From grades to learning: A philosophical enquiry into confirmation and evidence in educational measurement

Tony C. A. Tan

Centre for Educational Measurement

University of Oslo

UV9002: Philosophy of Science

Prof Sten R. Ludvigsen

23 March 2023

From grades to learning: A philosophical enquiry into confirmation and evidence in educational measurement

"A Pope is a Catholic" does not imply "a Catholic is a Pope". Symbolically,

$$\mathbb{P}\left(\text{Catholic} \mid \text{Pope}\right) \neq \mathbb{P}\left(\text{Pope} \mid \text{Catholic}\right)$$

where $\mathbb{P}\left(\text{event} \mid \text{information}\right)$ denote the probability of an event (being true) in light of some information. Mixing up the order of the conditional probability is a common mistake in reasoning, frequent enough to earn its name of "inverse probability fallacy" (Kalinowski et al., 2008) with severe, even tragic, consequences (Hill, 2005; Thompson & Schumann, 1987). I wish to argue in this paper that the same is true for the concept of evidence in educational measurement such that "learning, given grades" must not be confused with "grades, given learning".

As an integral part of educational practices, grades have long been used to quantify learning. As a measurement device, a test maps students' learning (the "cause") onto an ordinal scale (the "effect"), whose numeric readings are the grades. Map-makers report the goodness of their device by the degree to which the grades faithfully reflect the underlying learning — knowing a student has high competency or great amount of learning, the measurement device should report a high grade. Should the measurement procedure correspond the "effect" closely to its "cause", this device is said to be valid, a necessary but not sufficient condition for its use in educational practices. A valid device must also be reliable, that is, the grades reported by the device shall be consistent across repeated measurements (AERA et al., 2014).

Validity and reliability, however, jointly promote the accuracy of "grades, given learning" but *not* "learning, given grades" — while test-makers concern themselves with the former, society at large is more keenly interested in the latter. Such subtlety could be responsible for the news headline "COVID-19 is learning-neutral" after repeated efforts failed to demonstrate sizeable drops in "learning outcomes". An effect can remain stable even when the underlying cause shifts thanks to interventions of auxiliary measures — leniency and grade inflation being immediate candidates during COVID lockdowns, or the low sensitivity of the measurement device (e.g., early astronomers' failure of detecting parallax). Should one be

convinced by the incongruence between "grades, given learning" and "learning, given grades", a natural extension would be how to "flip" the causal arrow such that the latent construct learning can be inferred from the observed grades.

Enters Bayes formula, the mathematical logic for reasoning about conditional probabilities. Let L and G denote learning and grades, respectively, then the conditional probability of learning given grades can be inferred from the observable data through the bridge

$$\mathbb{P}(L \mid G) = \mathbb{P}(G \mid L) \cdot \frac{\mathbb{P}(L)}{\mathbb{P}(G)}$$

In practice, the denominator $\mathbb{P}(G)$ offers least insight as it is a constant for a given measurement device and can be numerically approximated by computer algorithms[1]. This reduces the Bayes reasoning at the philosophical-level to this core version

$$\mathbb{P}(L \mid G) \propto \mathbb{P}(G \mid L) \cdot \mathbb{P}(L)$$

where $\propto$ denotes "proportional to" or "subject to a normalising constant". There could be no dilemma between "grades, given learning" and "learning, given grades" as soon as both $\mathbb{P}(G \mid L)$ and $\mathbb{P}(L)$ become known, at least theoretically.

Practically, however, both terms demand additional examination. The $\mathbb{P}(G \mid L)$ component reflects properties of the mapping device whereas $\mathbb{P}(L)$ is the prior probability of learning, a measure of the existing or "base rate" of learning. The empirical estimation of $\mathbb{P}(G \mid L)$ is a technical one involving ascertaining the properties of the test instrument. The first two papers of my thesis focus on this aspect of the problem. The epistemological aspect of the Bayes reasoning resides in the final term $\mathbb{P}(L)$ since it declares that *all knowledge must be built based on existing knowledge*. Such demand opens up a philosophical debate on where, how, and who has the power to obtain the first piece of the prior knowledge $\mathbb{P}(L)$.

A second layer of the epistemological debate concerns itself with the objectivity of science. If all "hard evidence" (objectivity) must be blended with "prior knowledge" (subjectivity) before becoming "knowledge", does the resulting scientific knowledge remain independent from the observer/scientist at all—in the educational measurement context, one

[1]

$$\mathbb{P}(G = \boldsymbol{g}) = \int_{\boldsymbol{l} \in L} \mathbb{P}(G = \boldsymbol{g} \mid L = \boldsymbol{l}) \cdot \mathbb{P}(L = \boldsymbol{l}) \, \mathrm{d}\boldsymbol{l}$$

may question whether there exists learning in the absence of grades. If the answer is “no”, then the Bayes reasoning is not a scientific one at all. If the answer is “yes”, then the Bayes reasoning is a scientific one but the “learning” is not a scientific one. I wish to address this debate in the second stage of this course.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Hill, R. (2005). Reflections on the cot death cases. *Significance*, *2*(1), 13–16. https://doi.org/10.1111/j.1740-9713.2005.00077.x

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy. *Methodology*, *4*(4), 152–158. https://doi.org/10.1027/1614-2241.4.4.152

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, *11*(3), 167–187. https://doi.org/10.1007/BF01044641