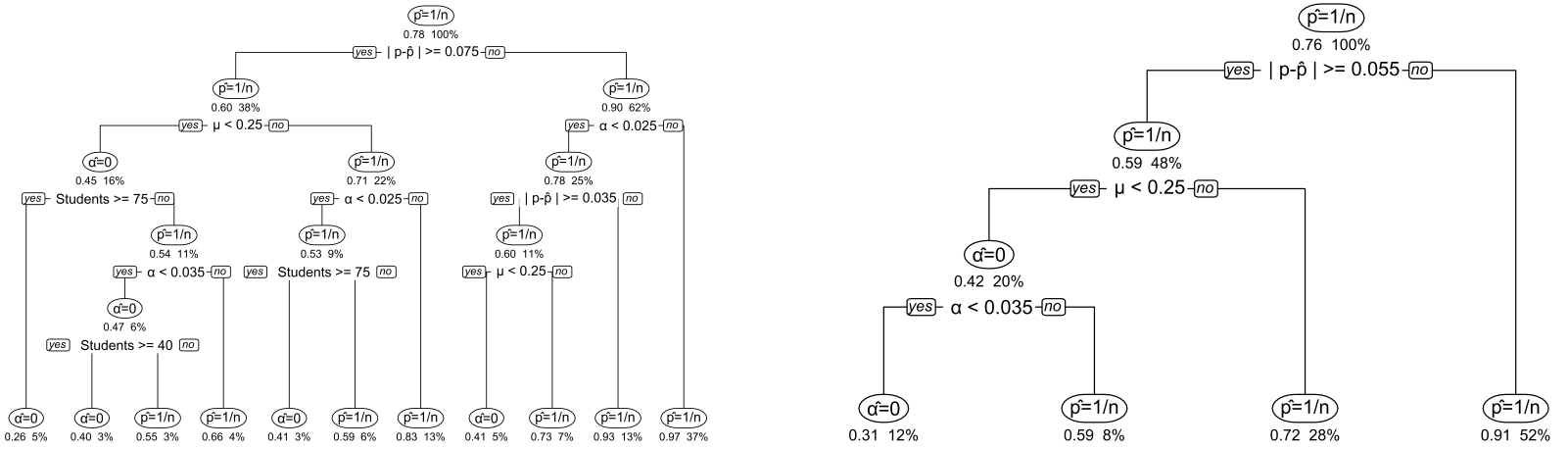
and potential solutions. *Educational assessment*, *10*(1), 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated irt model. *Journal of Educational Measurement*, *43*(1), 19–38.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, *27*(5), 357–371.

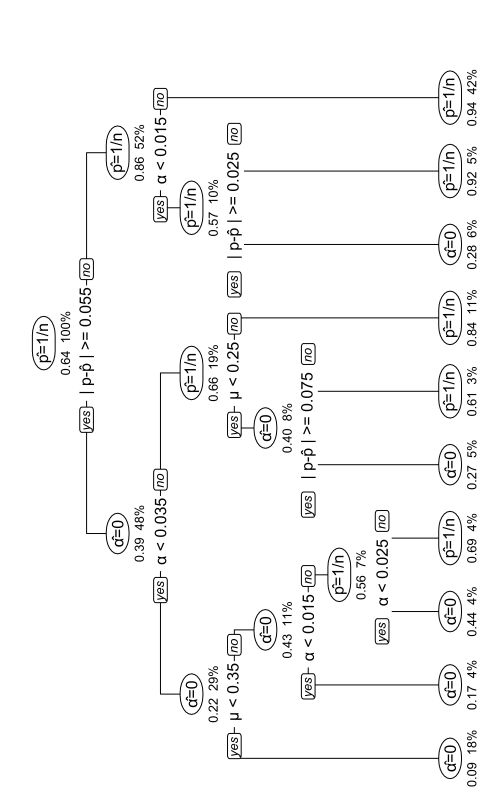
Appendix A

Recursive Partitioning Trees



(a) $\hat{p} = 0.20$ (b) $\hat{p} = 0.25$

Figure A1. The simulated $\hat{\gamma}$ distributions (`https://goo.gl/2f8NSa`) were used to fit a Recursive Partitioning tree (Breiman, Friedman, Stone, & Olshen, 1984) with an outcome variable of the minimum absolute error estimator. Features included were the absolute deviation from the true probability of guessing correct ($|p - \hat{p}|$), true underlying stock knowledge (μ), true underlying negative learning (α), and class size (*Students*); this last feature didn't appear in one of the trees as it did not provide sufficient information gain. Each node (oval) represents the best (lowest absolute error) estimation strategy at that point in the tree. Below each node is the proportion of the data in that node where $\hat{p} = 1/n$ is the preferred strategy (first number) followed by the proportion of the overall training data that node represents. The practitioner would proceed by answering the yes/no question and examining the resulting node. The practitioner should repeat this process until they encounter a terminal node or they are uncomfortable making an assumption about a yes/no question. Figure (a) is trained on all $\hat{\gamma}$ simulation results where $p = 1/5$ and Figure (b) is trained on all $\hat{\gamma}$ simulation results where $p = 1/4$.



(a) $\hat{p} = 0.20$

(b) $\hat{p} = 0.25$

Figure A2. The simulated $\frac{\hat{\gamma}}{1-\hat{\mu}}$ distributions (`https://goo.gl/2f8NSa`) were used to fit a Recursive Partitioning tree (Breiman et al., 1984) with an outcome variable of the minimum absolute error estimator. Features included were the absolute deviation from the true probability of guessing correct ($|p - \hat{p}|$), true underlying stock knowledge (μ), true underlying negative learning (α), and class size; this last feature didn't appear in the trees as it did not provide sufficient information gain. Each node (oval) represents the best (lowest absolute error) estimation strategy at that point in the tree. Below each node is the proportion of the data in that node where $\hat{p} = 1/n$ is the preferred strategy (first number) followed by the proportion of the overall training data that node represents. The practitioner would proceed by answering the yes/no question and examining the resulting node. The practitioner should repeat this process until they encounter a terminal node or they are uncomfortable making an assumption about a yes/no question. Figure (a) is trained on all $\frac{\hat{\gamma}}{1-\hat{\mu}}$ simulation results where $p = 1/5$ and Figure (b) is trained on all $\frac{\hat{\gamma}}{1-\hat{\mu}}$ simulation results where $p = 1/4$.

Appendix B

Monte Carlo Simulations of the Gain Measurement under the Null of Zero Learning

Table B1
90% and 95% critical values from simulated $\hat{\gamma}/(1-\hat{\mu})$ distribution where $n = 4$ $(p = 1/4)$ and underlying $\gamma = 0$ and $\alpha = 0$

Students	μ	$\frac{\hat{\gamma}}{1-\hat{\mu}}_{0.90}$ CV	$\frac{\hat{\gamma}}{1-\hat{\mu}}_{0.95}$ CV	$\hat{\mu}$ 95% CI	
15	0.1	0.238	0.333	-0.156	0.378
15	0.2	0.259	0.333	-0.156	0.556
15	0.3	0.259	0.333	-0.067	0.644
15	0.4	0.333	0.429	0.111	0.733
15	0.5	0.333	0.467	0.200	0.822
15	0.6	0.333	0.556	0.289	0.911
15	0.7	0.556	1.000	0.378	0.911
15	0.8	1.000	1.000	0.556	1.000
30	0.1	0.175	0.222	-0.111	0.333
30	0.2	0.185	0.238	-0.022	0.422
30	0.3	0.200	0.250	0.067	0.556
30	0.4	0.200	0.273	0.156	0.644
30	0.5	0.238	0.333	0.289	0.733
30	0.6	0.259	0.333	0.378	0.822
30	0.7	0.333	0.429	0.511	0.867
30	0.8	0.333	0.556	0.600	0.956
50	0.1	0.135	0.171	-0.067	0.280
50	0.2	0.140	0.183	0.013	0.387
50	0.3	0.147	0.200	0.120	0.493
50	0.4	0.158	0.200	0.227	0.573
50	0.5	0.175	0.228	0.307	0.680
50	0.6	0.200	0.259	0.413	0.760
50	0.7	0.238	0.333	0.547	0.840
50	0.8	0.333	0.333	0.653	0.920
80	0.1	0.103	0.136	-0.033	0.233
80	0.2	0.111	0.141	0.067	0.350
80	0.3	0.111	0.147	0.150	0.450
80	0.4	0.124	0.162	0.250	0.550
80	0.5	0.140	0.183	0.367	0.633
80	0.6	0.158	0.210	0.467	0.733
80	0.7	0.185	0.246	0.583	0.817
80	0.8	0.222	0.289	0.700	0.900
100	0.1	0.091	0.118	-0.027	0.227
100	0.2	0.096	0.127	0.067	0.333
100	0.3	0.103	0.133	0.173	0.427
100	0.4	0.111	0.145	0.267	0.533
100	0.5	0.124	0.162	0.373	0.627
100	0.6	0.140	0.179	0.480	0.720
100	0.7	0.158	0.212	0.587	0.800
100	0.8	0.200	0.259	0.707	0.893
300	0.1	0.052	0.066	0.031	0.169
300	0.2	0.056	0.074	0.129	0.276
300	0.3	0.059	0.076	0.227	0.373
300	0.4	0.065	0.083	0.324	0.476
300	0.5	0.069	0.091	0.427	0.573
300	0.6	0.082	0.106	0.529	0.667
300	0.7	0.093	0.118	0.636	0.760
300	0.8	0.111	0.145	0.747	0.853

Notes: Each row indicates the critical values at the 90% threshold ($\frac{\hat{\gamma}}{1-\hat{\mu}}_{0.90}$ CV) and 95% threshold ($\frac{\hat{\gamma}}{1-\hat{\mu}}_{0.95}$ CV). The 95% confidence interval of the estimated $\hat{\mu}$ values ($\hat{\mu}$ 95% CI) are provided to assist the practitioner or researcher in finding the appropriate $\frac{\hat{\gamma}}{1-\hat{\mu}}$ critical value for a given statement. A much larger set of critical values are available at `https://goo.gl/kKAySX`. All simulations run with 10,000 repetitions.

Table B2
90% and 95% critical values from simulated $\hat{\gamma}/(1-\hat{\mu})$ distribution where $n = 5$ $(p = 1/5)$ and underlying $\gamma = 0$ and $\alpha = 0$

Students	μ	$\frac{\hat{\gamma}}{1-\hat{\mu}}_{0.90}$ CV	$\frac{\hat{\gamma}}{1-\hat{\mu}}_{0.95}$ CV	$\hat{\mu}$ 95% CI	
15	0.1	0.205	0.271	-0.167	0.417
15	0.2	0.219	0.306	-0.083	0.500
15	0.3	0.231	0.306	0.000	0.583
15	0.4	0.250	0.375	0.083	0.667
15	0.5	0.286	0.375	0.167	0.833
15	0.6	0.375	0.464	0.333	0.917
15	0.7	0.375	0.583	0.417	0.917
15	0.8	1.000	1.000	0.583	1.000
30	0.1	0.145	0.188	-0.083	0.292
30	0.2	0.148	0.196	0.000	0.417
30	0.3	0.167	0.211	0.083	0.500
30	0.4	0.167	0.231	0.167	0.625
30	0.5	0.196	0.250	0.292	0.708
30	0.6	0.219	0.286	0.375	0.792
30	0.7	0.250	0.375	0.500	0.875
30	0.8	0.375	0.500	0.625	0.958
50	0.1	0.107	0.143	-0.050	0.250
50	0.2	0.113	0.152	0.050	0.375
50	0.3	0.125	0.167	0.125	0.475
50	0.4	0.135	0.167	0.225	0.575
50	0.5	0.145	0.196	0.325	0.675
50	0.6	0.167	0.226	0.425	0.750
50	0.7	0.196	0.250	0.550	0.850
50	0.8	0.250	0.375	0.675	0.925
80	0.1	0.087	0.113	-0.016	0.219
80	0.2	0.091	0.118	0.078	0.328
80	0.3	0.096	0.125	0.172	0.438
80	0.4	0.107	0.141	0.266	0.531
80	0.5	0.113	0.152	0.359	0.641
80	0.6	0.130	0.167	0.469	0.719
80	0.7	0.148	0.196	0.578	0.813
80	0.8	0.196	0.261	0.688	0.891
100	0.1	0.076	0.101	-0.013	0.213
100	0.2	0.083	0.110	0.088	0.313
100	0.3	0.089	0.116	0.175	0.425
100	0.4	0.094	0.122	0.275	0.525
100	0.5	0.103	0.135	0.375	0.625
100	0.6	0.113	0.152	0.488	0.713
100	0.7	0.135	0.179	0.588	0.800
100	0.8	0.167	0.219	0.700	0.888
300	0.1	0.044	0.057	0.038	0.167
300	0.2	0.048	0.063	0.133	0.267
300	0.3	0.051	0.064	0.229	0.371
300	0.4	0.054	0.071	0.329	0.471
300	0.5	0.060	0.077	0.433	0.571
300	0.6	0.065	0.084	0.533	0.667
300	0.7	0.078	0.102	0.638	0.758
300	0.8	0.094	0.122	0.746	0.850

Notes: Each row indicates the critical values at the 90% threshold ($\frac{\hat{\gamma}}{1-\hat{\mu}}_{0.90}$ CV) and 95% threshold ($\frac{\hat{\gamma}}{1-\hat{\mu}}_{0.95}$ CV). The 95% confidence interval of the estimated $\hat{\mu}$ values ($\hat{\mu}$ 95% CI) are provided to assist the practitioner or researcher in finding the appropriate $\frac{\hat{\gamma}}{1-\hat{\mu}}$ critical value for a given statement. A much larger set of critical values are available at `https://goo.gl/kKAySX`. All simulations run with 10,000 repetitions.

Appendix C

Derivations of Equations in the Article

Derivations of all equations in this paper are provided in this section. Alternatively, we have posted Mathematica files showing the same result: `https://goo.gl/qjCCUj`.

Derivations of Equation 4. If $\hat{\alpha} = 0$ is assumed, equation 2 can be set to zero and solved for $\hat{p}$.

$$\begin{aligned} 0 &= \frac{\hat{p}(\hat{\text{pl}} + \hat{\text{rl}} - 1) + \hat{\text{nl}}}{(\hat{p} - 1)^2} \\ 0 &= \hat{p}(\hat{\text{pl}} + \hat{\text{rl}} - 1) + \hat{\text{nl}} \\ -\hat{\text{nl}} &= \hat{p}(\hat{\text{pl}} + \hat{\text{rl}} - 1) \\ \hat{p} &= \frac{\hat{\text{nl}}}{1 - \hat{\text{pl}} - \hat{\text{rl}}} \end{aligned} \tag{12}$$

Derivations of Equation 5. Assuming $\hat{\alpha} = 0$, one can substitute equation 4 into equation 1 and simplify.

$$\begin{aligned} \hat{\gamma} &= \frac{\hat{p}(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}}{(\hat{p} - 1)^2} \\ (\hat{\gamma}|\hat{\alpha} = 0) &= \frac{\frac{\hat{\text{nl}}}{1-\hat{\text{pl}}-\hat{\text{rl}}}(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}}{\left(\frac{\hat{\text{nl}}}{1-\hat{\text{pl}}-\hat{\text{rl}}} - 1\right)^2} \\ (\hat{\gamma}|\hat{\alpha} = 0) &= \frac{\hat{\text{nl}}(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}(1 - \hat{\text{pl}} - \hat{\text{rl}})}{\left(\frac{\hat{\text{nl}}}{1-\hat{\text{pl}}-\hat{\text{rl}}} - \frac{1-\hat{\text{pl}}-\hat{\text{rl}}}{1-\hat{\text{pl}}-\hat{\text{rl}}}\right)^2 (1 - \hat{\text{pl}} - \hat{\text{rl}})} \\ (\hat{\gamma}|\hat{\alpha} = 0) &= \frac{\hat{\text{nl}}(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}(1 - \hat{\text{pl}} - \hat{\text{rl}})}{\left(\frac{\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1}{1-\hat{\text{pl}}-\hat{\text{rl}}}\right)(\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1)} \\ (\hat{\gamma}|\hat{\alpha} = 0) &= \frac{(\hat{\text{nl}}(\hat{\text{nl}} + \hat{\text{rl}} - 1) + \hat{\text{pl}}(1 - \hat{\text{pl}} - \hat{\text{rl}}))(1 - \hat{\text{pl}} - \hat{\text{rl}})}{(\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1)^2} \\ (\hat{\gamma}|\hat{\alpha} = 0) &= \frac{(\hat{\text{nl}}^2 + \hat{\text{nl}}\hat{\text{rl}} - \hat{\text{nl}} + \hat{\text{pl}} - \hat{\text{pl}}^2 - \hat{\text{pl}}\hat{\text{rl}})(1 - \hat{\text{pl}} - \hat{\text{rl}})}{(\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1)^2} \\ (\hat{\gamma}|\hat{\alpha} = 0) &= \frac{(\hat{\text{nl}} - \hat{\text{pl}})(\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1)(1 - \hat{\text{pl}} - \hat{\text{rl}})}{(\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1)^2} \\ (\hat{\gamma}|\hat{\alpha} = 0) &= \frac{(\hat{\text{nl}} - \hat{\text{pl}})(1 - \hat{\text{pl}} - \hat{\text{rl}})}{\hat{\text{nl}} + \hat{\text{pl}} + \hat{\text{rl}} - 1} \end{aligned} \tag{13}$$

Derivation of Equation 6. Using equations 1 and 3, one can find $\frac{\hat{\gamma}}{1-\hat{\mu}}$.

$$\begin{aligned}
\frac{\hat{\gamma}}{1-\hat{\mu}} &= \frac{\frac{\hat{p}(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(\hat{p}-1)^2}}{1-\frac{\hat{\text{nl}}+\hat{\text{rl}}-\hat{p}}{1-\hat{p}}} \\
\frac{\hat{\gamma}}{1-\hat{\mu}} &= \frac{\frac{\hat{p}(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(\hat{p}-1)^2}}{\frac{\hat{\text{nl}}+\hat{\text{rl}}-1}{\hat{p}-1}} \\
\frac{\hat{\gamma}}{1-\hat{\mu}} &= \frac{\frac{\hat{p}(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(\hat{p}-1)}}{\hat{\text{nl}}+\hat{\text{rl}}-1} \\
\frac{\hat{\gamma}}{1-\hat{\mu}} &= \frac{\hat{p}(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(\hat{\text{nl}}+\hat{\text{rl}}-1)(\hat{p}-1)}
\end{aligned} \tag{14}$$

Derivation of Equation 7. Using equations 6, we can find the value $(\frac{\hat{\gamma}}{1-\hat{\mu}}|\hat{\alpha}=0)$ by substituting equation 4 in for all instances of $\hat{p}$ and simplifying.

$$\begin{aligned}
\frac{\hat{\gamma}}{1-\hat{\mu}} &= \frac{\hat{p}(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(\hat{\text{nl}}+\hat{\text{rl}}-1)(\hat{p}-1)} \\
(\frac{\hat{\gamma}}{1-\hat{\mu}}|\hat{\alpha}=0) &= \frac{\frac{\hat{\text{nl}}}{1-\hat{\text{rl}}-\hat{\text{pl}}}(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(\hat{\text{nl}}+\hat{\text{rl}}-1)(\frac{\hat{\text{nl}}}{1-\hat{\text{rl}}-\hat{\text{pl}}}-\frac{1-\hat{\text{rl}}-\hat{\text{pl}}}{1-\hat{\text{rl}}-\hat{\text{pl}}})} \\
(\frac{\hat{\gamma}}{1-\hat{\mu}}|\hat{\alpha}=0) &= \frac{\hat{\text{nl}}(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}(1-\hat{\text{rl}}-\hat{\text{pl}})}{(\hat{\text{nl}}+\hat{\text{rl}}-1)(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)} \\
(\frac{\hat{\gamma}}{1-\hat{\mu}}|\hat{\alpha}=0) &= \frac{(\hat{\text{nl}}-\hat{\text{pl}})(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)}{(\hat{\text{nl}}+\hat{\text{rl}}-1)(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)} \\
(\frac{\hat{\gamma}}{1-\hat{\mu}}|\hat{\alpha}=0) &= \frac{\hat{\text{pl}}-\hat{\text{nl}}}{1-\hat{\text{nl}}-\hat{\text{rl}}}
\end{aligned} \tag{15}$$

Derivation of Equation 9. The following is the derivation of $\varepsilon_{\hat{\gamma}/(1-\hat{\mu})}$. This can be achieved by:

- Finding $\frac{\delta(\hat{\gamma}/(1-\hat{\mu}))}{\delta\Delta}$, which we will define as $g'(\Delta)$.
- Multiply $g'(\Delta)\frac{\Delta}{g(\Delta)}$.

Finding $g'(\Delta)$:

$$
\begin{aligned}
g(\Delta) &= \frac{(p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(p+\Delta-1)(\hat{\text{nl}}+\hat{\text{rl}}-1)} \\
g'(\Delta) &= \frac{1}{p+\Delta-1} - \frac{\hat{\text{pl}}+(\hat{\text{nl}}+\hat{\text{rl}}-1)(p+\Delta)}{(p+\Delta-1)^2(\hat{\text{rl}}+\hat{\text{nl}}-1)} \\
g'(\Delta) &= \frac{(p+\Delta-1)(\hat{\text{rl}}+\hat{\text{nl}}-1)-\hat{\text{pl}}-(\hat{\text{nl}}+\hat{\text{rl}}-1)(p+\Delta)}{(p+\Delta-1)^2(\hat{\text{rl}}+\hat{\text{nl}}-1)} \\
g'(\Delta) &= \frac{(\hat{\text{rl}}+\hat{\text{nl}}-1)((p+\Delta-1)-(p+\Delta))-\hat{\text{pl}}}{(p+\Delta-1)^2(\hat{\text{rl}}+\hat{\text{nl}}-1)} \\
g'(\Delta) &= \frac{1-\hat{\text{rl}}-\hat{\text{nl}}-\hat{\text{pl}}}{(p+\Delta-1)^2(\hat{\text{rl}}+\hat{\text{nl}}-1)}
\end{aligned} \tag{16}
$$

Multiply $g'(\Delta)\frac{\Delta}{g(\Delta)}$:

$$
\begin{aligned}
\varepsilon_{\hat{\gamma}/(1-\hat{\mu})} &= \left(\frac{1-\hat{\text{rl}}-\hat{\text{nl}}-\hat{\text{pl}}}{(p+\Delta-1)^2(\hat{\text{rl}}+\hat{\text{nl}}-1)}\right)\frac{\Delta}{\frac{(p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(p+\Delta-1)(\hat{\text{nl}}+\hat{\text{rl}}-1)}} \\
\varepsilon_{\hat{\gamma}/(1-\hat{\mu})} &= \frac{\Delta(1-\hat{\text{nl}}-\hat{\text{rl}}-\hat{\text{pl}})}{(p+\Delta-1)(p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}(p+\Delta-1)} \\
\varepsilon_{\hat{\gamma}/(1-\hat{\mu})} &= \frac{\Delta(1-\hat{\text{nl}}-\hat{\text{rl}}-\hat{\text{pl}})}{(p+\Delta-1)((p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}})} \\
\varepsilon_{\hat{\gamma}/(1-\hat{\mu})} &= -\frac{\Delta(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)}{(\Delta+p-1)((\hat{\text{nl}}+\hat{\text{rl}}-1)(\Delta+p)+\hat{\text{pl}})}
\end{aligned} \tag{17}
$$

The following is the derivation of $\varepsilon_{\hat{\gamma}}$. This can be achieved by:

- Finding $\frac{\delta\hat{\gamma}}{\delta\Delta}$, which we will define as $f'(\Delta)$.
- Multiply $f'(\Delta)\frac{\Delta}{f(\Delta)}$.

Finding $f'(\Delta)$:

$$
\begin{aligned}
f(\Delta) &= \frac{(p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(p+\Delta-1)^2} \\
f'(\Delta) &= \frac{\hat{\text{nl}}+\hat{\text{rl}}-1}{(p+\Delta-1)^2} - \frac{2((p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}})}{(p+\Delta-1)^3} \\
f'(\Delta) &= \frac{(\hat{\text{nl}}+\hat{\text{rl}}-1)(p+\Delta-1)-2((p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}})}{(p+\Delta-1)^3}
\end{aligned} \tag{18}
$$

Multiply $f'(\Delta)\frac{\Delta}{f(\Delta)}$:

$$
\begin{aligned}
\varepsilon_{\hat{\gamma}} &= \frac{(\hat{\text{nl}}+\hat{\text{rl}}-1)(p+\Delta-1)-2((p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}})}{(p+\Delta-1)^3}\frac{\Delta}{\frac{(p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}}{(p+\Delta-1)^2}} \\
\varepsilon_{\hat{\gamma}} &= \frac{\Delta((\hat{\text{nl}}+\hat{\text{rl}}-1)(p+\Delta-1)-2((p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}))}{(p+\Delta-1)(p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}}(p+\Delta-1)} \\
\varepsilon_{\hat{\gamma}} &= \frac{\Delta((\hat{\text{nl}}+\hat{\text{rl}}-1)((p+\Delta-1)-2(p+\Delta))-2\hat{\text{pl}})}{(p+\Delta-1)((p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}})} \\
\varepsilon_{\hat{\gamma}} &= \frac{\Delta((\hat{\text{nl}}+\hat{\text{rl}}-1)(-1-p-\Delta)-2\hat{\text{pl}})}{(p+\Delta-1)((p+\Delta)(\hat{\text{nl}}+\hat{\text{rl}}-1)+\hat{\text{pl}})} \\
\varepsilon_{\hat{\gamma}} &= -\frac{\Delta(\hat{\text{nl}}(\Delta+p+1)+2\hat{\text{pl}}+(\hat{\text{rl}}-1)(\Delta+p+1))}{(\Delta+p-1)((\hat{\text{nl}}+\hat{\text{rl}}-1)(\Delta+p)+\hat{\text{pl}})}
\end{aligned} \tag{19}
$$

Derivation of Equation 10.

$$
\begin{aligned}
R &= \frac{\varepsilon_{\hat{\gamma}/(1-\hat{\mu})}}{\varepsilon_{\hat{\gamma}}} = \frac{-\frac{\Delta(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)}{(\Delta+p-1)((\hat{\text{nl}}+\hat{\text{rl}}-1)(\Delta+p)+\hat{\text{pl}})}}{-\frac{\Delta(\hat{\text{nl}}(\Delta+p+1)+2\hat{\text{pl}}+(\hat{\text{rl}}-1)(\Delta+p+1))}{(\Delta+p-1)((\hat{\text{nl}}+\hat{\text{rl}}-1)(\Delta+p)+\hat{\text{pl}})}} \\
R &= \frac{\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1}{\hat{\text{nl}}(\Delta+p+1)+2\hat{\text{pl}}+(\hat{\text{rl}}-1)(\Delta+p+1)} \\
R &= \frac{\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1}{2\hat{\text{pl}}+(\hat{\text{nl}}+\hat{\text{rl}}-1)(\Delta+p+1)}
\end{aligned} \tag{20}
$$

Derivation of Equation 11. Given that $\hat{p} = p + \Delta$, all instances of $p + \Delta$ are replaced with equation 4 given we are assuming $\hat{\alpha} = 0$.

$$
\begin{aligned}
R &= \frac{\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1}{\hat{\text{nl}}(\Delta+p+1)+2\hat{\text{pl}}+(\hat{\text{rl}}-1)(\Delta+p+1)} \\
(R|\hat{\alpha}=0) &= \frac{\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1}{\hat{\text{nl}}(\frac{\hat{\text{nl}}}{1-\hat{\text{pl}}-\hat{\text{rl}}}+1)+2\hat{\text{pl}}+(\hat{\text{rl}}-1)(\frac{\hat{\text{nl}}}{1-\hat{\text{pl}}-\hat{\text{rl}}}+1)} \\
(R|\hat{\alpha}=0) &= \frac{\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1}{\hat{\text{nl}}(\frac{\hat{\text{nl}}+1-\hat{\text{pl}}-\hat{\text{rl}}}{1-\hat{\text{pl}}-\hat{\text{rl}}})+2\hat{\text{pl}}+(\hat{\text{rl}}-1)(\frac{\hat{\text{nl}}+1-\hat{\text{pl}}-\hat{\text{rl}}}{1-\hat{\text{pl}}-\hat{\text{rl}}})} \\
(R|\hat{\alpha}=0) &= \frac{(\hat{\text{pl}}+\hat{\text{rl}}-1)(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)}{-(\hat{\text{nl}}(1+\hat{\text{nl}}-\hat{\text{pl}}-\hat{\text{rl}}))+2\hat{\text{pl}}(\hat{\text{pl}}+\hat{\text{rl}}-1)-(\hat{\text{rl}}-1)(1+\hat{\text{nl}}-\hat{\text{pl}}-\hat{\text{rl}})} \\
(R|\hat{\alpha}=0) &= \frac{(\hat{\text{pl}}+\hat{\text{rl}}-1)(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)}{-\hat{\text{nl}}-\hat{\text{nl}}^2+\hat{\text{nl}}\hat{\text{pl}}+\hat{\text{nl}}\hat{\text{rl}}-2\hat{\text{pl}}+2\hat{\text{pl}}^2+2\hat{\text{pl}}\hat{\text{rl}}+1+\hat{\text{nl}}-\hat{\text{pl}}-2\hat{\text{rl}}-\hat{\text{nl}}\hat{\text{rl}}+\hat{\text{pl}}\hat{\text{rl}}+\hat{\text{rl}}^2} \\
(R|\hat{\alpha}=0) &= \frac{(\hat{\text{pl}}+\hat{\text{rl}}-1)(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)}{1-\hat{\text{nl}}^2-3\hat{\text{pl}}+\hat{\text{nl}}\hat{\text{pl}}+2\hat{\text{pl}}^2-2\hat{\text{rl}}+3\hat{\text{pl}}\hat{\text{rl}}+\hat{\text{rl}}^2} \\
(R|\hat{\alpha}=0) &= \frac{(\hat{\text{pl}}+\hat{\text{rl}}-1)(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)}{(2\hat{\text{pl}}+\hat{\text{rl}}-\hat{\text{nl}}-1)(\hat{\text{nl}}+\hat{\text{pl}}+\hat{\text{rl}}-1)} \\
(R|\hat{\alpha}=0) &= \frac{\hat{\text{pl}}+\hat{\text{rl}}-1}{2\hat{\text{pl}}+\hat{\text{rl}}-\hat{\text{nl}}-1}
\end{aligned} \tag{21}
$$

Appendix D

Simulations with Learning Correlated with Ability

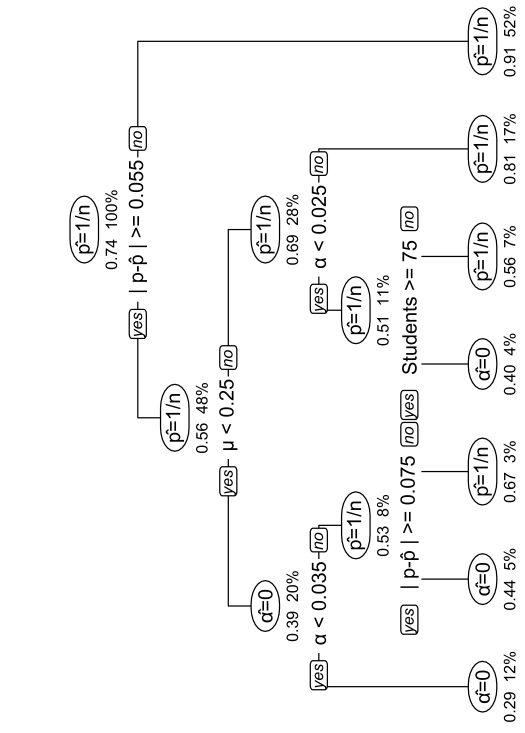
As a sensitivity check, we re-simulated every comparative Monte Carlo simulation presented in this paper. In the original simulations, we assumed a Binomial distribution to determine the student's level of stock knowledge (μ), positive learning (γ), and negative learning or forgetting (α). A core issue with these simulations is the implied assumption that there is no correlation between knowledge and learning. In this set of simulations we modify this assumption and assume that as stock knowledge increases, the probability of positive learning increases and the probability of forgetting (negative learning) decreases. Specifically, we use a logistic function ($\frac{L}{1+e^{-k(x-x_0)}}$) along with the difference between the seeded μ and the individual student's μ_i ($x - x_0 = \mu_i - \mu$ for positive learning; $x - x_0 = \mu - \mu_i$ for negative learning) to change the probability of the student learning or forgetting. We have set the limit on the logistic function such that the probability cannot exceed the bounds of $[0, 1]$, however, it spans as much of the range as possible while still being centered on the seeded value and approaches an appropriate boundary asymptotically.

We purposely made the correlation between μ_i and learning quite strong; we set the k value in the logistic function at 5. There is, of course, an infinite number of possible ways that μ_i could be correlated with the learning values. However, if our observations generated from comparative simulations where there is no correlation substantively match the observations when the simulations have a high degree of correlation then the observations at least appear to be generalizable.[1]

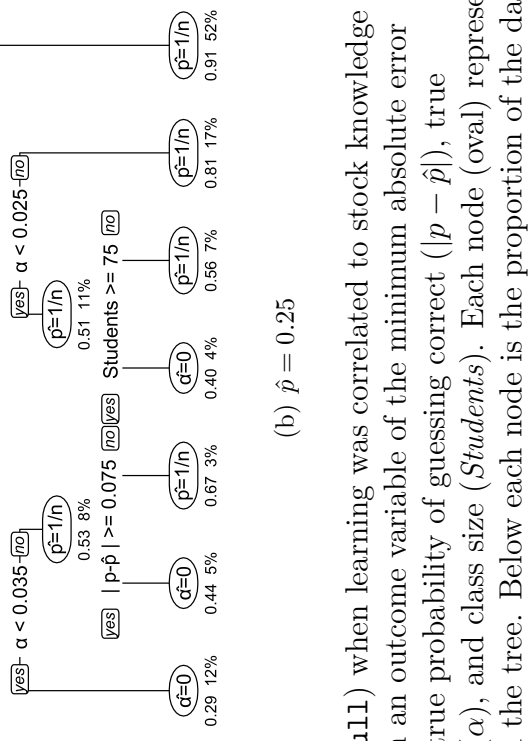
Using the simulations correlated to ability, we fit four Recursive Partitioning trees (Breiman et al., 1984) akin to Online Appendix A. Like the original trees, the outcome variable was whether the $1/n$ or $\hat{\alpha} = 0$ probability estimator produced more error in the

[1] Unlike the comparative simulations described in the main body of the paper, each distribution was created using only 1000 simulated classes (instead of 10,000). This allowed the procedure to complete in a reasonable amount of time.

learning estimate. While some break points changed slightly, the trees generated with this correlation assumption are remarkably similar to the those in Online Appendix A.

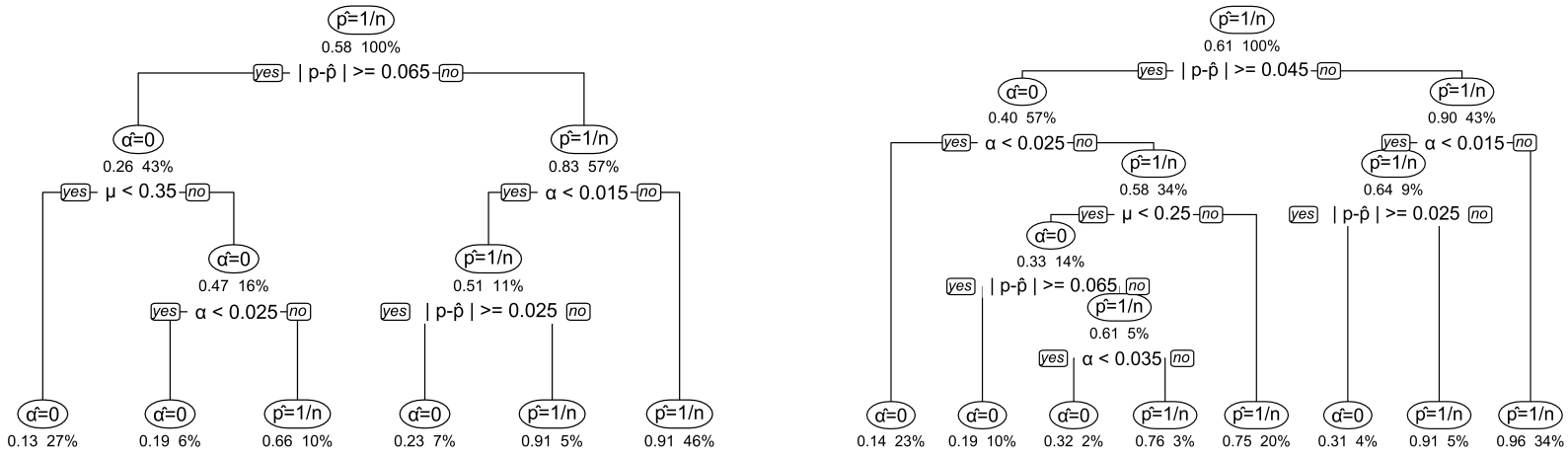


(a) $\hat{p} = 0.20$



(b) $\hat{p} = 0.25$

Figure D1. The simulated $\hat{\gamma}$ distributions (`http://bit.ly/MCSTFull`) when learning was correlated to stock knowledge were used to fit a Recursive Partitioning tree (Breiman et al., 1984) with an outcome variable of the minimum absolute error estimator. Features included were the absolute deviation from the true probability of guessing correct ($|p - \hat{p}|$), true underlying stock knowledge (μ), true underlying negative learning (α), and class size (*Students*). Each node (oval) represents the best (lowest absolute error) estimation strategy at that point in the tree. Below each node is the proportion of the data in that node where $\hat{p} = 1/n$ is the preferred strategy (first number) followed by the proportion of the overall training data that node represents. Figure (a) is trained on all $\hat{\gamma}$ simulation results where $p = 1/5$ and Figure (b) is trained on all $\hat{\gamma}$ simulation results where $p = 1/4$.



(a) $\hat{p} = 0.20$ (b) $\hat{p} = 0.25$

Figure D2. The simulated $\frac{\hat{\gamma}}{1-\hat{\mu}}$ distributions (`http://bit.ly/MCSTFull`) when learning was correlated to stock knowledge were used to fit a Recursive Partitioning tree (Breiman et al., 1984) with an outcome variable of the minimum absolute error estimator. Features included were the absolute deviation from the true probability of guessing correct ($|p - \hat{p}|$), true underlying stock knowledge (μ), true underlying negative learning (α), and class size. Each node (oval) represents the best (lowest absolute error) estimation strategy at that point in the tree. Below each node is the proportion of the data in that node where $\hat{p} = 1/n$ is the preferred strategy (first number) followed by the proportion of the overall training data that node represents. Figure (a) is trained on all $\frac{\hat{\gamma}}{1-\hat{\mu}}$ simulation results where $p = 1/5$ and Figure (b) is trained on all $\frac{\hat{\gamma}}{1-\hat{\mu}}$ simulation results where $p = 1/4$.

Appendix E

Comparative Monte Carlo Simulations

Table E1

γ estimator comparison when the $1/n$ estimator is deviated by 0.05 from the true probability of guessing correct

α	p	γ	μ	$\overline{\lvert\hat{\gamma}_{1/n} - \gamma\rvert}$	$\overline{\lvert\hat{\gamma}_Z - \gamma\rvert}$	$1/n$ Pref.
0.01	0.2	0.1	0.1	0.051	0.018	
0.01	0.3	0.1	0.1	0.043	0.022	
0.01	0.2	0.3	0.1	0.024	0.013	
0.01	0.3	0.3	0.1	0.019	0.017	
0.01	0.2	0.5	0.1	0.012	0.013	T
0.01	0.3	0.5	0.1	0.014	0.021	T
0.03	0.2	0.1	0.1	0.050	0.033	
0.03	0.3	0.1	0.1	0.043	0.037	
0.03	0.2	0.3	0.1	0.023	0.021	
0.03	0.3	0.3	0.1	0.020	0.024	T
0.03	0.2	0.5	0.1	0.012	0.016	T
0.03	0.3	0.5	0.1	0.015	0.024	T
0.05	0.2	0.1	0.1	0.050	0.053	T
0.05	0.3	0.1	0.1	0.044	0.057	T
0.05	0.2	0.3	0.1	0.024	0.033	T
0.05	0.3	0.3	0.1	0.020	0.035	T
0.05	0.2	0.5	0.1	0.012	0.017	T
0.05	0.3	0.5	0.1	0.015	0.026	T
0.01	0.2	0.1	0.3	0.036	0.016	
0.01	0.3	0.1	0.3	0.031	0.020	
0.01	0.2	0.3	0.3	0.013	0.010	
0.01	0.3	0.3	0.3	0.012	0.015	T
0.01	0.2	0.5	0.3	0.019		T
0.01	0.3	0.5	0.3	0.022		T
0.03	0.2	0.1	0.3	0.036	0.032	
0.03	0.3	0.1	0.3	0.031	0.036	T
0.03	0.2	0.3	0.3	0.012	0.014	T
0.03	0.3	0.3	0.3	0.012	0.017	T
0.03	0.2	0.5	0.3	0.020		T
0.03	0.3	0.5	0.3	0.022		T
0.05	0.2	0.1	0.3	0.036	0.052	T
0.05	0.3	0.1	0.3	0.031	0.056	T
0.05	0.2	0.3	0.3	0.013	0.022	T
0.05	0.3	0.3	0.3	0.012	0.023	T
0.05	0.2	0.5	0.3	0.020		T
0.05	0.3	0.5	0.3	0.022		T
0.01	0.2	0.1	0.5	0.022	0.014	
0.01	0.3	0.1	0.5	0.019	0.018	
0.01	0.2	0.3	0.5	0.010	0.017	T
0.01	0.3	0.3	0.5	0.011	0.026	T
0.03	0.2	0.1	0.5	0.022	0.031	T
0.03	0.3	0.1	0.5	0.019	0.034	T
0.03	0.2	0.3	0.5	0.010	0.023	T
0.03	0.3	0.3	0.5	0.011	0.033	T
0.05	0.2	0.1	0.5	0.022	0.051	T
0.05	0.3	0.1	0.5	0.019	0.055	T
0.05	0.2	0.3	0.5	0.010	0.027	T
0.05	0.3	0.3	0.5	0.011	0.036	T

Notes: These Monte Carlo simulations compare the true γ value to values produced by two different estimates of the probability of guessing correct. The $\hat{\gamma}_{1/n}$ estimator assumes the probability of guessing correct is equal to 1/4. The $\hat{\gamma}_Z$ estimator assume $\hat{\alpha} = 0$. These estimates of γ are compared to the true value in absolute terms and the averages across all 10,000 simulations are calculated. The column '$1/n$ Pref.' takes the value 'T' when the average absolute deviation of the $1/n$ estimator produced less average absolute deviation than the $\hat{\alpha} = 0$ estimator. In some cases, the $\hat{\gamma}_Z$ estimator can be very wrong: the blank cells indicate where at least one of the 10,000 simulations estimated the probability of guessing correct to be 100% (resulting in division by zero); this can occur by random chance alone. The results above are a subset of a much larger set of simulations available here: `https://goo.gl/2f8NSa`.

Table E2

γ estimator comparison when the $1/n$ estimator is deviated by 0.10 from the true probability of guessing correct

α	p	γ	μ	$\overline{\lvert\hat{\gamma}_{1/n} - \gamma\rvert}$	$\overline{\lvert\hat{\gamma}_Z - \gamma\rvert}$	$1/n$ Pref.
0.01	0.15	0.1	0.1	0.108	0.015	
0.01	0.35	0.1	0.1	0.079	0.025	
0.01	0.15	0.3	0.1	0.051	0.011	
0.01	0.35	0.3	0.1	0.030	0.019	
0.01	0.15	0.5	0.1	0.012	0.011	
0.01	0.35	0.5	0.1	0.022	0.027	T
0.03	0.15	0.1	0.1	0.108	0.032	
0.03	0.35	0.1	0.1	0.079	0.040	
0.03	0.15	0.3	0.1	0.051	0.020	
0.03	0.35	0.3	0.1	0.030	0.025	
0.03	0.15	0.5	0.1	0.012	0.013	T
0.03	0.35	0.5	0.1	0.022	0.030	T
0.05	0.15	0.1	0.1	0.108	0.052	
0.05	0.35	0.1	0.1	0.079	0.059	
0.05	0.15	0.3	0.1	0.051	0.032	
0.05	0.35	0.3	0.1	0.030	0.035	T
0.05	0.15	0.5	0.1	0.012	0.014	T
0.05	0.35	0.5	0.1	0.022	0.031	T
0.01	0.15	0.1	0.3	0.078	0.014	
0.01	0.35	0.1	0.3	0.056	0.023	
0.01	0.15	0.3	0.3	0.021	0.009	
0.01	0.35	0.3	0.3	0.013	0.018	T
0.01	0.15	0.5	0.3	0.037		T
0.01	0.35	0.5	0.3	0.043		T
0.03	0.15	0.1	0.3	0.077	0.031	
0.03	0.35	0.1	0.3	0.056	0.039	
0.03	0.15	0.3	0.3	0.021	0.013	
0.03	0.35	0.3	0.3	0.013	0.020	T
0.03	0.15	0.5	0.3	0.037		T
0.03	0.35	0.5	0.3	0.044		T
0.05	0.15	0.1	0.3	0.077	0.051	
0.05	0.35	0.1	0.3	0.056	0.058	T
0.05	0.15	0.3	0.3	0.021	0.021	
0.05	0.35	0.3	0.3	0.013	0.025	T
0.05	0.15	0.5	0.3	0.036		T
0.05	0.35	0.5	0.3	0.043		T
0.01	0.15	0.1	0.5	0.047	0.012	
0.01	0.35	0.1	0.5	0.033	0.020	
0.01	0.15	0.3	0.5	0.011	0.013	T
0.01	0.35	0.3	0.5	0.018	0.030	T
0.03	0.15	0.1	0.5	0.047	0.029	
0.03	0.35	0.1	0.5	0.033	0.036	T
0.03	0.15	0.3	0.5	0.011	0.019	T
0.03	0.35	0.3	0.5	0.018	0.036	T
0.05	0.15	0.1	0.5	0.047	0.050	T
0.05	0.35	0.1	0.5	0.033	0.057	T
0.05	0.15	0.3	0.5	0.012	0.023	T
0.05	0.35	0.3	0.5	0.018	0.040	T

Notes: These Monte Carlo simulations compare the true γ value to values produced by two different estimates of the probability of guessing correct. The $\hat{\gamma}_{1/n}$ estimator assumes the probability of guessing correct is equal to 1/4. The $\hat{\gamma}_Z$ estimator assume $\hat{\alpha} = 0$. These estimates of γ are compared to the true value in absolute terms and the averages across all 10,000 simulations are calculated. The column '$1/n$ Pref.' takes the value 'T' when the average absolute deviation of the $1/n$ estimator produced less average absolute deviation than the $\hat{\alpha} = 0$ estimator. In some cases, the $\hat{\gamma}_Z$ estimator can be very wrong: the blank cells indicate where at least one of the 10,000 simulations estimated the probability of guessing correct to be 100% (resulting in division by zero); this can occur by random chance alone. The results above are a subset of a much larger set of simulations available here: `https://goo.gl/2f8NSa`.

Table E3
$\gamma/(1-\mu)$ estimator comparison when the $1/n$ estimator is deviated by 0.05 from the true probability of guessing correct

α	p	γ	μ	$\overline{\left\lvert \frac{\hat{\gamma}_{1/n}}{1-\hat{\mu}_{1/n}} - \frac{\gamma}{1-\mu} \right\rvert}$	$\overline{\left\lvert \frac{\hat{\gamma}_{Z}}{1-\hat{\mu}_{Z}} - \frac{\gamma}{1-\mu} \right\rvert}$	$1/n$ Pref.
0.01	0.2	0.1	0.1	0.060	0.020	
0.01	0.3	0.1	0.1	0.059	0.026	
0.01	0.2	0.3	0.1	0.044	0.018	
0.01	0.3	0.3	0.1	0.044	0.022	
0.01	0.2	0.5	0.1	0.030	0.015	
0.01	0.3	0.5	0.1	0.030	0.018	
0.03	0.2	0.1	0.1	0.059	0.040	
0.03	0.3	0.1	0.1	0.059	0.044	
0.03	0.2	0.3	0.1	0.045	0.038	
0.03	0.3	0.3	0.1	0.045	0.041	
0.03	0.2	0.5	0.1	0.030	0.037	T
0.03	0.3	0.5	0.1	0.030	0.039	T
0.05	0.2	0.1	0.1	0.059	0.062	T
0.05	0.3	0.1	0.1	0.060	0.066	T
0.05	0.2	0.3	0.1	0.045	0.061	T
0.05	0.3	0.3	0.1	0.044	0.064	T
0.05	0.2	0.5	0.1	0.030	0.060	T
0.05	0.3	0.5	0.1	0.030	0.062	T
0.01	0.2	0.1	0.3	0.057	0.024	
0.01	0.3	0.1	0.3	0.057	0.031	
0.01	0.2	0.3	0.3	0.038	0.021	
0.01	0.3	0.3	0.3	0.039	0.025	
0.01	0.2	0.5	0.3	0.020	0.018	
0.01	0.3	0.5	0.3	0.020	0.019	
0.03	0.2	0.1	0.3	0.057	0.051	
0.03	0.3	0.1	0.3	0.057	0.056	
0.03	0.2	0.3	0.3	0.038	0.049	T
0.03	0.3	0.3	0.3	0.039	0.052	T
0.03	0.2	0.5	0.3	0.020	0.047	T
0.03	0.3	0.5	0.3	0.021	0.048	T
0.05	0.2	0.1	0.3	0.057	0.081	T
0.05	0.3	0.1	0.3	0.057	0.085	T
0.05	0.2	0.3	0.3	0.039	0.079	T
0.05	0.3	0.3	0.3	0.039	0.082	T
0.05	0.2	0.5	0.3	0.020	0.076	T
0.05	0.3	0.5	0.3	0.021	0.078	T
0.01	0.2	0.1	0.5	0.053	0.032	
0.01	0.3	0.1	0.5	0.054	0.041	
0.01	0.2	0.3	0.5	0.028	0.026	
0.01	0.3	0.3	0.5	0.029	0.030	T
0.03	0.2	0.1	0.5	0.053	0.072	T
0.03	0.3	0.1	0.5	0.054	0.079	T
0.03	0.2	0.3	0.5	0.028	0.068	T
0.03	0.3	0.3	0.5	0.028	0.072	T
0.05	0.2	0.1	0.5	0.054	0.114	T
0.05	0.3	0.1	0.5	0.054	0.122	T
0.05	0.2	0.3	0.5	0.027	0.110	T
0.05	0.3	0.3	0.5	0.029	0.114	T

Notes: These Monte Carlo simulations compare the true $\gamma/(1-\mu)$ value to values produced by two different estimates of the probability of guessing correct. The $\hat{\gamma}_{1/n}/(1-\hat{\mu}_{1/n})$ estimator assumes the probability of guessing correct is equal to 1/4. The $\hat{\gamma}_Z/(1-\hat{\mu}_Z)$ estimator assumes $\hat{\alpha}=0$. These estimates of $\gamma/(1-\mu)$ are compared to the true value in absolute terms and the averages across all 10,000 simulations are calculated. The column '$1/n$ Pref.' takes the value 'T' when the average absolute deviation of the $1/n$ estimator produced less average absolute deviation than the $\hat{\alpha}=0$ estimator. The results above are a subset of a much larger set of simulations available here: `https://goo.gl/2f8NSa`.

Table E4
$\gamma/(1-\mu)$ estimator comparison when the $1/n$ estimator is deviated by 0.10 from the true probability of guessing correct

α	p	γ	μ	$\overline{\left\lvert \frac{\hat{\gamma}_{1/n}}{1-\hat{\mu}_{1/n}} - \frac{\gamma}{1-\mu} \right\rvert}$	$\overline{\left\lvert \frac{\hat{\gamma}_{Z}}{1-\hat{\mu}_{Z}} - \frac{\gamma}{1-\mu} \right\rvert}$	$1/n$ Pref.
0.01	0.15	0.1	0.1	0.119	0.018	
0.01	0.35	0.1	0.1	0.119	0.029	
0.01	0.15	0.3	0.1	0.089	0.016	
0.01	0.35	0.3	0.1	0.089	0.024	
0.01	0.15	0.5	0.1	0.059	0.014	
0.01	0.35	0.5	0.1	0.059	0.020	
0.03	0.15	0.1	0.1	0.119	0.038	
0.03	0.35	0.1	0.1	0.119	0.047	
0.03	0.15	0.3	0.1	0.089	0.037	
0.03	0.35	0.3	0.1	0.089	0.043	
0.03	0.15	0.5	0.1	0.059	0.036	
0.03	0.35	0.5	0.1	0.059	0.040	
0.05	0.15	0.1	0.1	0.119	0.061	
0.05	0.35	0.1	0.1	0.118	0.069	
0.05	0.15	0.3	0.1	0.089	0.060	
0.05	0.35	0.3	0.1	0.089	0.066	
0.05	0.15	0.5	0.1	0.059	0.059	
0.05	0.35	0.5	0.1	0.059	0.063	T
0.01	0.15	0.1	0.3	0.115	0.022	
0.01	0.35	0.1	0.3	0.114	0.035	
0.01	0.15	0.3	0.3	0.076	0.019	
0.01	0.35	0.3	0.3	0.076	0.028	
0.01	0.15	0.5	0.3	0.038	0.017	
0.01	0.35	0.5	0.3	0.038	0.021	
0.03	0.15	0.1	0.3	0.114	0.049	
0.03	0.35	0.1	0.3	0.114	0.060	
0.03	0.15	0.3	0.3	0.076	0.047	
0.03	0.35	0.3	0.3	0.076	0.054	
0.03	0.15	0.5	0.3	0.038	0.046	T
0.03	0.35	0.5	0.3	0.038	0.050	T
0.05	0.15	0.1	0.3	0.114	0.079	
0.05	0.35	0.1	0.3	0.115	0.088	
0.05	0.15	0.3	0.3	0.076	0.077	T
0.05	0.35	0.3	0.3	0.076	0.084	T
0.05	0.15	0.5	0.3	0.038	0.075	T
0.05	0.35	0.5	0.3	0.038	0.079	T
0.01	0.15	0.1	0.5	0.107	0.029	
0.01	0.35	0.1	0.5	0.106	0.045	
0.01	0.15	0.3	0.5	0.053	0.024	
0.01	0.35	0.3	0.5	0.054	0.033	
0.03	0.15	0.1	0.5	0.107	0.069	
0.03	0.35	0.1	0.5	0.107	0.084	
0.03	0.15	0.3	0.5	0.053	0.066	T
0.03	0.35	0.3	0.5	0.053	0.074	T
0.05	0.15	0.1	0.5	0.107	0.111	T
0.05	0.35	0.1	0.5	0.107	0.126	T
0.05	0.15	0.3	0.5	0.053	0.108	T
0.05	0.35	0.3	0.5	0.054	0.116	T

Notes: These Monte Carlo simulations compare the true $\gamma/(1-\mu)$ value to values produced by two different estimates of the probability of guessing correct. The $\hat{\gamma}_{1/n}/(1-\hat{\mu}_{1/n})$ estimator assumes the probability of guessing correct is equal to 1/4. The $\hat{\gamma}_Z/(1-\hat{\mu}_Z)$ estimator assumes $\hat{\alpha}=0$. These estimates of $\gamma/(1-\mu)$ are compared to the true value in absolute terms and the averages across all 10,000 simulations are calculated. The column '$1/n$ Pref.' takes the value 'T' when the average absolute deviation of the $1/n$ estimator produced less average absolute deviation than the $\hat{\alpha}=0$ estimator. The results above are a subset of a much larger set of simulations available here: `https://goo.gl/2f8NSa`.

References

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and regression trees*. Taylor & Francis.