STRATEGIC OMISSION AND RISK AVERSION: A BIAS-RELIABILITY TRADEOFF

David Lang
Stanford University
david.nathan.lang@stanford.edu

ABSTRACT: Whether high-stakes exams such as the SAT or College Board AP exams should penalize incorrect answers is a controversial question. In this paper, we document that penalty functions can have differential effects depending on a student's risk tolerance. Moreover, literature shows that risk aversion tends to vary along other areas of concern such as race, gender, nationality, and socioeconomic status. In this article, we simulate Item Response Theory (IRT) data with and without a wrong answer penalty. In the presence of mild risk aversion, we find that students omit 12% more items than risk neutral individuals with identical ability. This translates into a nearly 2% difference in sum scores between the risk neutral and risk averse groups. We also find that penalty functions result in noisier estimates of student ability. These findings suggest that random guessing penalties should not be used in most circumstances, particularly for learning platforms.

Keywords: learning analytics, item response theory, risk aversion, differential item function, differential test function, simulation

1 MOTIVATION

In the past decade there have been notable shifts in the decision to penalize wrong answers in high-stakes testing. In 2010, the College Board removed its wrong answer penalty for the AP exams. The SAT has also removed this penalty from its exams in recent years.

In this paper, we explore whether learning platforms should follow suit. Many platforms implicitly or explicitly penalize guessing through either gamification mechanisms such as point systems or through hint generation. These designs often are associated with increased user engagement or performance but they may have downstream impacts on certain types of users (O'Rourke, Haimovitz, & Ballweber, 2014). Simulation may help us understand how these design features influence student behavior.

2 LITERATURE REVIEW

While much of the literature surrounding high-stakes testing has focused on bias in terms of gender and race/ethnicity, relatively little focus has been put forth into the effects of how random guessing penalties may mediate this bias. Past work points out that most exams with a penalty function are still designed so that a person who tries to maximize their average score will be indifferent to always guessing (Budescu & Bar-Hillel, 1993). Moreover, they point out that this penalty function introduces systematic biases for students. If students have a different objective (e.g. get a passing grade or get the top grade in the class), then these incentives may not hold. Other work found that there were substantial differences by gender in willingness to guess in the face of a penalty function (Baldiga, 2013). To date, there has been even less focus on how risk aversion affects the psychometric properties of these assessments.

2.1 Risk Aversion

There are three broad classifications of risk tolerance: risk-aversion, risk-preferring, and risk-neutrality. To understand these distinctions, consider a coin-flip bet where a person wins a dollar if the coin lands heads and loses a dollar if the coin lands tails. A risk averse person will never take a bet with an average payoff of zero. A risk-preferring person will always take this bet. The risk neutral person will be indifferent between taking this bet and not taking this bet.

In this paper, we model risk aversion using an exponential utility function:

$$U(points, risk_{tolerance}) = \frac{1 - e^{-points * risk_{tolerance}}}{risk_{tolerance}}$$

The components of the function are points (the number of points awarded or lost) and risk tolerance. Positive risk-tolerance parameters correspond to risk-aversion. Negative risk-tolerance parameters correspond to a risk-preferring behavior. In a testing framework, if the utility of attempting a question is positive, the examinee will attempt it. Otherwise, the examinee will omit it. This function exhibits several useful properties. First, it exhibits a constant coefficient of relative risk aversion. In decision analysis literature, this property is also known as the 'delta property' (Kirkwood, 1997). This property assures that an individual will have the same preferences regardless of their current wealth endowment. In a testing framework, this means that an individual's decision to omit a particular item will not depend on one's current score. This assumption is fairly reasonable for small scale decisions, such as one question on a forty-question exam. Additional benefits of this assumption are that it eliminates concerns with respect to item ordering effects interacting with risk aversion, and unlike other potential utility functions, this function can be transformed into a risk-averse/risk-preferring function simply by assigning a positive/negative risk tolerance value.

In terms of understanding what risk aversion looks like in the real world, most estimates suggest that individuals have positive risk tolerance and that a risk tolerance parameter of one is not unreasonable (Gandelman & Hernández-Murillo, 2014). Figure 1 shows that point estimates of risk aversion in the United States is around 1.5 . The most extreme countries are the Netherland with a risk tolerance of less than a quarter and Taiwan with a risk tolerance of nearly 2.5.

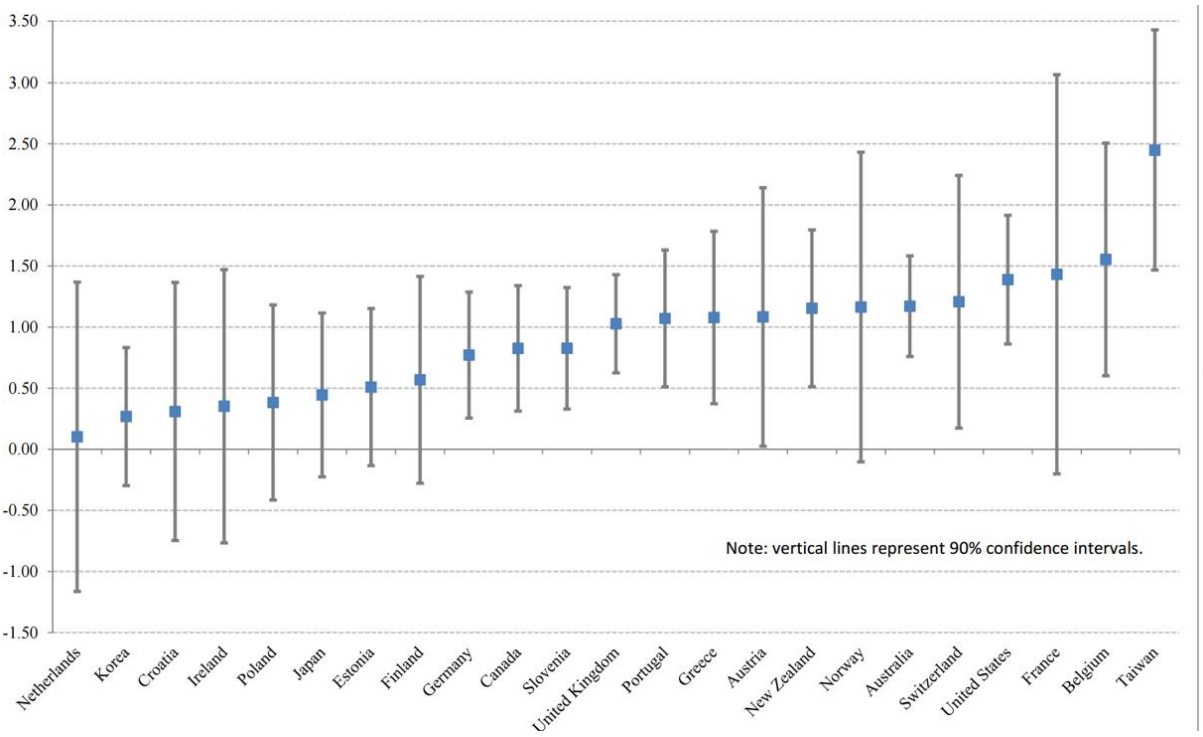

Figure 1 Relative Risk Aversion in Developed Countries (Source:St Louis Fed)

3 MODEL

To assess the question of omission on exams, we simulate a forty-question exam. The exam data is modeled as Rasch data such that each individual's true ability estimate is known to us. The probability that a student will answer an item correct can be expressed by the following formula where θ_i corresponds to the ability of student ***i*** and b_j corresponds to the difficulty of item ***j***:

$$\frac{1}{1+e^{(\theta_i - b_j)}}$$

We further assume that students are aware of their ability and item difficulty but are uncertain whether or not they get the specific item correct. We also assume that they are aware of a one-quarter point penalty if they answer a question incorrectly. In this case, the students will respond to an item only if the expression below holds:

$$\Pr\left(Correct|\theta_i, b_j\right) * U(1, risk_{tolerance}) + (1 - \Pr\left(Correct|\theta_i, b_j\right)) * U\left(-\frac{1}{4}, risk_{tolerance}\right) \geq 0$$

We then re-estimate a person's ability based on their responses under three separate scenarios: (1) no penalty, (2) risk-neutrality, (3) risk-aversion with a risk tolerance of 1. We then repeatedly estimate the difference between these three groups and our true ability measures to assess whether or not this biases estimates of test performances. The underlying data generation process assumes both ability and item difficulty follow the standard normal distribution.

Figure 2 illustrates the utility of responding to a question in which the student is aware of the probability they will get the question right. The horizontal line at zero identifies the locations at which students of varying risk tolerances will be indifferent to answering the question and omitting their response. Points above the zero line correspond to attempting the item. Points below the line correspond to omitting the item. The dashed-line corresponds to a risk neutral student. For risk-preferring students, students with a risk preference of three will "guess" if their probability of getting the question right is at least 3%. The most risk averse student would not respond unless they had at least a 55% chance of getting the question correct.

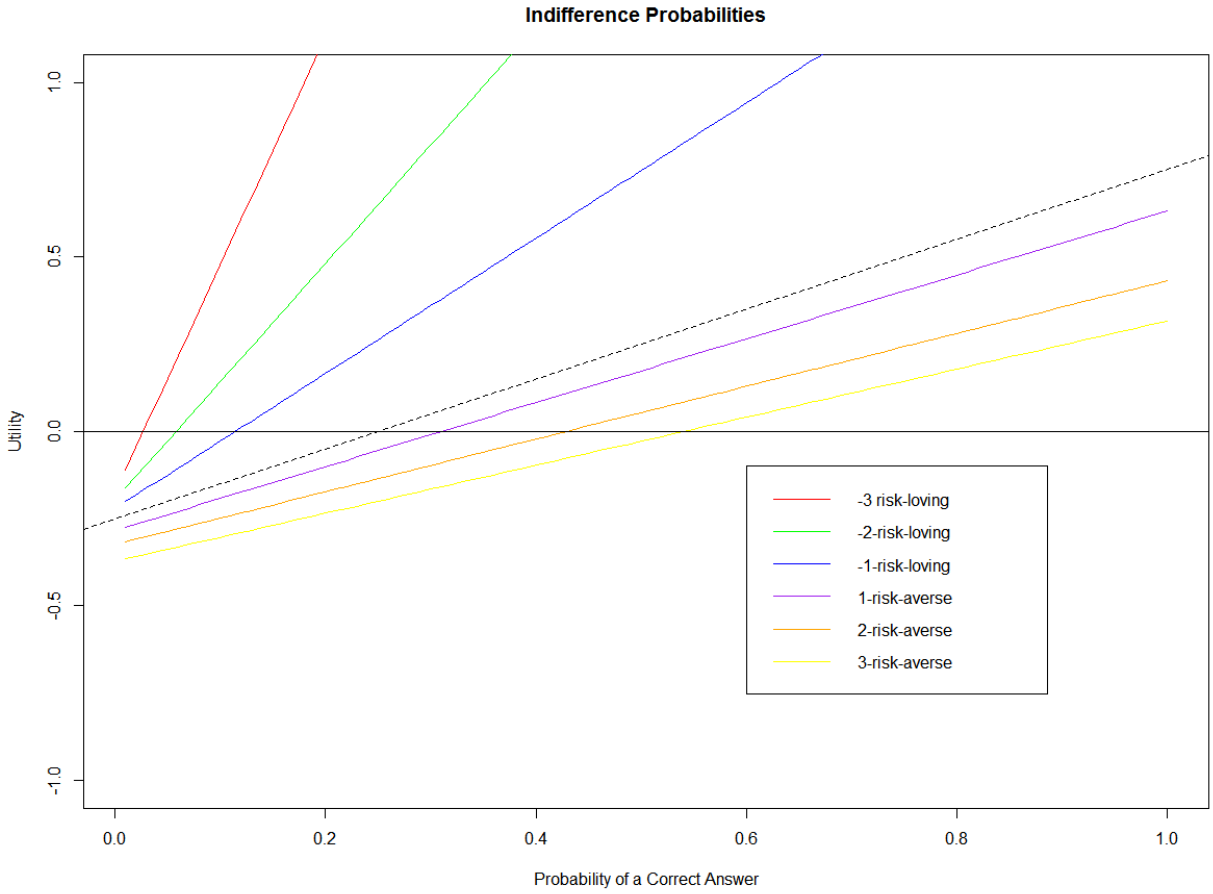

Figure 2 Indifference Probabilities and Utility

4 SIMULATIONS

A hundred bootstrapped simulations were run to better estimate the effects of strategic omission. Repeated simulations yields the omission rates plots below. On average, a risk-neutral simulation yields an omissions rate of 18%. In the risk-averse case, this omission rate jumps up to approximately 30%. Sum scores change relatively little with only a two percentage point difference in exam performance (Figure 3).

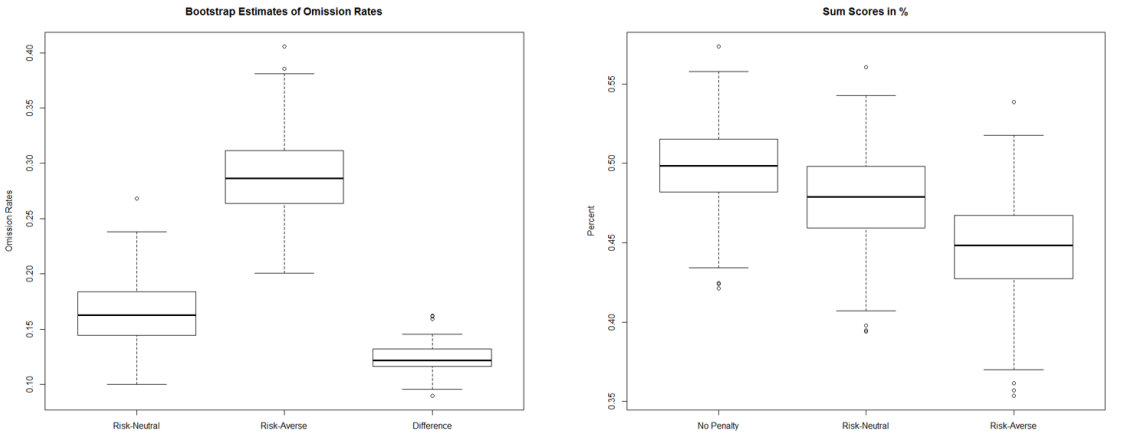

Figure 3 Bootstrapped Estimates of Sum Scores

4.1 Ability Measurement Error

By introducing a penalty, it introduces a large region where low ability individuals will not attempt certain items. This makes distinguishing between low ability people and very low ability people extremely difficult. From a maximum likelihood estimation perspective, this means that for each item there is a portion of the information curve where the estimate is completely flat. An illustration of that fact can be seen in Figure 4.

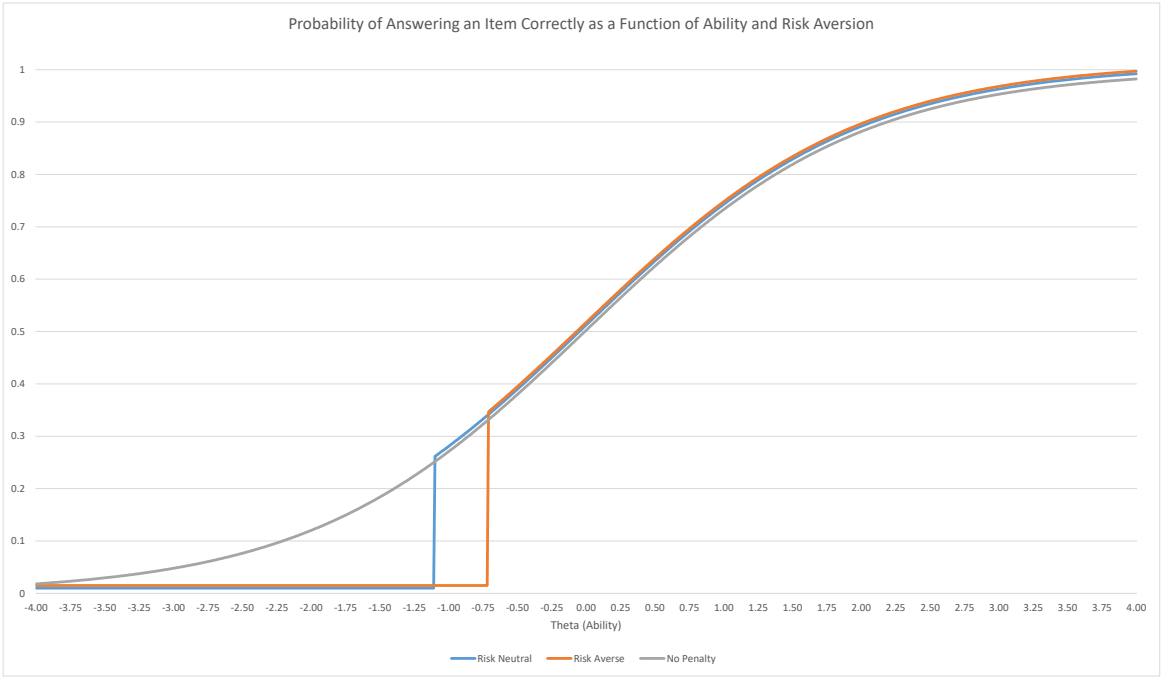

Figure 4 Probability of Answering an Item Correctly as a Function of Ability and Risk Aversion

We also recover individual ability estimates using a Rasch model and maximum likelihood. Estimates of these data yield unbiased estimates of an individual's ability (See Figure 3). The

mean absolute deviations of theta increases as the penalty function is introduced and as the risk-aversion increases. As such, the amount of error in ability measurements is nearly twice as large for a risk-averse population than if there were no penalties enacted on the same population of students (See Figure 5).

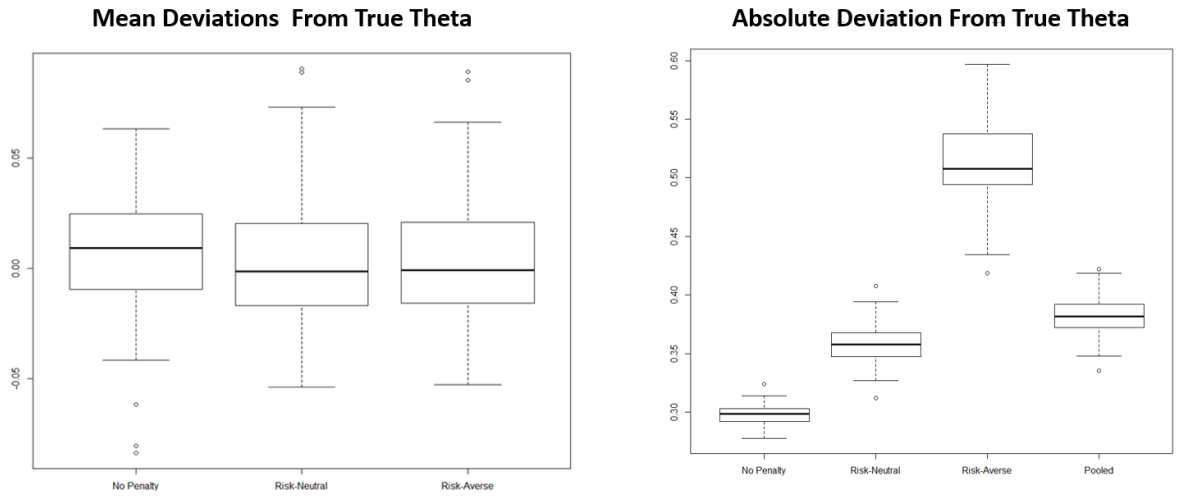

Figure 5 Absolute Mean Deviations of Ability

4.2 Reliability

So the fundamental question is why are these penalty functions used if it increases non-response rates and seems to introduce these potential claims of bias. One possible explanation is that improves measures of reliability. We compute the reliability of the generated exams using Cronbach's alpha (Cronbach, 1951). The boxplots below show that reliability increases if students are given an incentive to omit incorrect answers. This effect still holds even if one assumes heterogeneity of risk tolerance amongst users (See Figure 3). In effect, what happens is that users who have relatively low likelihood of getting an item correct through random guessing gets their answer compressed to zero in response to a penalty. This omission, in turn, increases the reliability of an exam.

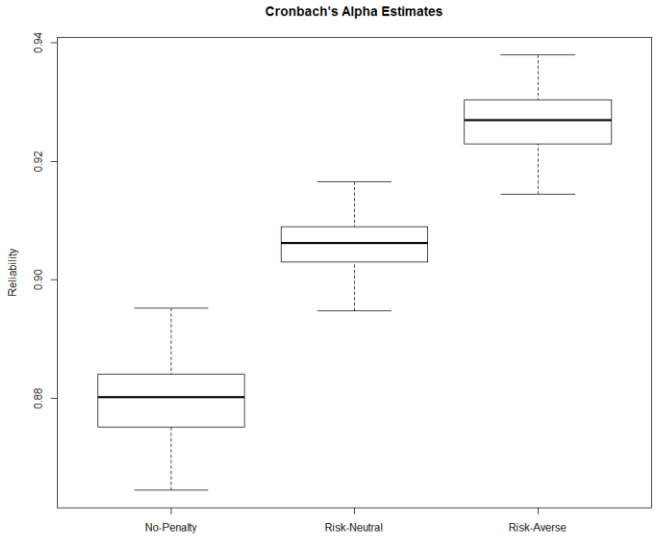

Figure 6 Cronbach's Alpha (Reliability)

5 DISCUSSION

From a reliability perspective, penalizing exams has some benefits. Introducing penalties tends to increase the reliability of the exams. This increase in reliability comes at the cost of certain measures becoming noisy. Further, if there's heterogeneity of risk aversion, it's possible that the rank ordering of students could jump noticeably when an exam switches from a penalty function to an exam without a penalty function. Strategic omission makes generating distinctions between the bottom-half of the distribution very difficult. To the extent that an exam is concerned with generating a precise estimate of ability, utilizing a penalty function is ill-advised.

The only cases where a guessing penalty could make sense are when risk tolerance is a parameter that is also being trained. For instance this type of penalty function could be useful when training actuaries, financial investors, or stockbrokers. The rationale for this is that their score would be both a composition of their true ability and their risk tolerance.

5.1 Implications for Learning Analytics and Platform Design

This works suggests that penalties should not be used for assessment purposes. If individuals are penalized for wrong answers, then risk-averse users will strategically omit more responses than risk-tolerant users. In turn, this means that learning platforms would direct risk-averse users into more remedial content than similar ability students who are risk-neutral. To the extent that these populations are underserved groups (females, underrepresented minorities, and low socioeconomic status), embedding penalties for random guessing could deter these groups from interacting with the platform and replicate existing inequalities. Further, our simulations suggest that guessing penalties may make it more difficult for learning platforms to distinguish between users in the lower end of the ability distribution. These are often the groups that are of focal interest to learning analytics researchers and policy makers.

Many learning platforms reward users with points or badges for engaging with the platform and penalize users for using built-in hint generation features. Removing penalties from these contexts

seem like a natural decision. Generally, these penalties should be removed when items are being used as part of a formative assessment.

If random guessing penalties are to be used in a summative assessment, there are approaches that mitigate the performance bias between risk-averse and risk neutral users. One of the design choices is to allow students to respond to multiple items before submitting a response for grading. This will allow rational agents to hedge their responses and makes risk-averse users more likely to respond so long as their knowledge is truly better than random guessing.

6 ACKNOWLEDGEMENTS

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant (#R305B140009). We also acknowledge thoughtful reviews from the conference organizers.

REFERENCES

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*. Retrieved from http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2013.1776

Budescu, D., & Bar-Hillel, M. (1993). To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring. *Journal of Educational Measurement*, *30*(4), 277–291. https://doi.org/10.1111/j.1745-3984.1993.tb00427.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Gandelman, N., & Hernández-Murillo, R. (2014). Risk Aversion at the Country Level. *FRB of St. Louis Working …*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2396103

Kirkwood, C. (1997). Notes on attitude toward risk taking and the exponential utility function. *Department of Management, Arizona State University …*. Retrieved from http://www.public.asu.edu/~kirkwood/DAStuff/refs/risk.pdf

O'Rourke, E., Haimovitz, K., & Ballweber, C. (2014). Brain points: a growth mindset incentive structure boosts persistence in an educational game. *Proceedings of The*. Retrieved from http://dl.acm.org/citation.cfm?id=2557157