philosophies

Article

On Educational Assessment Theory: A High-Level Discussion of Adolphe Quetelet, Platonism, and Ergodicity

Patrick Francis Bloniasz [1,2,3]

[1] Program in Neuroscience, Bowdoin College, Brunswick, ME 04011, USA; pblonias@bowdoin.edu or patrick.bloniasz@gmail.com
[2] Program in Digital and Computational Studies, Bowdoin College, Brunswick, ME 04011, USA
[3] Department of Education, Aeon for Ocean, Eden Prairie, MN 55347, USA

Abstract: Educational assessments, specifically standardized and normalized exams, owe most of their foundations to psychological test theory in psychometrics. While the theoretical assumptions of these practices are widespread and relatively uncontroversial in the testing community, there are at least two that are philosophically and mathematically suspect and have troubling implications in education. Assumption 1 is that repeated assessment measures that are calculated into an arithmetic mean are thought to represent some real stable, quantitative psychological trait or ability plus some error. Assumption 2 is that aggregated, group-level educational data collected from assessments can then be interpreted to make inferences about a given individual person over time without explicit justification. It is argued that the former assumption cannot be taken for granted; it is also argued that, while it is typically attributed to 20th century thought, the assumption in a rigorous form can be traced back at least to the 1830s via an unattractive Platonistic statistical thesis offered by one of the founders of the social sciences—Belgian mathematician Adolphe Quetelet (1796–1874). While contemporary research has moved away from using his work directly, it is demonstrated that cognitive psychology is still facing the preservation of assumption 1, which is becoming increasingly challenged by current paradigms that pitch human cognition as a dynamical, complex system. However, how to deal with assumption 1 and whether it is broadly justified is left as an open question. It is then argued that assumption 2 is only justified by assessments having ergodic properties, which is a criterion rarely met in education; specifically, some forms of normalized standardized exams are intrinsically non-ergodic and should be thought of as invalid assessments for saying much about individual students and their capability. The article closes with a call for the introduction of dynamical mathematics into educational assessment at a conceptual level (e.g., through Bayesian networks), the critical analysis of several key psychological testing assumptions, and the introduction of dynamical language into philosophical discourse. Each of these prima facie distinct areas ought to inform each other more closely in educational studies.

Keywords: educational assessment theory; dynamical educational assessment; ergodicity; critical psychology; psychological testing; Adolphe Quetelet; educational Platonism; assessment validity

check for updates

Citation: Bloniasz, P.F. On Educational Assessment Theory: A High-Level Discussion of Adolphe Quetelet, Platonism, and Ergodicity. *Philosophies* **2021**, *6*, 46. https://doi.org/10.3390/philosophies6020046

Academic Editor: Andrew Stables

Received: 26 April 2021
Accepted: 2 June 2021
Published: 4 June 2021

1. Introduction

Education is a curious discipline in that it does not, in the strictest sense, have its own clear methodology. What cohesion does exist in the field has come from the fusion of educational practitioners and various disciplines including, but not limited to, history, psychology, sociology, and philosophy [1]. Further, to a large extent, contemporary education has only come to know itself through the language of business after the rise of neo-liberal philosophies (e.g., one 'delivers' curriculum, content ought to be 'efficient').[1] Ultimately, I believe such an interdisciplinary approach is a strength of educational studies. Still,

[1] For critical educational anthology, see [2].

with so many actors using different methodologies with different ideological structures in the Ricoeurian sense [3], coupled with education having its norms dictated within the political sphere, there are many problematic features of education, even when using critical pedagogy, that have failed to be adequately discussed.

One of the many 'arms' of education that helps push the discipline forward as an accumulating body of diverse perspectives is educational assessment—with the most recognizable form of assessment in contemporary times being seen in standardized tests. Standardized educational assessment carries at least two features largely ignored by contemporary practitioners—both seemingly problematic. The first is a Platonistic-esque assumption found in standardized educational assessment (e.g., College Board's SAT) and cognitive psychology that has yet to be fully grappled with after psychology's empirical turn. The assumption can be traced back to at least Adolphe Quetelet's (1796–1874) explicitly Platonistic work; Quetelet is a Belgian mathematician and astronomer who drastically impacted the social sciences even though he is largely ignored in education [4]. I argue that the assumption, namely that *the arithmetic mean from psychological tools 'picks out' some stable and quantitative psychological property*, is derivative from Platonism, but was never properly addressed in psychology due to Quetelet's work being swept up in several historical movements (e.g., Mental Testing and Eugenics); as such, Quetelet became understudied until recently [5].[2]

This assumption about the arithmetic mean essentially modeling some stable and quantitative psychological property is closely related to a consequential difference in the definition of 'measurement', which is discussed thoroughly by Michell [6]. It might be helpful to understand the split definitions for clarity. In the natural sciences, measurement is of quantitative attributes where "[q]uantitative attributes are distinguished from non-quantitative attributes by the possession of additive structure" ([7], p. 401). For realists about measurement, like myself, it is typically assumed that the question of whether a given attribute has additive structure requires experimental data. For example, does the systematic assignment of the unit kilogram (kg) to the hypothesized attribute of an object's mass make sense as something that is additive (assuming we did not already use mass as a quantitative metric)? Suppose an object, O1, has what is defined as 1 kg in mass and a duplicate of O1, O2, is combined with O1. If the expected unit mass is additive to twice itself, and indeed is so empirically irrespective of irrelevant conditions (e.g., the time they are stacked does not matter, nor does the time of day), this would be empirical evidence that mass is quantitative. It follows here that "[t]he issue of whether or not any attribute is measurable is an empirical one" ([7], p. 401). This, however, is distinct from the other definition of measurement that is attributable to thinkers like Stevens (e.g., [8,9]) and Kelley [10] where measurement is merely the systematic (i.e., rule-based) assignment of numbers to objects or events ([11], pp. 650–659 for discussion). In fact, Kelley, a student of the arguable founder of modern educational psychology, Edward Lee Thorndike, claimed the following:

> Our mental tests measure something, we may or may not care what, but it is something which it is to our advantage to measure, for it augments our knowledge of what people can be counted on to do in the future. The measuring device as a measure of something that it is desirable to measure comes first, and what it is a measure of comes second ([10], p. 86).

In this way, it has been claimed that "[a]s Pythagoreans [who believe that nature is quantitative in structure *simpliciter* and such structure can be uncovered via mathematics], psychometricians accepted it as an article of faith that psychological tests must measure something, even if it is not known what" ([11], p. 655). In terms of the history of philosophy, (Neo)pythagoreanism was mostly absorbed into (Neo)platonism in 4 CE [12]. In engaging with Quetelet's work, I will show that assumption one is not merely an ancient idea reborn

[2] Note that "Platonism" and "Neoplatonism" are being used here for the sake of simplicity, but Quetelet can be considered as a mix between Neoplatonist and Platonism as they are traditionally defined; his position will become clearer through the quotations of his writing later in this work.

in the mid-20th century, but rather that the assumption has persisted through time and through various intellectual movements.

As I will show, this distinction in the definition of measurement is relevant to modern assessment because, for someone like Michell, it must be demonstrated empirically that there is some structure that is explicitly additive to make claims about some stable psychological trait prior to creating an operational definition, whereas for Stevens or Kelley it is enough to systematically assign values to a system to uncover patterns. For example, for Michell it would not make sense to measure intelligence prior to knowing what intelligence is and showing the structure itself is quantitative, whereas for Kelley it makes sense to guess at what intelligence is and try to measure it with a systematic but arbitrary rule in order to help predict the actions or ability of a given person in the context of group trends—though more on this in Section 3. This difference shows that those like Stevens have an extra burden of proof on their rule-based approach beyond whether or not a tool is considered valid and therefore their assumption is not applicable everywhere. The general approach being critiqued is that of those who take measurement in the sense of Kelley and S.S. Stevens; as such, the word 'measure' should be taken in this sense and not in Michell's sense, meaning that the object of measurement could have attributes with a justifiably quantitative structure (e.g., distance), but it is not necessarily so.[3]

The second problematic feature of standardized educational assessment discussed comes from the fact that ergodic theory and dynamical systems theory have failed to make their way into education, even though many cognitive scientists think of human cognition in dynamical terms [13–15] and make claims as if they took such a dynamical model seriously. If human cognition is dynamic, we cannot take ergodicity, a dynamical systems concept that allows us to make interpretations from aggregated group data down to an individual, for granted in education and psychology. As such, I claim many of our interpretations of summative and normalized standardized assessments, such as the College Board's SAT, are at risk of being invalid.

This paper does not intend to break new ground on pushing back views of Platonism, nor do I intend to flesh out fully an account regarding the direction we ought to move to integrate ergodic theory into education. Rather, the purpose of the present paper is two-fold. First, I intend to historically trace back one (of many) theoretical threads of where educational assessment originated from in order to substantiate Michell's claim that some contemporary definitions of measurement originated in ancient ideas and have persisted through time, with Quetelet being one example. What I do not intend to claim is Quetelet's work is the same or the primary foundation for contemporary psychometrics or that this critique is applicable to psychological measurement of all kinds. If the first discussion is successful, it will contribute to the literature the idea that the problem of the definition of measurement in educational psychometrics is not primarily contemporary, but has persisted explicitly through time.

Second, I will discuss some of the measurement logic that primarily exists in the present educational testing industry and how, even though it is qualitatively different from Quetelet, it suffers from many long-standing issues that are continuously challenged by dynamical systems theory. This final thread intends to contribute a call for educational psychometrics to borrow from the dynamical systems paradigm. I also hope to establish the connection between ergodic theory, which typically is only discussed in physical systems, and educational assessment.

The paper can be outlined as follows. It begins with a 'user-friendly' introduction to assessment and how it relates to Adolphe Quetelet's work. I then introduce the normal distribution, how it was originally used in mathematics, and how Quetelet reinterpreted its use for the social sciences. I then briefly connect Quetelet's ideas to the mental testing fervor of the late 19th century that was eventually normalized in psychology. I briefly push

[3] It should be noted here that patterns, whether numerical (i.e., continuous or discrete) or qualitative and described by some arbitrary value, will always form some sort of a distribution or pattern. As such, seeing a pattern should not be considered the sine qua non of quantitative structure.

back on Quetelet's own argument for how the normal distribution ought to be interpreted using Field's [16] classic critique of Platonism before discussing how similar logic plays out in modern cognitive psychology today—though instead of taking place in Plato's forms (or something of the sort) it takes place in operational definitions. Finally, assuming that the first assumption is philosophically justified or completely side-stepped in assessment today, I show that much more work needs to be done to demonstrate that aggregated data can make claims about a specific person in relation to the group. Broadly, this paper is in an effort to encourage more future scholarship in the intersection of philosophy, neuroscience, and dynamic systems that will ultimately help improve educational assessment theory (e.g., addressing what sort of 'dynamical' system is human cognition and how it relates to personal identity).[4]

It should be noted that some of the claims offered in the paper, particularly those that are 'philosophically loaded', are not thoroughly fleshed out due to the limiting nature of a mere journal article that is covering such a great range of scholarship; as such, those who study a particular area mentioned in-depth might find some claims underwhelming or under-contextualized. However, one does not have to subscribe to many of the claims and interpretations offered to be convinced that dynamical systems and educational assessment ought to come into contact more often using philosophical analysis. It should also be noted that 'psychometrics' is not being used as the proper diverse discipline of psychological measurement; it would be both incorrect and irresponsible to claim that psychometricians, as an entire professional cohort, are mistaken in their work. The present critiques could not possibly cover the immense scale of approaches in the field that are carried out in diverse contexts and with diverse assumptions. 'Psychometrics' is being used to specifically pick out standardized educational assessments that define measurement to be the systematic assignment of numbers to psychological events without explicitly justifying what that event is and where its limits are. Practically speaking, most of these psychometric tools can be found in the for-profit testing industry, which might be small in number relative to existing psychometric literature, but hugely influential. In contrast to the psychometric positions I am discussing, see [18–20] for some examples of scholarship that would presumably escape the critique.

The clearest place to begin is in considering the meaning of 'assessment'. In the simplest sense, assessment is any "procedure for eliciting evidence that can assist in educational decision-making" ([21], p. 5). Taking this definition, it is then typically assumed that the assessment is distributed, empirical results are collected, those results are interpreted, and educational action is taken (e.g., distributing the mark, evaluating and rewarding teacher performance, providing extra resources to achieve student proficiency). Unsurprisingly, the aforementioned cadence of assessing students is far from being so clean. One reason for the messiness is that, even in the context of educational theory, assessment as a practice is essentially hermeneutical [21] and as such lent itself to a troubled past when discussed uncritically—both in terms of its role in socio-political movements and in terms of foundational methodologies as an 'evidence-based' practice [22].[5]

Much of the educational common sense that we have today surrounding what assessment mechanically does can be traced back at least until 1835 in Belgium when mathematician and astronomer Adolphe Quetelet published his most popular work titled *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale* [23].[6,7] In this text and

[4] A philosophical example of what I take to be a 'dynamical' system in human cognition is the very system Eric T. Olson claims is the qualitative "characterization question" in personal identity [17].

[5] See especially chapters 6 and 7 in [22].

[6] See also [24,25].

[7] Quetelet's "obsession [of the average man] led him to develop statistical profiles of the average physical, intellectual, and social features of various populations. . . Quetelet thought of his 'averages' as properties of particular people, groups, or races, that is, as things waiting to be discovered and not just abstract concepts. What he called his 'Average Man' was nothing more or less than the central tendency of a distribution of some human trait with deviations around the central tendency that looked like a normal bell-shaped curve" ([23], p. 66). In his lifetime, "Quetelet applied [his statistical] concepts to statistics on crime, mortality, weight, disease, intelligence, occupations, levels of education, religious affiliation, economics, and other phenomena" ([23], p. 66).

several of his following works through mid-century, Quetelet argues for the application of the newly, at the time, developed mathematical tool called the Gaussian curve to culture and society. The Gaussian curve is also called the normal distribution or bell curve. In fact, Quetelet is attributed as the source for two ideas that are taken as assessment common sense today: (1) there is such a thing as the empirical "average human being" (e.g., in terms of academic performance) and (2) the error distribution or normal curve can stand in to describe the behavior of dissimilar entities (e.g., the assumption that drives the fact that most state and private assessments are artificially designed to fall into the shape of a normal curve). Quetelet was the "father of the 'average man'", says statistician Simon Raper, in the sense that Quetelet "pioneered the application of statistical methods to social data", which "radically reinterpreted Gauss's error curve" to be applicable to group metrics like crime, health, intellect, and poverty ([26], p. 15). To appreciate how profound of a shift this was, it is important to understand what exactly the normal curve is, what it is designed to do in the world of measurement, and how it fits into science's quest for accuracy.

2. Adolphe Quetelet and the Normal Distribution

The most common knowledge people tend to have of the world of mathematics is that the discipline is an excellent tool for analyzing the characteristics of systems by simplifying conditions. We can never have perfect knowledge of a system, but numerical analysis, descriptive analysis, and dynamical modeling can help us get the gist of a system by making a few assumptions about the world. After all, accurate measurements are notoriously difficult to achieve.

When we measure the temperature of a room, there are many different facts to take into account; in the most basic sense, there are different temperatures in essentially every area of the room. Similarly, we weigh different amounts at different times during the day, when we wear certain clothes, and when we travel to certain points on earth. The same could be said for how tall we are, since our spine is compressed more under certain levels of gravity on earth and in the morning we are a bit taller. We could go on and on with these sorts of examples. In a colloquial sense, when we talk about the measurement of something, we typically mean the true value or the average value; typically, those two ideas are conflated without much loss of understanding. In other words, we care about what value something generally is most of the time. I am about six feet one inch, the room is around 22 °C, I am about 82 kg, and so forth.

When it comes to science though, precision and accuracy are critical. One of the first scientific disciplines to run into measurement error in a way that produced astronomically different results (both literally and figuratively) was astronomy. When measuring the orbit of a planet, for example, the time something took to get to one place or another could differ drastically depending on the limitations of tools or the difference in measurement technique. What early scientists intuitively struggled with was the following question: what do they do when they take different measurements of the same object under nearly identical conditions? In other words, if all measured values essentially have the same claim of objectivity for being the true value of a phenomenon (e.g., the true distance of an object from the earth), do you keep all of the values or do you combine them in some way?

Today we know that it is commonsensical to combine measurements in most cases[8] into some sort of central tendency—typically the arithmetic mean (henceforward referred to simply as the 'mean'). In the case of measurement, the mean is the summation of the given measured values divided by the total quantity of measurements. As mentioned, the first science to adopt the mean as a descriptive statistic was astronomy, merely because it was one of the few sciences from the 1600s–1700s that regularly endorsed taking multiple

[8] For the purposes of this paper, I intend to only focus on the arithmetic mean. There is an extraordinary number of methodologies for treating data. Even a high school statistics course teaches students that a distribution that is not roughly normal, via several tests to verify assumptions, should be treated with the median and the IQR instead of the arithmetic mean and standard deviation. Still, it is difficult to argue against the arithmetic mean being the most pervasive metric of the central tendency of a distribution with the standard deviation as the description of variation.

measurements and, thus, needed a way to combine them;[9] surprisingly, using a metric like the mean was largely controversial until the late 1770s, because it artificially erased the contextual factors present at the time of measuring an entity.[10] Still, with the mean finally achieving the wide appeal it has today with the turn of the 19th century, its use was for single object measurements. In other words, if I were to measure my weight repeatedly, 19th century scientists would be comfortable describing my weight with the mean because it, theoretically, should be roughly the same under common circumstances.

Right around this time, the mean also teamed up with Laplace's Central Limit Theorem (CLT), giving us the classic bell curve shape we are familiar with today.[11] The bell curve itself essentially tells us the following:

1. Smaller errors in measurement, such as being off by a few decimals when measuring weight, are more likely than larger errors, like being off by 30 kg.
2. The likelihood of some "overshooting" error ϵ is the same as some "undershooting error" $-\epsilon$ (i.e., the normal curve is symmetrical around some center point).
3. In measuring the same entity over many measures, the most probable value for the entity is the average of the measurements (i.e., the mean is the center point).

Still, neither the normal distribution nor the mean were intended to describe a group of dissimilar items. In Raper's words, "nowhere was there any inkling that an average might express a measurement across many dissimilar cases. That idea, with its implied wastefulness of simply throwing away all the information that pertained to individual cases, remained beyond consideration" ([26], p. 14). That is, until we turned to Quetelet's mid-19th century revolution. Raper recount's the turning point:

> The pivotal moment came while Quetelet was examining a set of data giving the chest sizes of 5738 Scottish soldiers. As he examined the distribution of the chest measurements, he did something unheard of for the time: he calculated the average chest size across all the unique and physically distinct Scottish soldiers ([26], p. 15).

Quetelet would then go on to notice that the data he was examining were "[i]n modern terms…normally distributed" ([28], p. 108). While the data he was looking at does, at least, look like a normally distributed model, it is interesting to note that, strictly speaking, the data are not normally distributed.[12]

While in one sense, the fact that he combined dissimilar objects into a single number was already a new concept, he also offered a strange interpretation for what that mean value meant. As educational professor Todd Rose notes in an article adopted from his book *The End of Average: How We Succeed in a World That Values Sameness*:

> After Quetelet calculated the average chest circumference of Scottish soldiers, he concluded that each individual soldier's chest size represented an instance of naturally occurring "error," whereas the average chest size represented the size of the 'true' soldier—a perfectly formed soldier free from any physical blemishes or disruptions, as nature intended a soldier to be ([30]).

It is in this study where Quetelet's work would go on to have a profound impact on the social sciences, by not only providing the suggestion that any measurements conceptually similar enough can be combined into the mean with a normal distribution, but also that there is such a thing as a 'perfect' average human represented by the mean.

[9] One of the opponents of estimating the 'true value' of a measurement, regardless of the technique of estimation, was Chemist Robert Boyle. Boyle boldly argued against combining measurements due to losing the value of individual context from experiments, saying "… experiments ought to be estimated by their value, not their number; "… a single experiment … may as well deserve an entire treatise… As one of those large and orient pearls … may outvalue a very great number of those little … pearls, that are to be bought by the ounce … " ([27], p. 158).

[10] See [28] for full discussion.

[11] See [29]) for discussion on CLT.

[12] See ([28], pp. 108-9) for analysis and discussion.

This is not an exaggeration. Quetelet was advocating for a position where the normal distribution and the errors that physical items had in deviating from the mean were directly tied to objective social value. In Quetelet's words:

> If the average man were completely determined, we might. . . consider him as the type of perfection; and everything differing from his proportions or condition, would constitute deformity and disease; everything found dissimilar, not only as regarded proportion or form, but as exceeding the observed limits, would constitute a monstrosity ([31], p. 99).

In other words, Quetelet was outlining the following assumptions: (1) somewhat related entities that have roughly similar properties can be measured and those measurements can be combined to find a perfect central tendency and that humans are sufficiently similar and stable enough both physically and mentally for this operation, (2) all deviations from the central tendency are not an error of measurement, but rather a flaw in that object relative to the perfect central tendency, and (3) this information could be used to make deterministic predictions regarding groups (i.e., his "social physics" as he discusses in his 1835 work [32]). Determinism is used here in the sense that these traits are essentially unalterable via actions the person might take in the world. It should be noted briefly, before being delt with fully in a moment, that Quetelet is basing his interpretation of the normal curve on a crude form of Platonism—a particular form that Aristotle would have potentially rejected (e.g., [33]). Platonism "is the view that there exist such things as abstract objects—where an abstract object is an object that does not exist in space or time and which is therefore entirely non-physical and non-mental" [34].

Quetelet's work was taking off just at the right time for Sir Francis Galton, the founder of Eugenics, to combine the assumptions of Quetelet's dream of a "social physics" and experimental psychologist Alfred Binet's first mental examination into a socio-political ideology and infrastructure that would produce nearly a half-century of political unrest.[13] While the history of Eugenics[14] and the foundations of mental testing are beyond the scope of this paper, Quetelet's influence is still distantly related to how we interpret educational assessment today because loose interpretations of work like Quetelet's were built into key assumptions in sociology and psychology that were unchallenged for decades [23].

It might at first be difficult to tie Quetelet's claims about physicality to the mental testing of educational assessment, where his view does not consider it a positive outcome to be above the mean academic performance. To be clear, Quetelet made the claim that any person who deviated from the mean was flawed in part because he was originally concerned with physicality[15]; as his career moved forward, he turned his attention to mental attributes ([28], p. 109; [23], p. 66). Quetelet's claims of the physical average being perfect makes more sense given that Quetelet was the creator of the body mass index: if someone's chest becomes too big, it is a good sign they are out of shape and if someone's chest becomes too small, it is a good sign they have a frail stature in terms of muscularity. Because of aforementioned socio-political ideologies, this claim that the numerical values on the error distribution were tied to social values was then reinterpreted when it came to education and mental testing. If the mean person's intelligence is the most perfect form of the human animal at some particular time and any deviation from the mean was an error, the right side of the normal distribution was seen as a positive, accidental error that unfolded over time.

As such, the 'positive' error surfaced in early 20th century Northern America, where some early strands of biological positivist philosophy[16] were beginning to mix up in social discourse with a form of social Darwinism (i.e., "natural selection" applied to humans,

[13] See [35] for a comprehensive account on the relationship between this statistical interpretation, educational examination, and Eugenics.

[14] See [36] for a comprehensive historical account of Eugenics.

[15] Recently, there has been extraordinary work done in the social sciences to demonstrate the difficulty of measuring non-physical attributes [37–39]. What is interesting is that Wiliam [21] argues against even using physical measurement of length as an "ideal" that education ought to hold itself to. In his words, "[i]t is time that educational assessment stopped trying to 'be a science', and found its own voice" ([21], p. 18).

[16] For description of biological positivism, see Chapter 3 in [40].

saying that only the physically and mentally 'strong' should reproduce and lead society to avoid extinction) that Quetelet helped inspire (e.g., [41]). During this time, intelligence was thought of as an intrinsic trait and it would make sense to the classical Eugenicist that a premium was placed on finding those individuals at the higher end of the distribution and that such individuals should be met with more resources to succeed and then such people would be encouraged to reproduce with similar people. It is this brief summary of trends that led to the explosive use of educational assessment. As Allen notes:

> [t]he examination evolved with cynical speed beyond its initial conception as imposed artifice, to become a reflection of naturally occurring, competitive forces. It fell within the ambit of a sociological naturalism, according to which in its artificial forms examination merely formalises what is already going on. Examination becomes unavoidable, inescapable ([35], p. 101).

Though some of the historical events related to educational assessment are troubling, we still ought to consider whether much of educational common sense today is merely reinterpretation of much of this sort of work.

3. Pathology in Contemporary Educational Assessment

When we say "all students should have the right to learn and succeed in school", we are advocating for the lurching of students below the mean to either be at the average or exceed the average over time. This is a harmless normative claim prima facie. We have long since shed the assumption of exclusively intrinsic, heritable intelligence and acknowledge that there might even be multiple views through which we think about intelligence. Still, we in the United States assess nationally and at the state-level over K–12 grades in order to identify weaknesses in student performance and then provide targeted educational, non-profit, and for-profit interventions to improve those test scores. This fits squarely within the Northern American educational enterprise which, fundamentally, tries to bring the whole distribution mean assessment score up, depending on the assessment being used, in order to improve the future of the country, which is typically tied with the idea that the success of a nation is contingent on the accrued skills and credentialism of its youth. The latest iteration of this foundation is the relatively new Every Student Succeeds Act (ESSA), which is currently the national educational law in the U.S that aims at achieving hyper-accountability[17] for measurable, improved results in state level metrics by race and gender. States have a wide range of freedom in choosing the actual metrics that they are accountable to, such as state-chosen assessments (which are almost always based on a fitted normal curve), improved high school graduation rates, or decreased disciplinary action taken against students.[18,19]

When we go through the cadence of education (i.e., teach, assess, report, and repeat), assessments have a particular purpose especially in the age of ESSA: from generalized group data, we make claims about individuals. When we want to know how well New Hampshire students over a decade are doing on a specific mathematics assignment, we collect data, create a normal distribution, and then interpret that distribution of group data to then intervene back down into the individual level. In doing so, we are essentially trying to track (1) where the high achievers are, (2) where the low achievers might need improvement, and (3) where the mean performance of the group rests in relation to years past to see if the program is really effective at improving performance. In the clearest form, tests like the ACT or College Board's redesigned SAT—which are used for college admissions (being either mandatory or optional) and sometimes for U.S. state-wide assessments (such as New Hampshire)—proceed with the following cadence: students

[17] For discussion on this paradigm of assessment, see [42] and Chapter 2 in [43].

[18] For policy analysis of ESSA and examples of targeted non-profit interventions, see [44].

[19] It has been argued elsewhere, practically speaking, the hyper-accountability that was supposed to make ESSA successful can easily be worked around to produce the intended measurable goals, without actually qualitatively improving educational outcomes; an example of this would be through test score pollution or teaching to the test which both artificially inflate scores or graduating students, especially students who have a unique learning style, before they're ready [44–46].

sit for a normalized exam, that data confronts a normal distribution, from that normal distribution scores are given to students, those students submit the scores to colleges, those colleges then compare the students' performances to the collective performance of those test takers to influence decision making.

Notice an assumption here: while the mean in a normal distribution does not describe any one student in particular, it is essentially a good estimate of the true mean of the group's real academic performance on the skills for which the examination was a proxy (i.e., whatever the assessment defined as the testing domain). Implicit is the idea that someone's 'academic performance' does not need to have an empirically tested quantitative structure prior to measurement, as long as the test is systematic and forms a distribution—which, as was shown, was a sentiment shared by Quetelet over a hundred-fifty years earlier. The assumption continues by implying that this group data can then be used to interpret the merits of someone's performance relative to the group mean. This claim seems commonsensical and indisputable and has been argued to be such. A controversial defense can be seen in *The Bell Curve*, where the mean and normal curve combination is described as "one of nature's more remarkable uniformities" ([47], p. 557). A much more tamed account, which is generally in-line with mainstream educational testing, can be found in *Measuring Up: What Educational Testing Really Tells Us*, where Koretz describes that it would be impractical to test someone's educational skill in a particular area over a lifetime, so the results of a given test are a relatively reasonable proxy, provided the proper context, to establish one's level of knowledge ([48], pp. 35–45). Regardless, the above assumption, which helps us interpret the meaning of an assessment's results, can directly be traced back to the work of Quetelet. A given measurement can meaningfully be compared to some non-real mean performance, because Quetelet helped establish that it is reasonable for the given test performance of one student to be compared to that of another distinct person since they both are instances of some shared deeper source structure—namely intelligence or ability. Does his logic hold up? Let us investigate it on two fronts: the philosophical merits and the dynamical systems theory merits.

In interpreting his quasi-normally distributed data, Quetelet is alluding to Plato's philosophy of the forms to ground his second aforementioned assumption (Section 2). Plato's forms claim there is essentially a 'hidden reality' where the one true form or idea of a given object resides and that the world of 'appearances' that you and I live in is constituted with merely imperfect, error filled copies. For example, consider a tree. In the view of a crude account of Platonism, the perfect "tree" exists beyond time and physicality; it is inaccessible to experience. The trees that we engage with are the failed, mangled copies of that perfect tree; the trees we engage with, in fact, are not really trees at all, but rather are derivations of the one true tree in the abstract realm. As we will discuss later, the modern day derivative of this Platonism in the context of education is that psychometricians who have students sitting for a standardized test are making an assumption that, given some group of students who were subject to the same standard-learning criteria, there is some real stable mean performance of that group that students can objectively or quasi-objectively be compared to in order to characterize their competence. Further, the test distributed is an valid proxy for the competence measurement. This stable mean performance is, of course, the assumption much of psychological measurement 'gets at' with constructs like 'intelligence', 'personality', and so forth—the mean is some sort of causal, stable trait being picked up by the test even though we do not really know what it might be and have not confirmed that its structure is quantitative as opposed to qualitative.[20]

Before dealing fully with the modern day, let us continue with exploring Quetelet. While the above view of there being some real, inaccessible, yet causal, intelligence trait seems quite metaphysically mysterious, Quetelet is in line with how many mathematicians viewed numbers and subsequent calculations, as well as how Stevens and Kelley

[20] It should be noted there is no such thing as a "valid test", but rather a valid interpretation of the findings of the test that line up with our theoretical constructs ([49], p. 119).

think about measurement. A narrower and more sophisticated version of Platonism is particularly common among logicians and mathematicians and is known as mathematical Platonism—though it seems to be dwindling in recent years. Mathematical Platonism argues that claims such as "there are infinitely many transfinite cardinal numbers" is literally a true statement about an abstract object that exists, with such a statement having equal reality with the claim that "there is a coffee cup to my right" [34]; as such, that means when we are saying "13 is a prime number", we are not making the claim about an idea—namely that 13 is a prime number in the same sense as my birthday is an idea dependent on human thought—nor a physical object, but rather a new class of immaterial object that exists irrespective of mental processes. Several prominent accounts in defense of mathematical Platonism can be attributed to Gödel ([50], pp. 515–525) and Putnam ([51], see chapters 3–5 and 8–9). John Burgess [52], in the spirit of Gottlob Frege [53], has a strong account that many find attractive, because he attempts to show that mathematical truths operate just as language is thought to function: via classical semantics. Classical semantics is the view that the semantic function of a given term refers to an object and a quantifier (e.g., $\exists x$, "there is such an x" and $\forall x$, "for all x") range over objects cleanly. It has been argued elsewhere that this view is at odds with our current conception of how we understand language [54]; however, even though there are respectable accounts in defense of these views, their proponents are often careful about the implications of their accounts.

In fact, "[m]any philosophers who defend Platonism in this purely metaphysical sense would reject the additional epistemological claims" where we can come to know anything about what the properties of abstract entities are in virtue of existing [55]. Rather, we say that whatever we do in mathematics can have bearing on reality and what is more important is our justification of the assumptions we make and how wrong they end up being. Still, I do not find mathematical Platonism particularly attractive. Consider the following argument similar to Field [16][21] :

[1]	(1)	Human beings exist entirely within spatio-temporal dimensions.	Scientific Premise
[2]	(2)	If abstract objects exist, then they exist entirely outside of space-time and are not contingent on cognition (i.e., beyond interaction with spatio-temporal objects).	Platonist Premise
[3]	(3)	To acquire knowledge of something, human beings become aware of those properties via interactive experience of some sort.	Scientific Premise
[1,2]	(4)	If abstract objects exist entirely outside of space-time, human beings cannot interact with abstract objects.[22]	Truth Functionally follows from (1,2)
[1,2,3]	(5)	Either we know nothing about mathematics or numbers are not abstract objects outside space-time.	Truth Functionally Follows from (3,4)

Considering that mathematics, while beautiful and interesting in its own right, is tied to the sciences, the mathematical Platonist cannot reject premise one or three.[23] Most

[21] I concede that the discussion around Platonism is far from a consensus, though, while I am not particularly sure I am committed to nominalism, I cannot seem to justify any recent accounts of Platonism, nor any ancient ones. For some negative accounts on Platonism see [56] and more recently [57]. For a direct negative account to [56], see [58].

[22] A true Platonist would try to reject the truth-functional inference 4 by rejecting premise 1 or 2, because it rejects the portion of the human soul that makes the transition from the world of physical realities to the world of forms. Access to the world of forms was thought to be achievable through dialectics and dieresis. While this objection is admissible, it is not entirely clear that Plato would endorse the view as it relates to Quetelet or mathematical Platonism, even though Plato's teachings were inexorably linked to mathematics [34]. See [59] for a helpful look at Plato's full works. If the objection was raised, while I am remaining neutral on the existence of souls in the current paper, the soul-view of personal identity remains a particularly demanding view of reality and is not something I would endorse [60,61].

[23] It should be noted that Gödel rejected premise 1, because he believed we acquired mathematical knowledge via intuitionism, which is a property that humans supposedly have and is a property that is immaterial. Intuitionism has fallen out of favor in the philosophy of mind for any number of reasons, but none more than the fact that intuitions are not necessarily reflective of reality. However, this does not outright disqualify the view. Suppose I have the intuition that the next time I sit down in my chair, it will not collapse under my weight. This intuition is from the fact that I have sat down in it thousands of times and it did not collapse. Thus, I have the intuition that it ought to not break, but, as it turns out, the chair does in fact break. Naturally, this informal thought experiment shows that intuitionism is logically similar, if not identical, to scientific induction. David Hume famously argued that there is no logical connection between what is true now and what ought to be the case in the future (i.e., is/ought distinction); there has yet to be a sufficient reply to this argument in a few hundred years.

logicians and mathematicians who endorse Platonism do so in order to demonstrate that science has a privileged access to universal knowledge, because mathematics would be built directly into the universe as opposed to being created by humans. Similarly, mathematicians and logicians would not want to reject the claim that we have mathematical knowledge or otherwise their work would be meaningless. Most contemporary mathematicians reject Platonism as something to be defended, but instead endorse "working realism", which is the methodological view that says mathematics should proceed as if numbers, calculations, and theorems were real abstract objects (i.e., have real implications in our lives), but does not actually require a metaphysical defense; in other words, they are "not real" in the sense that mathematical models are "not real" ([62], pp. 38–44).

Even if the reader is sympathetic to some form of mathematical Platonism that avoids the above epistemological argument, Quetelet and his interpretations of normal distributions regarding human assessment patently cannot escape, because he claims that the average person in some measurement is both an abstract object (i.e., perfect form) and that we have knowledge of them within our space-time from the normal distribution. In his words, he asks us to consider a perfect body of a Gladiator relative to measurements of the chest of a random person; he then asks us to consider statues that copied that gladiator and he makes the claim that our intuition about the error of statues is the same as how we should measure people:

> Let's change our hypothesis again, and assume that we used a thousand statues to copy the gladiator with all the care imaginable. [You] certainly would not think that the thousand copies were exactly like the model, and that the successive measurements, the thousand that I obtained, will also be concordant as if I had taken measure of the gladiator itself. . . but if [the copier's] inaccuracies are accidental, the thousand measures, grouped in order of severity ["grandeur"], will still present with remarkable regularity and will succeed each other in the order assigned to them by the law of possibility. . . Yes, truly, we measured more than one a thousand copies of a statue that I am to assure you were not to be that of the Gladiator, but which, in any case, is not far from it ([63], pp. 135–136; my translation).

In this sense, Quetelet is subject to the deduction above and his argument becomes logically untenable because he is assuming that, by taking the average, we are getting at the perfect, abstract object outside of space-time by interacting exclusively with the spatio-temporal world (i.e., negating the second disjunct from the conclusion) while also claiming the mathematical knowledge of Laplace in 1810 regarding the properties of error curves (i.e., negating the first disjunct from the conclusion).

It is at this point where we have achieved the first purpose of the paper. From thinkers like Stevens, Koretz, and Kelley and how we think about education today, this thread was traced back to Quetelet to demonstrate that skipping the justification of quantitative structure (rather than assuming it is built into the universe) can lead to trouble. It was also shown that the question of whether or not dissimilar attributes should be averaged and interpreted—at least the rigorous formulation of the question—is not strictly contemporary and can be traced back to the early 1800s. As such, scholarship in this area must extend beyond the 20th century. We have also seen that, for what its worth, Quetelet's own project, though it was left behind for other reasons, fails on its own merits.

The issue at hand, however, is not merely a claim that Platonism disqualifies Quetelet's work—though it does seem that he would have some logical issues to address. In many ways we have moved away from the exact argument of Quetelet because his idea of intelligence measurement was unsuccessful even then ([5], footnote 9); it is here where one might ask 'what is the point of even looking at Quetelet at all?'. Averaging the way we do in cognitive psychology is descendant from Quetelet's work in the sense that we are getting at some stable and real abstract trait, albeit imperfectly, in the same way Quetelet thought we measured people and the stars: there is some true value from a quantitative structure that is inaccessible directly via measurement, but if we collect enough data by systematically assigning numerals to a system by rule, the true value leaves imprints on the measured value plus some error. What is the justification for such a claim?

As Speelman and McGann [38] note, cognitive psychology believes that the tools it uses to capture a psychological construct (which is already a proxy of some biological property) is the true value plus error. This is certainly so in psychological test theory, since it has an axiom that a test is sensitive to some trait with some noise, just as we have when we measure physical phenomenon like temperature in the room or something of the sort [64]. Recalling from earlier, one way to get rid of this issue is to either artificially fit scores from a psychological assessment to a normal distribution or to collect enough responses to create a stable scoring output using the law of large numbers. Given that there is some mean, the likelihood of an overshooting and undershooting error under standardized conditions is symmetrical (i.e., equally likely) and therefore these errors cancel out and just leave the true value of the exam falling into a distribution of test takers. We often soften this interpretation by saying there is still noise in the exam where students or study participants are producing the true value of the trait as it is picked up by the particular instrument. Therefore, a poor score on a national assessment is not that the student is an academic failure, but rather they just did not perform well in the area that the test was sensitive toward.

We tend to have no issue with this assumption, but upon closer reflection, there is a serious challenge to this point of view regularly, and in some subfields entirely, ignored. Consider just one individual taking the United States organization College Board's SAT. Students are given their actual score and a confidence interval that tells them how they would have performed if they had taken the exam over and over again under the same conditions without additional learning. Colleges in the United States often encourage a student to take the exam as many times as they need to get a score that represents their academic achievement because taking the exam more times will supposedly remove the 'noise' surrounding the circumstances of the exam (e.g., being stress free, or not getting enough sleep). This should be fine because if we took the exam many times our performance would eventually settle around a central tendency.

But why would we not take the opposite approach, in that the test performance is a snapshot of the cognitive system and is the specific trait as it is at that moment, rather than saying many error filled trials produces a mostly error-free trait or level of achievement? On this alternative view, there is no stable quantitative trait, but the output of the tool is just a snapshot of emergent behavior from a model's rendering (i.e., the model being the exam construct) of a complex, dynamical system. Speelman and McGann [38] lobbed this same question at certain psychological constructs around taking many trials of response times and then averaging them to get some mean response time of supposedly a stable trait that is someone's reflex speed in context. This is much more serious in educational assessment because these tools are used, as mentioned, for educational decision-making and often lead to interpretations that deal with moral and practical claims about the social worth of an individual in the professional world or general job market—as such, we ought to care just how good of a proxy these tests are for the object of study in question.

If, after several test iterations, you average out some individual's performance to be at some quartile in a distribution, what exactly is that biologically picking out relative to the distribution? If the human brain is malleable (i.e., via plasticity), how would we be able to tell what the student's mean score is actually picking out, if anything at all, and should the trait being picked up not be an instantaneous snapshot of cognitive pathways interacting as opposed to a stable trait? If the reason we even repeat tests or worry about having a sufficient sample size is because we believe the assessment tools are flawed at giving us a perfect measurement, the idea that "taking the average of measures from repeated trials will provide a reflection of some stable element of the cognitive system seems fanciful given that the system could not be stable if we keep giving it experiences" [38].

This, of course, is to ignore that there is no reason to believe that taking a snapshot of a cognitive system should be quantitative at all (i.e., as opposed to just an unstable qualitative trait), because it has never been justified [6] and was propagated by Quetelet and his contemporaries before being normalized into educational theory. If the trait is qualitative, should we not be more concerned with what that test says about the cultural infrastructure surrounding that student and not be worried about that student's capability relative to her society? If the brain is something that is affected by education—such as the fact it grows, changes, and so forth—then, in fact, the very act of measuring the supposedly stable trait would influence the properties of that trait. This is because the student is reinterpreting information on the test within their cognitive models just the same as they would doing a homework assignment or reading a book. Their interpretation of the information is culturally mediated and, unless the exams pose no intellectual challenge whatsoever, we should not be so infatuated with the information these exams produce. As such, while I do not raise these questions to 'take down' educational assessment as a general practice, the interpretations—which typically stem from political discourse and pop-psychology—of what exactly these assessments say about a general population are highly suspect to say the least (e.g., in terms of a population being 'smarter', this school is 'better', some country is 'uneducated') and it is doubtful whether many psychometricians, when this challenge is pointed out, would endorse these claims of cross-context comparability. While this paper must leave the raised questions open, a recent full review [65] has unpacked a number of these issues with rigorous care.

4. Ergodicity and Individual Assessment Performance

Let us consider that eventually we are able to justify that the mean is the true value of an individual's ability plus some error for a given trait and that such a result comes from an intrinsically quantitative structure comparable across students. Would we be comfortable with the above methodology of taking snapshots of these performances? As mentioned, when we try to use social science research to generate aggregated data and then make inferences from statistical tests conducted on that data, we are making the assumption that meta-group trends can be reciprocally applied to understand the individual properties of the members of that same measured group. The problem is that "statistical findings at the interindividual (group) level only generalize the intraindividual (person) level if the processes in question are ergodic" [66,67]. An ergodic process is, informally for our purposes, any instance in data analysis concerning dynamical systems where we can generalize the results from group data to individual data ([68], p. 106). A bit more specifically, a system is ergodic if a dynamic system's time average and the expected observational value are the same. A time average is essentially the mean of some function over T iterations from some starting point within the measure space.[24] In psychology (and education for that matter), the time average would be the exam mean for one iteration over time and the observational value would be the deviation across student performances at a given time step—for instance, variance—from the mean. As such, in these humanistic fields, ergodicity can only occur when the mean and variance of groups and individuals remain consistent over time naturally without leaving the measure space. As Fisher et al. point out, "[b]ecause psychological and biological phenomena are organized within persons over time, generalizations that rely on group estimates are non-ergodic if there are individual exceptions" ([66], p. 106). If ergodicity is not present then dynamical descriptions (like human cognition) cannot be replaced with simple probabilistic tools like the normal distribution; this is an emerging problem in other fields as well, such as economics [70].

To show this idea a bit in practice, Aaron Fisher and his team of neurologists and psychologists ran symmetric comparisons of interindividual and intraindividual variation across six studies with repeated-measure design in both social and medical research. They found that, while the mean was relatively similar across their analysis constructs, the "variance around the expected value was two to four times larger within individuals than within groups. This suggests that literature in social and medical sciences may overestimate

[24] For more comprehensive description, see [69].

the accuracy of aggregated statistical estimates" ([66], p. 106). While this is one reality in psychological and medical research at large, it is much more problematic in educational assessment because of the way these assessments are defined, typically by forcing results into a normally distributed range.

Theoretically, if we retained the early 20th century idea that intelligence is an intrinsic and permanent feature of human beings that is not changed via education, we would escape this charge. While arguing against this view is beyond the scope of the current paper, we will continue with the assumption that intelligence is, in fact, not a stable, intrinsic human feature and that people are able to improve given their access to educational opportunities and support.[25]

In briefly returning to Quetelet's impact on modern day assessments, "in positing the 'average man' Quetelet was explicitly pinning the properties of a group onto an individual, even if that individual was ideal rather than actual. . . Quetelet has the same picture as his [modern] predecessors of what measurement fundamentally is, only he is carrying out his measurement on an ideal object" ([26], p. 16). Educational psychometricians today claim that they are, in the best case, measuring a snapshot of the student's academic capability at a given instant in time—not perfectly, but within reason. The problem with this is that education is fundamentally tasked with changing the behavior of the students, creating exception cases and often having different measure spaces. We then use that data to make educational decisions about the students in question. In other words, temporal and reflexive dynamics matter in educational assessment and cannot be assumed away in order to use clean probability distributions, even with non-psychological data as trivial as attendance records [72]. 'Snapshot assessment' is, therefore, fundamentally inconsistent via ergodic theory, because educational assessment in most cases is used either to (1) again, use the assessment to influence policy to affect individual students (the system) or (2) position an individual in relation to the mean (even though students are often dynamic exception cases).

There is nothing wrong with a given test per se, but the tendency for both trained professionals and non-academics to make critical mistakes in interpreting what educational statistics mean and, even if interpreted well, are used in improper ways to overstate conclusions of value, competence, or efficiency—especially about the individual from aggregate data. Raper reflects that "perhaps the original mistake [Quetelet made] is built right into us, as what Richard Dawkins called, in an influential article, 'The Tyranny of the Discontinuous Mind'" where people have an overt preference to discrete categories over continuous categories: above or below average, male or female, gay or straight, embryos are alive or dead, you're a Democrat or Republican, and so forth [73].

The irony here, of course, is that normal distributions are only for continuous random variables by definition, so it is unsurprising that scholars of assessment have known for a long time that psychological phenomena do not inherently distribute normally since assessments typically are discrete (i.e., you usually can get a 90% or a 91%, but not a 90.123213. . . %) [74–79]. As Dudley-Marling puts it:

> [h]uman behaviors are always socially and culturally mediated and, therefore, never occur randomly, a conclusion supported by an overwhelming body of theory and research. Yet the myth of the normal curve as a model of human behavior continues to exert a powerful influence on theory and practice in education and the social sciences, an instance of scientific groupthink that misrepresents the human experience ([74], p. 204).

[25] For example, see [71].

Dudley-Marling continues in saying “[t]he fetishization of the mean [and our efforts to move individuals beyond it] has the effect of masking the range of human differences that are always present in any population of students, perverting educational decision-making in the process” ([74], p. 208).[26]

5. Concluding Thoughts and the Future of Educational Psychometrics

While some troubling assumptions were tackled, there are many reasons to be hopeful about psychometrics broadly in the future. This is important to note in this heavily critical paper as psychometrics receives a disproportionate amount of criticism because of its relation to past social movements. In recent years, many researchers have moved away from the over-emphasis of the individual in relation to the aggregated data and have embraced the dynamical paradigm. For example, Zoannetti and Pearce [82] demonstrated that Bayesian networks are “theoretically well-supported” for committee-based programmatic assessment, because they “can ensure precedents are maintained and consistency occurs over time”.[27] Programmatic assessment is “an approach in which routine information about the learner’s competence and progress is continually collected, analyzed and, where needed, complemented with purposively collected additional assessment information, with the intent to both maximally inform the learner and their mentor and allow for high-stakes decisions at the end of a training phase” ([84], p. 211). As such, it is different in that there are no “high-stakes” test-taking or decision-making, but rather individual assessments provide feedback that then allows the learner to analyze "their own performance, formulate concrete learning goals and demonstrably attain them", which is opposed to the previously discussed assessment structure where the assessment *is* the evidence of learning or the lack thereof ([84], p. 211). Further reading in this area should be directed to [85–88].

Several key takeaways can be emphasized from the present analysis. First, dispite the more recent scholarship, educational assessment ought to seriously consider moving mathematically, psychologically, and philosophically into more dynamic areas of inquiry at a more broad level. As we have seen, some of the troubling assumptions in one form or another have been around in their rigorous statistical form at least since the early 1800s and can be traced in its loose form all the way back to Platonism and, as some claim, even before that period [89]. As such, psychology in general ought to be more careful about justifying the quantitative structure of the constructs it measures and, if they are not quantitative, invest more energy in developing and normalizing qualitative research methodologies. It was also shown that even if we sidestep the assumptions about measurement and aggregation of data, that does still not avoid the issue of ergodicity which is an under-discussed and under-appreciated mathematical concept in both psychometrics and education. This paper contributes one of the first explicit calls for the connection between education and ergodic theory and hopes to continue the recent trend toward the dynamical system paradigm being used in psychometrics.

There remain a few other future directions from the present work. Philosophically, accounts of personal identity that pitch the person as a dynamic narrative being could be potentially fruitful—especially in answering the ‘characterization question’ which typically asks “what sorts of beings are we in the world” [90,91], which can help formulate what characteristics are interesting to include in Bayesian models. The analytic tradition has failed to adequately interact with these accounts, which has resulted in work that focuses too much on what it means to be a numerically identical person over time (i.e., focusing on hyper-individualistic notions), rather than noticing that the two different

[26] It was discussed by [80] that ergodicity is sufficient by not necessary for group-to-individual generalizability. While this is true, the conditions where mental data of individuals is explained by group data are rarely demonstrated in psychological or educational literature. As the researchers’ reply to [80] explains: “[p]ractically, we are concerned that group models and research designs are often easier to power, perform, and incentivize (e.g., fund). Given the fact that mental, physiological, and behavioral processes manifest within people over time, it is prudent to assume that group models do not explain individual-level processes until it has been demonstrated. The burden of proof should thus fall on the group model to describe individuals. Each individual system may be quantitatively or qualitatively unique” [81].

[27] For full discussion on Bayesian networks, please see [83].

strands of thought are compatible.[28] This opens up a clear space for hermeneutical study to enter education and, more specifically, educational assessment paired with Bayesian statistics. In regard to psychology, it ought to be more critical of the assumptions that it has taken with it from the late 19th and early 20th centuries and instead turn closer to techniques in dynamical neuroscience to ask more questions about instantaneous brain states rather than trying to locate traits we cannot even be sure are there. In focusing so closely on the quantitative, scientific aspects of psychology, often the discipline can be distracted from issues of its paradigmatic foundations. Finally, education itself needs a better relationship with computational modeling and mathematics in general, such as the aforementioned Bayesian models playing out in medical school admissions or recent advances in probabilistic modeling; while I conceded that education is not an 'evidence-based' practice in the sense it produces objective and universal knowledge, nor should it be [21], a firm understanding of mathematics puts certain methodological assumptions into clear view—allowing for a more critical reading of otherwise commonsensical techniques.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable for the present study.

Informed Consent Statement: Not applicable for the present study.

Data Availability Statement: Not applicable for the present study.

Acknowledgments: I would like to thank Fernando Nascimento, for his continued mentorship and friendship. I would also like to offer thanks to Khushi Patel for her feedback on an earlier version of this paper. Thank you to my loving family for their continual support and for putting up with me through the COVID-19 pandemic. Finally, a special thanks to the anonymous peer reviewers for their helpful feedback and to Bowdoin College for affording me the opportunity to conduct my scholarship.

Conflicts of Interest: The author declares no conflict of interest.

References

1. McCulloch, G. 'Disciplines Contributing to Education?' Educational Studies and the Disciplines. *Br. J. Educ. Stud.* **2002**, *50*, 100–119. [CrossRef]
2. Themelis, S. *Critical Reflections on the Language of Neoliberalism in Education: Dangerous Words and Discourses of Possibility*; Routledge: London, UK, 2020.
3. Ricoeur, P. *Lectures on Ideology and Utopia*; Columbia University Press: New York, NY, USA, 1986.
4. Jahoda, G. Quetelet and the Emergence of the Behavioral Sciences. *Springerplus* **2015**, *4*, 473. [CrossRef] [PubMed]
5. Mosselmans, B. Adolphe Quetelet, the Average Man and the Development of Economic Methodology. *Eur. J. Hist. Econ. Thought* **2005**, *12*, 565–582. [CrossRef]
6. Michell, J. Is Psychometrics Pathological Science? *Measurement* **2008**, *6*, 7–24. [CrossRef]
7. Michell, J. Reply to Kline, Laming, Lovie, Luce and Morgan. *Br. J. Psychol.* **1997**, *88*, 401–406. [CrossRef]
8. Stevens, S.S. On the Theory of Scales of Measurement. *Science* **1946**, *103*, 677–680. [CrossRef]
9. Stevens, S.S. Measurement, Statistics, and the Schemapiric View. *Science* **1968**, *161*, 849–856. [CrossRef]
10. Kelley, T.L. *Scientific Method: its Function in Research and in Education*; Ohio State University Press: Columbus, OH, USA, 1929.
11. Michell, J. Quantitative Science and the Definition of Measurement in Psychology. *Br. J. Psychol.* **1997**, *88*, 355–383. [CrossRef]
12. Huffman, C. Pythagoreanism. In *The Stanford Encyclopedia of Philosophy*, 2019th ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2019.
13. Smith, L.B. Cognition as a Dynamic System: Principles from Embodiment. *Develop. Rev.* **2005**, *25*, 278–298. [CrossRef]
14. Barrett, N.F. A Dynamic Systems View of Habits. *Front. Hum. Neurosci.* **2014**, *8*, 682. [CrossRef]
15. Shine, J.M.; Breakspear, M.; Bell, P.T.; Ehgoetz Martens, K.A.; Shine, R.; Koyejo, O.; Sporns, O.; Poldrack, R.A. Human Cognition Involves the Dynamic Integration of Neural Activity and Neuromodulatory Systems. *Nat. Neurosci.* **2019**, *22*, 289–296. [CrossRef]
16. Field, H.H. *Realism, Mathematics & Modality*; Blackwell: Oxford, UK, 1989.
17. Olson, E.T.; Witt, K. Narrative and Persistence. *Can. J. Philos.* **2019**, *49*, 419–434. [CrossRef]
18. Schoenherr, J.R.; Hamstra, S.J. Psychometrics and its Discontents: An Historical Perspective on the Discourse of the Measurement Tradition. *Adv. Health Sci. Educ. Theory Pract.* **2016**, *21*, 719–729. [CrossRef]

[28] For example, see [17] to discuss the plausibility of the characterization question.

19. Wilson, M. Seeking a Balance between the Statistical and Scientific Elements in Psychometrics. *Psychometrika* **2013**, *78*, 211–236. [CrossRef]
20. Pearce, J. In Defence of Constructivist, Utility-Driven Psychometrics for the 'Post-Psychometric Era'. *Med. Educ.* **2020**, *54*, 99–102. [CrossRef]
21. Wiliam, D. Towards a Philosophy for Educational Assessment. In Proceedings of the British Educational Research Association's 20th Annual Conference, Oxford, UK, 1994; pp. 1–23.
22. Kelly, A.V. *The Curriculum: Theory and Practice*, 6th ed.; SAGE Publications Ltd: New York, NY, USA, 2009.
23. Wright, J.D. The Founding Fathers of Sociology: Francis Galton, Adolphe Quetelet, and Charles Booth: Or What Do People You Probably Never Heard of Have to Do with the Foundations of Sociology? *J. Appl. Soc. Sci.* **2009**, *3*, 63–72. [CrossRef]
24. Johnson, N.L.; Kotz, S. (Eds.) *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present (Wiley Series in Probability and Statistics)*, 1st ed.; Wiley: Hoboken, NJ, USA, 1997.
25. Eknoyan, G. Adolphe Quetelet (1796–1874)–the Average Man and Indices of Obesity. *Nephrol. Dial. Transpl.* **2008**, *23*, 47–51. [CrossRef]
26. Raper, S. The Shock of the Mean. *Significance* **2017**, *14*, 12–17. [CrossRef]
27. Boyle, R. *Certain Physiological Essays and Other Tracts Written at Distant Times, and on Several Occasions by the Honourable Robert Boyle ; Wherein Some of the Tracts Are Enlarged by Experiments and the Work Is Increased by the Addition of a Discourse about the Absolute Rest in Bodies*; Henry Herringman: London, UK, 1669.
28. Stahl, S. The Evolution of the Normal Distribution. *Math. Mag.* **2006**, *79*, 96–113. [CrossRef]
29. Kwak, S.G.; Kim, J.H. Central Limit Theorem: The Cornerstone of Modern Statistics. *Korean J. Anesthesiol.* **2017**, *70*, 144–156. [CrossRef]
30. Rose, T. How the Idea of a 'Normal' Person Got Invented. *Atlantic* **2016**, *22*, 2017.
31. Quetelet, A. *A Treatise on Man and the Development of His Faculties, Trans*; Burt Franklin Philosophy Monograph Series #15; Burt Franklin: New York, NY, USA, 1842; Volume 247.
32. Quetelet, A.J. *Sur l'Homme et le Développement de ses Facultés, ou Essai de Physique Sociale*; Hachette Livre (Bibliothèque nationale de France): Paris, France, 1835.
33. Hasper, P.S. Aristotle's Argument From Universal Mathematics Against The Existence Of Platonic Forms. *Manuscrito* **2019**, *42*, 544–581. [CrossRef]
34. Balaguer, M. Platonism in Metaphysics. In *The Stanford Encyclopedia of Philosophy*, 2016th ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2016.
35. Allen, A. *Benign Violence: Education in and Beyond the Age of Reason*; Palgrave Macmillan: London, UK, 2014.
36. Kevles, D.J. *In the Name of Eugenics: Genetics and the Uses of Human Heredity*, reprint ed.; Harvard University Press: Harvard, UK, 1998.
37. Salzberger, T. Attempting Measurement of Psychological Attributes. *Front. Psychol.* **2013**, *4*, 75. [CrossRef]
38. Speelman, C.P.; McGann, M. How Mean is the Mean? *Front. Psychol.* **2013**, *4*, 451. [CrossRef]
39. Schmittmann, V.D.; Cramer, A.O.J.; Waldorp, L.J.; Epskamp, S.; Kievit, R.A.; Borsboom, D. Deconstructing the Construct: A Network Perspective on Psychological Phenomena. *New Ideas Psychol.* **2013**, *31*, 43–53. [CrossRef]
40. Heidt, J.; Wheeldon, J.P. *Introducing Criminological Thinking: Maps, Theories, and Understanding*, 1st ed.; SAGE Publications, Inc.: New York, NY, USA, 2014.
41. Galton, F. *Inquiries Into Human Faculty and Its Development*; Macmillan: London, UK, 1883.
42. Falabella, A. The Seduction of Hyper-Surveillance: Standards, Testing, and Accountability. *Educ. Adm. Q.* **2020**, 57, 113–142. [CrossRef]
43. Falchikov, N. *Improving Assessment through Student Involvement: Practical Solutions for Aiding Learning in Higher and Further Education*; Taylor & Francis: Abingdon, UK, 2005.
44. Bloniasz, P.F. *Case Study: Every Student Succeeds Act (ESSA) and National Service—A Review of Educational Support Needs and Curricula Development*; America's Service Commissions: Washington, DC, USA, 2019.
45. Haladyna, T.M.; Nolen, S.B.; Haas, N.S. Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution. *Educ. Res.* **1991**, *20*, 2–7. [CrossRef]
46. Ahlquist, R.; Gorski, P.; Montaño, T. (Eds.) *Assault on Kids: How Hyper-Accountability, Corporatization, Deficit Ideologies, and Ruby Payne Are Destroying Our Schools (Counterpoints)*, 1st ed.; Peter Lang Inc., International Academic Publishers: New York, NY, USA, 2011.
47. Herrnstein, R.J.; Murray, C. *The Bell Curve: Intelligence and Class Structure in American Life*; Free Press: New York, NY, USA, 1994; Volume 1.
48. Koretz, D. *Measuring Up: What Educational Testing Really Tells Us*, illustrated ed.; Harvard University Press: Cambridge, MA, USA, 2009.
49. Sullivan, G.M. A Primer on the Validity of Assessment Instruments. *J. Grad. Med. Educ.* **2011**, *3*, 119–120. [CrossRef]
50. Gödel, K. What is Cantor's Continuum Problem. *J. Symb. Log.* **1947**, *55*, 515–525.
51. Putnam, H. *Philosophy of Logic*; Allen & Unwin: London, UK, 1971.
52. Burgess, J.P. Why I Am Not a Nominalist. *Notre Dame J. Form. Log.* **1983**, *24*, 93–105. [CrossRef]

53. Frege, G. *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number*, 2nd ed.; Northwestern University Press: Evanston, IL, USA, 1884.
54. Bloniasz, P.F. On Cognition and the Tension of Live Metaphors. *Meta: Research in Hermeneutics, Phenomenology, and Practical Philosophy* **2020**, *7*, 499–516.
55. Linnebo, Ø. Platonism in the Philosophy of Mathematics. In *The Stanford Encyclopedia of Philosophy*, 2018th ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2018.
56. Liggins, D. Is There a Good Epistemological Argument Against Platonism? *Analysis* **2006**, *66*, 135–141. [CrossRef]
57. Leng, M. Does 2 + 3 = 5? In Defence of a Near Absurdity. *Math. Intell.* **2018**, *40*, 14–17. [CrossRef]
58. Kasa, I. On Field's Epistemological Argument Against Platonism. *Studia Log.* **2010**, *96*, 141–146. [CrossRef]
59. Kraut, R. Plato. In *The Stanford Encyclopedia of Philosophy*, 2017th ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2017.
60. Berger, J. A Dilemma for the Soul Theory of Personal Identity. *Int. J. Philos. Relig.* **2018**, *83*, 41–55. [CrossRef]
61. Patrick, B. Concerning Theories of Personal Identity. Ph.D. Thesis, University of South Florida, Tampa, FL, USA, 2004.
62. Shapiro, S. *Philosophy of Mathematics: Structure and Ontology*; Oxford University Press: New York, NY, 1997.
63. Quetelet, A. *Lettres à S.A.R. le duc Régnant de Saxe-Coburg et Gotha, sur la Théorie des Probabilités, Appliquée aux Sciences Morales et Politiques*; M. Hayez: Bruxelles, Belgium, 1846.
64. Novick, M.R. The Axioms and Principal Results of Classical Test Theory. *J. Math. Psychol.* **1966**, *3*, 1–18. [CrossRef]
65. Michell, J. Representational Measurement Theory: Is Its Number Up? *Theory Psychol.* **2021**, *31*, 3–23. [CrossRef]
66. Fisher, A.J.; Medaglia, J.D.; Jeronimus, B.F. Lack of Group-to-Individual Generalizability is a Threat to Human Subjects Research. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E6106–E6115. [CrossRef]
67. Molenaar, P.C.M. A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement* **2004**, *2*, 201–218. [CrossRef]
68. Gray, R.M. *Probability, Random Processes, and Ergodic Properties*; Springer: Boston, MA, USA, 2009.
69. Moore, C.C. Ergodic Theorem, Ergodic Theory, and Statistical Mechanics. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 1907–1911. [CrossRef]
70. Peters, O. The Ergodicity Problem in Economics. *Nat. Phys.* **2019**, *15*, 1216–1221. [CrossRef]
71. Ng, B. The Neuroscience of Growth Mindset and Intrinsic Motivation. *Brain Sci.* **2018**, *8*, 20. [CrossRef]
72. Koopmans, M. When Time Makes a Difference: Addressing Ergodicity and Complexity in Education. *Complicity* **2015**, *12*, 5–25. [CrossRef]
73. Raper, S. An average understanding. *Significance* **2017**, 14, 13–16. [CrossRef]
74. Dudley-Marling, C. The Tyranny of the Normal Curve: How the "Bell Curve" Corrupts Educational Research and Practice. In *Groupthink in Science: Greed, Pathological Altruism, Ideology, Competition, and Culture*; Allen, D.M.; Howell, J.W., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 201–210.
75. Fashing, J.; Goertzel, T. The Myth of the Normal Curve a Theoretical Critique and Examination of its Role in Teaching and Research. *Humanit. Soc.* **1981**, *5*, 14–31. [CrossRef]
76. Bradley, J.V. *Distribution-free Statistical Tests*; Cliffs, N.J., Ed.; Prentice-Hall: Englewood, NJ, USA, 1968.
77. Walberg, H.J.; Strykowski, B.F.; Rovai, E.; Hung, S.S. Exceptional Performance. *Rev. Educ. Res.* **1984**, *54*, 87–112. [CrossRef]
78. Cronbach, L.J. *Essentials of Psychological Testing*, 1st ed.; Harper: New York, NY, USA, 1949.
79. Wechsler, D. *The Range of Human Capacities*; Williams & Wilkins Co: Baltimore, MD, USA, 1935; Volume 159.
80. Adolf, J.K.; Fried, E.I. Ergodicity is Sufficient but Not Necessary for Group-to-Individual Generalizability. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 6540–6541. [CrossRef]
81. Medaglia, J.D.; Jeronimus, B.F.; Fisher, A.J. Reply to Adolf and Fried: Conditional Equivalence and Imperatives for Person-level Science. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 6542–6543. [CrossRef]
82. Zoanetti, N.; Pearce, J. The Potential Use of Bayesian Networks to Support Committee Decisions in Programmatic Assessment. *Med. Educ.* **2020**. Available online: https://doi.org/10.1111/medu.14407 (accessed on 2 June 2021).
83. Niedermayer, D. An Introduction to Bayesian Networks and Their Contemporary Applications. In *Innovations in Bayesian Networks: Theory and Applications*; Holmes, D.E.; Jain, L.C., Eds.; Springer: Berlin, Germany, 2008; pp. 117–130.
84. Schuwirth, L.; van der Vleuten, C.; Durning, S.J. What Programmatic Assessment in Medical Education can Learn from Healthcare. *Perspect. Med. Educ.* **2017**, *6*, 211–215. [CrossRef]
85. Heggarty, P.; Teague, P.A.; Alele, F.; Adu, M.; Malau-Aduli, B.S. Role of Formative Assessment in Predicting Academic Success Among GP Registrars: A Retrospective Longitudinal Study. *BMJ Open* **2020**, *10*, e040290. [CrossRef]
86. McMahon, C.J.; Tretter, J.T.; Redington, A.N.; Bu'Lock, F.; Zühlke, L.; Heying, R.; Mattos, S.; Krishna Kumar, R.; Jacobs, J.P.; Windram, J.D. Medical Education and Training within Congenital Cardiology: Current Global Status and Future Directions in a Post-COVID-19 World. *Cardiol. Young* **2021**, 1–13. [CrossRef] [PubMed]
87. Heeneman, S.; Oudkerk Pool, A.; Schuwirth, L.W.T.; van der Vleuten, C.P.M.; Driessen, E.W. The Impact of Programmatic Assessment on Student Learning: Theory Versus Practice. *Med. Educ.* **2015**, *49*, 487–498. [CrossRef] [PubMed]
88. Reichenberg, R. Dynamic Bayesian Networks in Educational Measurement: Reviewing and Advancing the State of the Field. *Appl. Meas. Educ.* **2018**, *31*, 335–350. [CrossRef]
89. Michell, J. Normal Science, Pathological Science and Psychometrics. *Theory Psychol.* **2000**, *10*, 639–667. [CrossRef]

90. Ricoeur, P. *Oneself as Another*, reissue ed.; University of Chicago Press: Chicago, IL, USA, 1995.
91. Taylor, C. *Sources of the Self: The Making of the Modern Identity*; Harvard University Press: Cambridge, MA, USA, 1992.